Nonprofit Sector Rationalization:
Measurement and implications for nonprofit finance and evaluation

Francisco Javier Santamarina

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Mary Kay Gugerty, Chair

David Suárez

Rachel Fyall

Benjamin Brunjes

Program Authorized to Offer Degree:

Public Policy and Governance

©Copyright 2023

Francisco Javier Santamarina

University of Washington

Abstract

Nonprofit Sector Rationalization:
Measurement and implications for non-financial contributions and impact evaluation standards

Francisco Javier Santamarina

Chair of the Supervisory Committee:
Mary Kay Gugerty
Evans School of Public Policy and Governance

This dissertation explores the relationship between process formalization or rationalization and resources among U.S. nonprofits and non-governmental organizations (NGOs) in international contexts. The first paper defines rationalization and explores the rationalization of rationalization itself, or meta-rationalization, by analyzing global standard-setting documents. I use structural topic modeling to identify six expressions of meta-rationalization and analyze distinctions across four domains of focus. The findings suggest that competing accountabilities would impose resource constraints when NGOs comply with multiple standards. The second paper proposes a rationalized approach to classifying in-kind donations of goods and services. The proposed taxonomy is designed to provide data that enhances understanding of what resources are necessary to deliver programs, improve program replication and scaling, compare subjective valuations, and improve measurement and theory building. In the third paper, I present a flexible method of measuring rationalization using publicly available data. I apply this method to a set of health nonprofits to identify what contributed service resources would most benefit performance.

Dedication

To the many random people who saw the value of my work and its uses.

To my family and friends for supporting me, and to my dogs for forcing me to get up, go outside, and breathe.

And, from now until the end, to my wife.

Acknowledgments

I would like to extend my thanks to my dissertation committee: my chair, Mary Kay Gugerty, David Suárez, Rachel Fyall, and Benjamin Brunjes. Mary Kay has worked with me since the very beginning, and she and David reviewed countless iterations of this work and both helped to shape it ever since I started to write my MAP. Rachel and Ben's guidance was paramount for me to see new perspectives. I am extremely grateful to Noah Smith, who served as both my GSR and as a guide and sounding board for my methods. His feedback at several times throughout my process helped to steer me towards the cleanest methodological paths.

Sharon Kioko has been a supporter since before I started my program, and I dearly appreciate her patience and generosity with me. Each of the faculty members at the Evans School during my time there were always kind and willing to share their time and thoughts with me – my thinking and my research reflects their communal willingness to grow new scholars, and I am grateful to each of them.

Much of this framing and work would not exist without my fellow PhD students. Conversations with Gowun "Gonnie" Park shaped my research from my MAP all the way through my dissertation, and everyone has and continues to provide so much friendship, rich discussion, space to ask questions and to be myself, and challenge me to think about my research in novel ways. To my D&D group: adventures with you all were a warm solace in a dark time, literally – the best part of winter was rolling dice with you.

Comments received from various faculty and fellow students at the conferences and student workshops at ARNOVA, ISTR, West Coast Nonprofit Data Conference, and the AOM PNP community were immensely helpful in shaping papers 1 and 2. Special thanks to David Frumkin

and Chao Guo for challenging some of my assumptions: your comments provided a critical inspiration.

I would like to especially recognize Jeff Brudney, whose guidance and seemingly off-hand comments profoundly shaped my research multiple times, peace be upon him.

Thank you to my parents and siblings. My mother supported me, even when I was supposed to be supporting her. My father gave me tireless guidance, insight, and time, and his formatting, visual guidance, feedback, and advice was essential to shaping every iteration that I presented of this research. They continue to shape the person that I am, even as an adult. My brother and sister always shared their thoughts and experiences, and never hesitated to point out an opportunity to make my research stronger.

I am deeply indebted to the countless people who spoke with me about their professional experiences, their research ideas, and shared their thoughts with a random PhD student who contacted them.

Finally, thank you to Samantha Lara, Pascal, and Pierre. Without their support and love, I would not have been able to finish this journey. Samantha – after all of your editing, feedback, thinking, encouragement, and strength, this PhD is as much yours as it is mine.

Table of Contents

List of Tables

Paper 1

Paper 2

Paper 3

Conclusions

List of Figures

Introduction

Paper 1

Paper 2

Paper 3

Conclusions

Introduction

Organizational processes and behaviors have become more structured and formalized in today's society. This long-observed phenomena within and across sectors has especially pronounced implications for nonprofits because of their role in society, which in turn produces burdens and expectations on how nonprofits deliver services, i.e., their processes.

Nonprofit theory suggests that U.S. nonprofits and non-governmental organizations (NGOs) in international contexts work to respond to perceived government or market failures and provides for societal or community needs not met by the other two sectors (ex. Salamon, 1987). In providing for these needs, they consume resources, transforming them from inputs into outputs and, so they hope, into broader social outcomes and impacts (W. K. Kellogg Foundation, 2004). The provided services and intended social outcomes can be targeted at virtually any level of society – from a neighborhood block or community all the way to national and global stages. As such, how efficiently and effectively nonprofits and NGOs' processes convert resources into outputs and outcomes can have significant implications for recipient communities. The relationship between this conversion by nonprofits and outcomes, or the effects experienced by communities, is not well understood, resulting in the dependence on proxy measures such as overhead expense (cost) ratios for performance (Coupet & Berrett, 2019). Unfortunately, such proxy measures yield detrimental effects for nonprofits: because these organizations are already often under-resourced, efforts that rely on proxy measures to improve nonprofits can starve them by forcing managers to reallocate resources in suboptimal ways that ultimately reduce the nonprofits' potential social impacts (Lecy & Searing, 2015). Improving the understanding of process and resource efficiencies, related performance, and how it is measured and understood can improve nonprofit theory as well as yield

positive effects for nonprofits and NGOs' programmatic outcomes and larger-scale societal impacts.

To explore this relationship between process formalization, resources, and outcomes, I use the perspective of rationalization. I define rationalization as the formalization of core (service delivery) and support (management) processes within a nonprofit (Dart, 2004; Hwang & Powell, 2009; Suárez & Hwang, 2013; Maier et al., 2016). This perspective is drawn from sociological institutionalism, in which formalization of processes can occur as a result of various isomorphic pressures (DiMaggio & Powell, 1983). The utility of understanding and intentionally transforming processes via formalization has long been explored in industrial engineering and business literature, for example, process design concepts (Sakamoto, 1989), business process redesign (Davenport & Short, 1990), and business process reengineering (Hammer, 1990; O'Neill & Sohal, 1999), to name just a few. While applications of these concepts have occurred across for-profit, public, and nonprofit organizations, the broader phenomenon of rationalization and its relationship with performances, i.e., resources consumed, outputs produced, and outcomes achieved, calls for further exploration in the nonprofit sector.

As of February 14, 2023, the IRS has records for 1,841,649 tax exempt organizations (Internal Revenue Service, 2023). This figure does not include the various community-based organizations, volunteer and social groups, and other expressions of collective action all across the United States that, while not legally registered as nonprofits, are critical elements of society at every level. Beyond the United States, the definition of and requirements for NGOs varies by country, and so understanding how third-sector organizations work in transnational and global contexts becomes further complicated. The unifying perspective that I draw from to discuss such organizations is the structural-operational definition of nonprofits (Anheier & Salamon, 1992; Schepers et al., 2005):

nonprofits have some minimal degree of organization, are private entities distinct from governments, are self-governing with respect to other organizations, do not distribute revenues outside of the organization, and do not wield compulsory or coercive authority. I also add to these elements that nonprofits, and to an extent NGOs broadly, are mission-oriented organizations for which profit maximization is intended to be subordinated to mission achievement while acting separate and distinct from government.

This definition establishes why rationalization is important to understand. Nonprofits and NGOs fulfil critical social roles, are often under-resourced, and face various pressures to allocate resources that may not allow them to focus on achieving their missions in the most effective ways. The degree of rationalization within the organization affects the volume and kinds of resources that they need to achieve their missions and intended social impact. Too little rationalization increases the required number of resources in ways that can reduce the size of their impact. Too much rationalization can reduce their ability to grow, to respond to change, and to reflect on whether some processes are outdated, i.e., inefficient. Ultimately, some degree of rationalization is necessary to reduce resource waste and allow nonprofits and NGOs to approach their theoretical, maximum ability to achieve their missions, while not passing the point of marginal returns at which increasing rationalization reduces that ability, as illustrated in Figure 1.

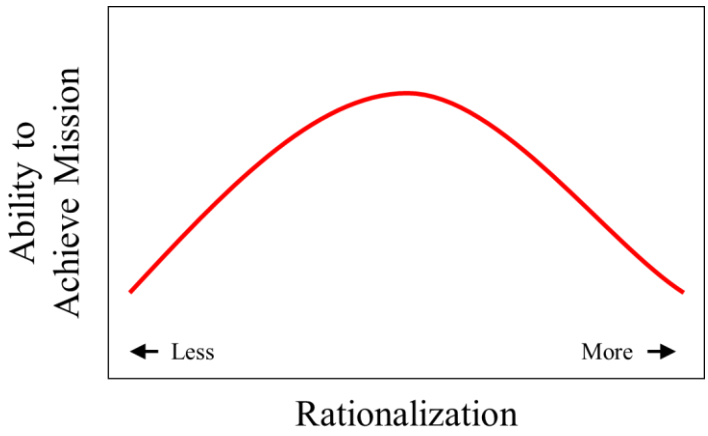

Figure 1. Proposed relationship between rationalization and mission achievement for nonprofits and NGOs.

Given my identified definition of nonprofits and NGOs, I propose that the ability to achieve mission is the ultimate performance measure for such organizations: implemented performance measures are intended to capture some element of or proxy this. This is in part because the ideal performance or peak ability and the point of marginal returns from increasing rationalization will vary by a nonprofit and NGO's internal and external factors. That said, I expect that there is some relatively ubiquitous minimum threshold for rationalization that allows an organization to be able to connect resources used as inputs to outputs, and possibly even outcomes. The goal of this dissertation is to build a foundation towards understanding the benefits from approaching that peak, consequences from exceeding it or falling short, and possible reasons that some minimum for rationalization can be good for organizations. I intend for this foundation to be applicable to any nonprofit or NGO. I use three papers to explore how expressions of rationalization can be understood and their potential consequences for NGOs, how rationalization can be used to address current knowledge gaps, and how rationalization can be measured across nonprofits at scale.

Overview

How might standard-setting documents cause NGOs to rationalize processes through formalizing impact evaluation?

The first paper explores this question by defining categories of rationalization and the rationalization of rationalization itself, referred to as meta-rationalization. Rationalization can result from external pressures, which can be pronounced for nonprofits and NGOs (organizations) because of their dependence on external actors for resources. These external actors can leverage resources to pressure organizations into rationalizing certain kinds of processes, which in turn can lead to other kinds of rationalization within the organization. In this way, external actors can have a significant secondary influence on how rationalization occurs and is expressed within organizations subject to those actors' pressures. These intended, primary pressures and unintended, secondary influences on rationalization have consequences for the organization's ability to achieve its mission (Figure 1). One category of such pressures is evaluation standards, which capture a kind of accountability relationship between NGOs and external actors that demands for NGOs to demonstrate or meet certain performance standards.

For this reason, I explore how global evaluation standards impose normative expectations around impact evaluation, itself a loosely and inconsistently defined concept. I begin by reviewing the literature and synthesizing a novel definition for impact evaluation. I then establish what rationalization is, and how it is distinct from the broader phenomenon of formalization and related phenomena such as managerialism and professionalism. I leverage a dataset of standards compiled as part of a previous research effort (Gugerty et al., 2021) and apply a novel text analysis technique, structural topic modeling. I explore the latent constructs uncovered via the modeling for trends in

the dataset and discuss generalized, practical findings as well as theoretical implications from and the utility of exploring meta-rationalization. Some of these findings include that most standards of impact evaluation identify the need to create change and to understand interventions and effects. The other four identified topics of compliance, producing and identifying community benefits, establishing systems, and engaging with data are not nearly as prevalent, and thus not as rationalized by standards. The prevalence of topics varies widely by standard and domain of focus: type of organization that produced the standard (by location in the development aid chain), regional focus of the standard (national vs. international), sectoral origin of the standard-setting organization ([inter]governmental vs. NGO), and the type of standard (self-regulation vs. third party). The findings suggest that, for a signatory organization to sign on to multiple standards, they would have to invest resources in rationalizing numerous distinct and non-overlapping processes that could limit their ability to be compliant, accountability to the standards and to stakeholders, and how well they can deliver services.

What are ways to value in-kind contributions of goods and services that can improve how nonprofits understand their outputs and outcomes?

The second paper proposes a rationalized approach to classifying in-kind donations of goods and services, in an effort to shed light on possible responses to the question above as well as build towards a more complete understanding of programmatic efficiency and performance. Current methods and understanding of classification of in-kind donations face various theoretical, legal, and practical limitations. As a result, they represent a class of underreported asset, an unobserved and empty space in the set of assets that nonprofits control and use to deliver services. Without this full picture of what is required to deliver services, and ultimately to achieve a nonprofit's mission, it is impossible to understand what it costs to achieve a particular social outcome and how

efficient a program is in achieving its outcomes, regardless of how near or far it is from the maximum impact that it could have produced. While it is likely impossible to achieve the theoretical maximum ability of a nonprofit to achieve its mission, incomplete understanding of what are the necessary inputs prevents meaningful progress towards that end. It also has implications for the most meaningful ways to rationalize the organization. The end result is that there can be little confidence in knowing where a nonprofit lies on a graph such as Figure 1, and thus efforts to replicate or improve what works well will always be flawed without that complete inventory of necessary resources as inputs.

I begin the paper with an overview of the current classification challenges and valuation methods. I then propose a novel solution, a modular taxonomy consisting of eight dimensions that can be applied as needed by researchers and practitioners to build a resource portfolio or inventory. In building inventories at the program level, researchers and practitioners alike will be able to perform more accurate analyses of true program-related expenses and resource consumption. This in turn will better inform efforts to replicate programs in other contexts and scaling programming, by anticipating potential resource constraints (such as availability, costs, necessary quantities) that are not traditionally captured or quantifiable. Furthermore, subjective valuations of resources by a nonprofit using the taxonomy can be compared to such valuations by other nonprofits, enhancing understanding of what nonprofits not only need but value and see as critical to their delivery of services. Finally, taxonomy-derived analyses can be used to generate performance measures that are not solely financial, externally valid, and can be reliably implemented. I then present applications of the taxonomy in generating sample propositions to better understand impact evaluations focused on efficiency (cost oriented) and efficacy (outcome oriented).

How can we better measure rationalization within nonprofits?

In the third paper, I set out to explore the proposed question in the context of donated services that may further rationalize organizations, and under what conditions that may or may not be desirable. Nonprofits accept donated services for a variety of reasons, e.g., reducing costs for necessary services, expanding the organization without reallocating existing resources, and opportunities to convert volunteers into fiscal donors to receive benefits over the long term. Donated services are often acquired in contexts external to the recipient nonprofit, and often to the nonprofit sector, and operate on assumptions about organizational operations and best practices that may not overlap with the capacities or mission of the nonprofit. At the same time, donated services rationalize the nonprofit by enhancing existing knowledge or introducing novel knowledge. Understanding the costs and benefits of this kind of rationalization is necessary to equip nonprofits with the knowledge of when to accept, reject, or redirect offers of donated services, in the service of approaching and not exceeding the theoretical peak indicated in Figure 1.

Before proceeding further, I realized that I need a means of measuring rationalization that can be easily and flexibly applied to large numbers of nonprofits. Current measures that scholars use for similar purposes include a variety of financial and non-financial measures with limited consistency across studies and utility as proxy measures, and the most specific rationalization measure has its own limits, including effort necessary to implement and ability generate using commonly available and public data.

I respond to these challenges with a proposed method of measuring rationalization that is flexible towards the data one has available. I apply a dimensional reduction technique known as the lasso to explore which variables among health-focused nonprofits' tax filings have the highest correlation with an implementation of the current dominant rationalization measure, as well as identify a subset that seem likely by drawing on literature. After identifying the algorithm- and

literature-derived variables, I test the performance of measures using each set as well as a combined set. I conclude with reflections on the performance of the flexible rationalization score and the findings related to health-focused nonprofits. Specifically, there are four observed modes or concentration of rationalization among this set of nonprofits, one above average, one at, and two below. More rationalized organizations would benefit from contributed services that increase novel knowledge and reduce rationalization such that the organization becomes more flexible in its processes to better achieve intended outcomes. Nonprofits with an average degree of rationalization would benefit from contributions that enhance existing knowledge, or increase rationalization, as well as provide novel knowledge and new processes (decreasing rationalization), depending on the combination of organizational need and content or focus of the contribution. Nonprofits with less rationalization would benefit from contributions that overlap heavily with existing knowledge, rationalizing current processes to reduce variation and related resource inefficiencies. Overall, the method used to produce the flexible rationalization score seems to work well, and the intentionally included areas for improvement (ex. using a more sophisticated algorithm, less heterogeneous dataset) will undoubtedly improve performance in future studies. Including literature-derived variables alongside algorithm-derived variables will have minimal negative effects on performance while also increasing theoretical validity and strength.

These three papers shed light on how rationalization affects resource demands, can impose constraints on what resources nonprofits and NGOs can use, and its implications for their performance. The first paper explores implications of rationalization, not at the level of an organization, but within individual components and processes of an organization. The second paper applies rationalization to fill a gap in contemporary literature and practice around in-kind

donations of goods and services, an underreported class of firm assets and resources with potentially large effects on an organization's performance and service impacts. The third paper develops a way to measure nonprofit rationalization flexibly and using existing data, a critical preliminary step for understanding how contributed services and external influences can rationalize organizations. The papers individually and collectively provide a set of evidence and tools to further progress towards answering my three driving questions, and further enhancing our knowledge of the relationship between rationalization of processes and resource use. In particular, we now know of meta-rationalization expressions among impact evaluation that could negatively impact NGOs in numerous ways. Understanding and accounting for meta-rationalization will yield benefits for organizations by allowing them to best distribute and allocate resources preemptively rather than in reaction to standards. Future studies can expand this by interviewing signatory organizations to explore how processes are explicitly affected by meta-rationalization and exploring this phenomenon in other industries and sectors. The taxonomy could be implemented and used to test theoretical assumptions about performance and necessary resources for all kinds of nonprofit programs. Its utility as a sector-agnostic tool, e.g., use for governments and businesses, can also be explored and tested via creating resource portfolios. This would allow scholars and researchers alike to better understood performance of any kind of organization or program. We can also now understand the influence of rationalization on how changes in the resource portfolio can affect performance via the flexible rationalization measure. This method could be tested across categories of nonprofits as well as geographies to push its boundaries, identify limitations, and create appropriate corrections.

While the three papers do not fully answer the proposed overarching questions, they do make progress towards understanding how influences from outside actors affect NGOs' processes, how

nonprofits serve their communities through consuming and allocating resources, and understanding how the degree of rationalization at a nonprofit affects their activities as well as what is needed for improvement.

References

Coupet, J., & Berrett, J. L. (2019). Toward a valid approach to nonprofit efficiency measurement. *Nonprofit Management and Leadership*, *29*(3), 299-320.

Dart, R. (2004). Being "business-like" in a nonprofit organization: A grounded and inductive typology. *Nonprofit and voluntary sector quarterly*, *33*(2), 290-310.

Davenport, T. H., & Short, J. E. (1990). The new industrial engineering: information technology and business process redesign.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American sociological review*, 147-160.

Gugerty, M. K., Mitchell, G. E., & Santamarina, F. J. (2021). Discourses of evaluation: Institutional logics and organizational practices among international development agencies. *World Development*, *146*, 105596.

Hammer, M. (1990). Reengineering work: Don't automate, obliterate. *Harvard business review*, *68*(4), 104-112.

Hwang, H., & Powell, W. W. (2009). The rationalization of charity: The influences of professionalism in the nonprofit sector. *Administrative science quarterly*, *54*(2), 268-298.

Internal Revenue Service. (2023, February 21). *Exempt Organizations Business Master File Extract (EO BMF)*. Retrieved March 7, 2023, from https://www.irs.gov/charities-non-profits/exempt-organizations-business-master-file-extract-eo-bmf

Lecy, J. D., & Searing, E. A. (2015). Anatomy of the nonprofit starvation cycle: An analysis of falling overhead ratios in the nonprofit sector. *Nonprofit and Voluntary Sector Quarterly*, *44*(3), 539-563.

Maier, F., Meyer, M., & Steinbereithner, M. (2016). Nonprofit organizations becoming business-like: A systematic review. *Nonprofit and Voluntary Sector Quarterly*, *45*(1), 64-86.

O'Neill, P., & Sohal, A. S. (1999). Business Process Reengineering A review of recent literature. *Technovation*, *19*(9), 571-581.

Sakamoto, S. (1989). Process design concept: A new approach to IE. *Industrial Engineering*, *21*(3), 31-34.

Salamon, L. M. (1987). Of market failure, voluntary failure, and third-party government: Toward a theory of government-nonprofit relations in the modern welfare state. *Journal of voluntary action research*, *16*(1-2), 29-49.

Salamon, L., & Anheier, H. (1992). In search of the non-profit sector. I: The question of definitions. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *3*(2), 125-151. https://doi.org/10.1007/BF01397770

Schepers, C., De Gieter, S., Pepermans, R., Du Bois, C., Caers, R., & Jegers, M. (2005). How are employees of the nonprofit sector motivated? A research need. *Nonprofit Management and Leadership*, *16*(2), 191-208.

Suárez, D. F., & Hwang, H. (2013). Resource constraints or cultural conformity? Nonprofit relationships with businesses. *Voluntas: International Journal of Voluntary and Nonprofit Organizations, 24*(3), 581-605.

W. K. Kellogg Foundation. (2004). Logic model development guide: Using logic models to bring together planning, evaluation, and action. Battle Creek, MI: Author.

Paper 1. Meta-Rationalization of Impact Evaluation

Introduction

In the field of international development, evaluation is the use of scientific methods to try to estimate the outcomes of implemented programs, attempted efforts, and organizations (Bingham & Felbinger, 2002). They are often an accountability mechanism for non-governmental organizations (NGOs) to themselves and to others (ex. donors, stakeholders; Ebrahim, 2003). Because evaluation is such an important accountability mechanism for international development, institutional actors have sought to codify evaluation expectations and requirements into standards. Standards consist of mandatory and voluntary self-regulatory programs that capture expectations for norms and behaviors of NGOs that have adopted or sign-on to a given standard, i.e., "signatories" (AbouAssi & Bies, 2018; Gugerty et al., 2021). NGOs commit to evaluation standards as accountability mechanisms to signal quality in their work or in response to external threats such as scandals or potential government regulation (Ebrahim, 2003; Gugerty, 2009). Evaluations within international development can vary widely in complexity and compliance requirements, which in turn can place steep resource demands on organizations (Cumming, 2008; Prakash & Gugerty, 2010). Furthermore, NGOs experience numerous isomorphic pressures to become signatories to evaluation standards, including dependence on external resources and aligning with expected behaviors (coercive isomorphism), emulating partner NGOs (mimetic isomorphism), and through learning networks and membership groups (normative isomorphism) – though only normative isomorphism has been shown to have a statistically significant relationship with standard adoption (AbouAssi & Bies, 2018; Bromley & Orchard, 2016; DiMaggio & Powell, 1983).

Evaluation standards may lead to increased formalization of organizational characteristics and processes by requiring signatories to perform compliant evaluations, regardless of whether such compliance aligns with signatories' existing practices or not at all. Such pressures have implications for how signatories allocate resources to respond to competing accountabilities, though they can vary by kind of evaluation, e.g., formative, process-, or impact-oriented (Cumming, 2008). In particular, impact evaluations yield not only challenges of isomorphism and complexity but also numerous benefits: they can be used to measure program performance and societal effects, propose accountability for long-term effects, and allow organizations to learn and change their activities as appropriate (Ebrahim, 2003; Rossi, Lipsey, & Freeman, 2004). Because of these same reasons, impact evaluation standards can be highly prescriptive to ensure these benefits are achieved.

This paper studies potential for standards to trigger *rationalization*, operationalized as organizations' formalization of their processes in compliance with evaluation standards. Research on rationalization gained relevance within the field of evaluation during the past two decades, where it is explored from the perspective of managerialism. In the context of civil society organizations (CSOs[1]), Hvenmark (2016) defines managerialism as the "ideology prescribing that organizations ought to be coordinated, controlled, and developed through corporate management knowledge and practices" (p. 2849).[2] Multiple scholars have drawn on Hvenmark's (2016) definition to ground studies and explorations of managerialism in various NPOs and NGOs (see: Maier et al., 2016; Mitchell, 2018). This definition of managerialism as focus on the formalization

[1] Which I interpret as roughly analogous to NGOs in the context of the Hvenmark (2016) article.

[2] "Corporate" in this context is analogous to private or business, as implied by Hvenmark's (2016) presentation of the three sectors of private/business, government/public, and nonprofits/NGOs as "the corporate world, public agencies or CSOs" (p. 3825), where civil society organizations (CSOs) is one of the names used to reference the nonprofit sector (Anheier and Salamon, 2006).

of organizational management knowledge and practices, however, has been criticized as unspecific and too broad (Roberts et al., 2005; Suárez & Hwang, 2013; Suárez & Gugerty, 2016).

In light of this critique, this research goes beyond the corporate-derived domains of managerialism (Mitchell, 2018); I focus instead on the more specific concept of rationalization. I draw on definitions by Dart (2004), Hwang & Powell (2009), Suárez & Hwang (2013), and Maier et al. (2016) to frame rationalization as *the formalization of core (service delivery) and support (management) processes within a nonprofit*. In this context, processes refer to the activities that organizations undertake to transform resources into some desired output or outcome, in line with the notion of "activities" in a logic model (W. K. Kellogg Foundation, 2004). The proposed definition accounts for managerialism and focuses on the broader phenomenon of formalization, regardless of sectoral source of influence.

Existing literature on rationalization within impact evaluation suffers from siloed studies (by country or region), limited sample size (e.g., respondents), and limited focus on impact evaluation (Castro et al., 2016; Cummings, 2008; Podems, 2014; Raina, 1999; Roberts et al., 2005; Schwandt, 2017). Several articles reference internal and external pressures for more impact measurement as a contributing factor to increases in expressions of managerialism, including rationalization and professionalization (Roberts et al., 2005; Suárez & Gugerty, 2016). Examples of those expressions include the task description in European evaluation job solicitations (Castro et al., 2016), a call from Russian groups for the government to "institutionalize regulatory impact assessment" (Podems, 2014, p. 131), and French development NGOs' inclusion of tools used in impact evaluation (i.e., feasibility studies and logical frameworks) in their proposals to the French Foreign Ministry for engaging in fieldwork abroad (Cumming, 2008).

In this paper, I explore how the language used in standard-setting documents for impact evaluation establishes patterns of expectations around rationalization at multiple levels within NGOs. In doing so, standards are not just accountability mechanisms but also are vehicles through which pressure is exerted on NGOs. Standards can trigger or cause changes to NGO processes either directly or indirectly, i.e., modifying an existing, internal source of pressure. While accountability mechanisms exist to ensure and improve "justice in the use of the world's power and resources" (Murtaza, 2012, p. 122), they can be resource-intensive for NGOs and divert resources away from serving program participants (Ebrahim, 2003; Epstein & Klerman, 2013). In addition, NGOs are increasing compelled to develop an understanding of their programmatic effects at scale (ex. across a society or time period) using various kinds of impact evaluations, which they may perform to comply with external pressures, improve internal processes, or both. Given the scale of impact evaluation requirements, the resource demands to achieve these complex and sometimes nebulous measurements can be taxing on NGOs. I examine rationalization within the context of impact evaluation because of the high importance placed on attempts to understand societal impact, the evolving expectations around how to achieve this, and the ensuing pressures and demands experienced by NGOs.

This paper focuses explicitly on impact evaluation and on standard-setting documents as potential mechanisms for rationalization of related processes. Where previous literature aims to explore the factors affecting standard adoption, such as isomorphic pressures from the institutional environment (ex. Bromley & Orchard, 2016), I aim to identify potential organizational implications from adopting standards – what does isomorphism imply for organizational processes, rather than how or when does it occur. I also use the focus on impact evaluation processes to explore the potential for internal changes to organizations. I build on the previous

literature by applying an interpretivist, inductive approach to a dataset of standards that reflect a broad range of organizations within the international development space. I use structural topic models to identify the most dominant expressions of proposed formalizations for impact evaluation within the dataset, referred to as latent constructs (or topics). I then explore and analyze those topics to generate observations about dominant trends among impact evaluation standards and possible implications for signatories based on what kind of organization created the standard, its regional focus, the nature of its work, and the kind of standard that it produced. This study will help donors to make more informed decisions regarding evaluation requirements and capacity demands, empower NGOs to better anticipate evaluation expectations, and shed a novel light on expectations for international development actors and their ability to adapt to their work.

Defining Impact Evaluation

"Impact evaluations are designed to answer the question: 'What was the effect of an intervention on an outcome?'" (International Initiative for Impact Evaluation, n.d.). Impact evaluation is a complex process that can vary in its specificity and composition – what it refers to can vary widely by scholars, standard-setting organizations, and implementing NGOs. To understand the range of dominant framings of the concept, I reviewed prominent articles and grey literature produced by influential organizations in this space and I identified 13 sources[3] that present explicit definitions which vary widely in the concepts that are presented and emphasized. I categorize my results according to the elements that consistently appeared across sources: randomized control trials (RCTs) and counterfactuals, responses to or rejections of RCTs and counterfactuals, and

[3] I identified these sources by starting with influential organizations, scholars, and articles in the space, then seeing who they cited and who cited them. I stopped including additional sources once I achieved a high degree of data saturation such that any further definitions provided minimal additional, novel detail.

formalized understandings of the world through logic models and theories of change. Afterwards, I present a formal definition of impact evaluation that synthesizes and addresses the identified elements. For a more detailed discussion of each section below, please refer to Appendix 1A: Impact Evaluation.

RCTs and Counterfactuals

RCTs and counterfactuals as methods dominate the practice of impact evaluation with their use of treatment and control (or comparison) groups to establish counterfactuals for causal attribution of program (intervention) effects and outcomes (Bernard, Delarue, & Naudet, 2012; Gertler et al., 2016; Gibson & Sautmann, 2021; Ravallion, 2001; Shah et al., 2015). Counterfactuals are emphasized regardless of whether group assignment is random (see Gertler et al., 2016; Gibson & Sautmann, 2021). They provide an observable, measurable, and testable basis for comparing how program recipients (treatment group) and non-recipients (control group) fared during and after implementation, which provides a positivist determination of a program's effect on outcomes, or its impact. Once produced, counterfactuals and causal attributions are used to explore program effectiveness and theories of change (Gibson & Sautmann, 2021).

While quantitative data is an explicit focus for such impact evaluations (see Gibson & Sautmann, 2021), mixed-methods studies that use quantitative and qualitative are also seen as viable impact evaluations (Shah et al., 2015). More broadly, Ravallion (2001) notes that "evaluation is essentially a problem of missing data" (p. 137); they and others identify the utility of alternative methods to construct or analyze counterfactuals that compensate for missing data and allow for causal attribution, such as quasi-experimental designs (International Initiative for Impact Evaluation, n.d.; Ravallion, 2001).

Some see RCTs as but one step of impact evaluation, and impact evaluation as but one part of implementation. For example, it can be a learning mechanism among pilot programs before scaling up, produce evidence that can influence policies, and contribute to wide-reaching efforts to reduce poverty (Bernard et al., 2012; Innovations for Poverty Action, n.d.-a; Innovations for Poverty Action, n.d.-b).

Challenging the "When" and "What" of Impact Evaluations

"When" to conduct

Impact evaluations are not always appropriate. For example, it is a resource-intensive process, so it should be applied when there is a need for outcome-related information and the program is well-structured and defined (Rossi et al., 2004). They are best for new(-ly growing) programs and programs with uncertain effectiveness, and should generate generalizable knowledge beyond a single program (Savedoff et al., 2006).

"What" & how to define

Similarly, there is no consensus among authors and organizations as to how define impact evaluation. Some question whether "attribution" is achievable and note that it implies an exclusive relationship between interventions and impact that is unlikely (Stern et al., 2012). Others broaden the scope to the measurement of programmatic effects on societal conditions (Rossi et al., 2004). A group of NGOs proposed a definition oriented on "lasting or significant changes-positive or negative, intended or not-in people's lives" (Roche, 2000, p. 546) where contextual information and judgments would drive what was and was not considered impact.

Logic models and theory of change

There are efforts to ground impact evaluation within the language of theory of change (TOC) and logic models. Schaffer (2011) identifies impact evaluation as focused on the last two elements, outcomes and impacts, referring to "longer-term effects of projects usually on some dimension of well-being" (p. 1621). Stern et al. (2012) explicitly note that impact evaluations should be consistent with TOC logic.

Identifying gaps in definitions

There are also gaps in the common definitions indicated by the literature used to train evaluators. In their analysis, Gugerty et al. (2021) identify certain frequent concepts: internal validity, disclosure of methodological limitations, theory-based evaluation using formalized concepts and methods (ex. theory of change, logical frameworks), attempts to establish causal attribution, using both qualitative and quantitative methods (or even explicit "mixed methods"), and the use of a counterfactual. While not all of the concepts are presented in the cited literature here, their presence in discourse suggests that each of these concepts compose some part of impact evaluation.

Proposing a Definition of Impact Evaluation

Given this wide variety of ways in which the concept of "impact evaluation" is defined and operationalized across literatures and discourses, I synthesize my own definition to use as an analytical foundation:

> Impact evaluation is a systematic process to identify the relationship between programs and their effects through collecting evidence and making claims about the relationship between program-specific activities and observed changes in the world.

This definition is meant to address the key elements that appeared consistently in the definitions above as well as elements that, while not common, were highly salient or emphasized when they did appear. It allows for a broader perspective of evaluation that is more global in scope: it is less rooted in specific methodologies and disciplinary approaches; it allows for culturally specific interpretations of relationships between activities and changes; it makes space for the various approaches of implementing organizations, which may be more fluid and less prescriptive in their internal procedures around performing impact evaluations. Finally, the definition allows for both attribution and contribution: as noted by Stern et al. (2012), there is a tension as to whether evaluation should measure "the intervention as the cause of the impact" and that specific relationship (*attribution*) or if it should measure the extent to which the intervention was one of multiple factors that may yield or contribute to an observed impact (*contribution*; p. 38). The extent to which impact evaluation should pursue one versus the other is something that I make space for the data to indicate, rather than impose my own beliefs. I do not use the word "evaluate" in the definition because the positivist implication of words such as "evaluate" and "assess" found in common definitions (ex. Gertler et al., 2016, p. 7) may preclude discussions of impact evaluation that are less focused on assessing the validity of assumed causal relationships and more focused on sense- and meaning-making. With the definition in hand, I can apply it to the data to identify the various ways in which standard-setting organizations discuss impact evaluation.

Meta-rationalization: the rationalization of rationalization

By evaluating impact, one is evaluating the relationship between processes of a program and observed changes in the world.[4] But what is a process? I propose that a *process* is a series of

[4] Or however the organization chooses to define impact.

sequential actions, tasks, or steps performed to achieve a desired output, end state, or goal. This is done through the consumption of resources, such as staff time, using computers to send and receive information, organizational reputation, and other tangible and intangible resources.

There are commonly multiple permutations of a process that can yield the same output. I identify *formalization* as the reduction of those permutations, with the goal of codifying and standardizing the steps performed to achieve the output. Figure visually presents two expressions of process formalization, professionalization and rationalization. I consider professionalization to be the formalization of *what* is to be performed; once specified, this is then used to determine *who* can perform a given set of steps (see Figure , Step C filled with vertical lines). Only certain roles are permitted to perform a portion of or all steps in a process. This draws from discussions of professionalization where members of a profession attempt to define their work, place controls on its production, and legitimate it, and commonly refers to substantive, discipline-specific professions (DiMaggio & Powell, 1983; Hwang and Powell, 2009). For example, an organization may require that the employees preparing financial statements as specified by an accounting board are certified accounting professionals: the entire process is now limited to certain roles. In terms of international development, an organization may require that a theory of change and related tools, such as logic models or logical frameworks, must be prepared by a staff member who has taken a related training course. Only a theory of change created by the "approved" staff member will produce the desired output – with implications for the NGO's workforce. Any number of steps up to an entire process may be formalized, not solely the piecemeal example provided in Figure . A focus on professionalization would result in processes being formalized in accordance with the roles able or permitted to complete steps.

In contrast, rationalization is the formalization of *how* a given set of steps is performed (see Figure , Steps A & D filled with upward diagonal stripes). Only certain steps can be performed, and the order is predefined. For example, the United Nations Evaluation Group's Norms and Standards for Evaluation (2016) identifies that "the evaluators or the evaluation teams must be selected through a transparent and competitive process" (p. 25). Performing the steps in reverse or alternate order would not yield the expected or desired output: an organization that designs a transparent process after the evaluation team is selected would not produce a transparently selected team. In line with the nonprofit literature, I categorize processes relative to the instrumental nature of nonprofits as core (or service delivery) and support (or management). The focus on how such processes are formalized thus aligns with the literature-derived, proposed definition of rationalization as "the formalization of core (service delivery) and support (management) processes within a nonprofit." [5] Core, service delivery processes are often synonymous with mission-oriented activities, as these activities consist of processes around delivering services that directly align with the core of the organization, i.e., its mission or stated purpose. Support, management processes are often distinct from mission-oriented activities because they consist of processes that do not provide direct service delivery, but rather enable or facilitate such actions. For example, while proper financial accounting processes may not be directly related to a development NGO's mission-oriented activities, they are essential to providing the inputs required for those processes. Because impact evaluation is not focused on delivering direct services but understanding how to measure and understand those services, it is primarily considered a support process.

[5] For the rest of this paper, I will use formalization of processes to refer to the "how," or rationalization, rather than the "what/who," or professionalization.

Impact evaluation can lead to formalization because it yields information as to how well processes are able to achieve desired, observed changes in the world. This information can be used by an organization to reduce the permutations of processes that compose a program and further specify how it should be implemented moving forward or in the future. Some examples of formalization resulting from impact evaluation include organizational streamlining and benefits such as reduced costs and wasted resources, improved production times, greater consistency in process outputs, and greater ease in making adjustments, e.g., incorporating a new step, across all permutations.

When focused on specifying the steps or "how" of the process, impact evaluation can yield rationalization, with different types of impact evaluation methods varying in their degree of potential rationalization. While impact evaluation may not always cause changes in practice, I interpret such changes to be part of the fundamental intent, and so I consider impact evaluation to be, at its core, a method of rationalization.

Similarly, standard-setting documents are vehicles of formalization because they attempt to codify and standardize signatories' processes. When they mandate that signatories perform impact evaluation in a certain way, such standards establish expectations that, for compliance to be met, a signatory may need to further formalize a process. From one perspective, the standard establishes expectations for how a signatory should rationalize their impact evaluation processes. In other words, to be compliant with the standard, the signatory will experience a reduction in the possible ways that it can conduct impact evaluation. Returning to the depiction in Figure , this would mean that the signatory's possible permutations of steps, or process, is reduced from 3 to 2 to achieve a compliant output, or complete impact evaluation.

What happens when the output of one process affects the possible steps for another? Impact evaluation, a support-oriented process, can yield rationalization for an organization's mission-

oriented process. Thus, a standard addressing impact evaluation has implications for formalizing how the signatory rationalizes its affected mission-oriented processes, such as limiting or perceiving a different range of steps that could be formalized. For the sake of clarity, I refer to this second perspective as *meta-rationalization* (see Figure , Step G with downward diagonal stripes). Figure suggests that professionalization, rationalization, and meta-rationalization do not overlap. The distinction is merely for the sake of illustration; I do not anticipate mutual exclusivity, but rather some degree of overlap between professionalization and the other two expressions. Table 1 presents a comparison of what the three levels of analysis of process, rationalization, and meta-rationalization look like in the context of international development NGOs.

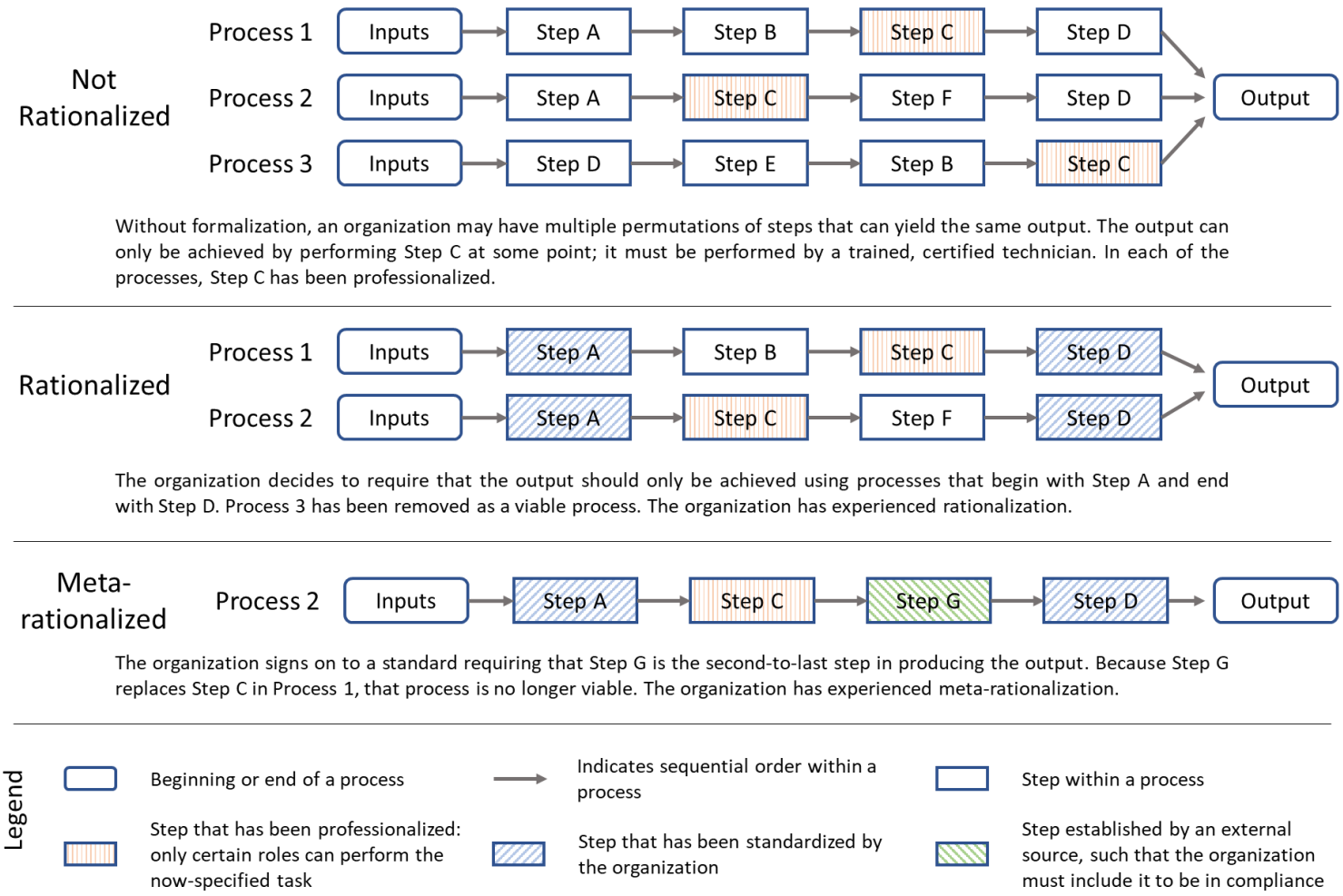

Figure 1. Comparison of professionalization, rationalization, and meta-rationalization showcasing permutation reduction and related modifications of intra-process steps. Numbers and letters are meant to distinguish processes and steps, not to establish an order. Order among steps is indicated solely by the directional arrows. The distinction between the kinds of formalization is purely illustrative and not meant to imply exclusivity.

Consider the following example of a hunger-reduction program. In response to famine experienced by certain rural communities, an NGO develops a program to distribute packages of rice (Table 1, level 1). Because of various factors, these communities have never been part of this kind of hunger-reduction program before. As part of its mission and core values, the NGO wants to ensure that the most vulnerable members of the affected communities have sufficient access to the rice. They

also want to determine if families are seeing a reduction in hunger, the current distribution methods are culturally appropriate, the rice is being bartered in ways that undermine or even hurt the local economies, and other possible impacts. To evaluate such impacts, the NGO has decided to try and get a range of beneficiary feedback. They perform interviews with community leaders, supplemented by a survey of served household by community, depending on staffing and time constraints; not all communities are surveyed (Table 1, level 2). The NGO begins to modify rice distribution methods in response to the ongoing, interview-focused evaluations, and it establishes distribution sites at key community landmarks identified in the evaluations (ex. religious centers, homes of village leaders).

Halfway through the evaluation, the NGO signs on to a globally recognized international development standard. The standard is quite rigorous, with topics ranging from data privacy to how to conduct impact evaluations. To be in compliance, the NGO must use a participatory-focused approach to impact evaluation centered on voices from all served communities. The standard specifies that surveys are preferable to interviews but recognizes interviews as useful sources of supplemental data. Because of staffing and time constraints, the NGO decides to focus solely on conducting surveys instead of the more staff-intensive interviews (Table 1, level 3). As a result of the survey responses from the evaluation, the NGO modifies its service delivery activities and begins to distribute the rice at different community landmarks (ex. sources of water, markets).

Table 1. Levels of analysis for meta-rationalization

	Level of Analysis	Description	Contextualized to NGOs & Impact Evaluation	International Development Example
1	Process	A series of sequential tasks or steps performed to achieve a desired output, end state, or goal	The series of implemented steps that an organization evaluates. It is often a mission-oriented service.	An NGO develops a program to distribute rice in rural communities that are experiencing famine.
2	Rationalization	The extent to which a process is formalized, or the formalization of a process	The methods used by the organization to evaluate and reduce variation in the process. It is often a support-oriented service.	The NGO evaluates the impacts of its rice distribution program using a combination of interviews with local leaders and, if time and resources permit, community surveys.
3	Meta-rationalization	The extent to which rationalization is formalized	The extent to which the organization must follow certain steps required by a standard when evaluating the process. The standard originates outside of the organization or unit that is implementing the process.	The NGO has signed on to a standard that specifies using a participatory-focused approach to impact evaluation, and so the NGO must survey all served communities as part of the evaluation process to be in compliance. This new focus does not leave enough time or resources to also conduct interviews. How the NGO distributes rice is affected.

Standards of Impact Evaluation

In this paper, I explore the language for impact evaluation used in standards to identify patterns of meta-rationalization, which in turn may demonstrate isomorphic pressures experienced by signatories. “Standards” can be mandatory or voluntary and serve to signal quality and ethical behavior as well as to enhance accountability and performance (Ebrahim, 2003; Gugerty, 2009). NGOs sign on to standards due to isomorphic pressures from dependence on external resources (coercive isomorphism), participation in networks and membership groups (mimetic isomorphism), and partnerships with other NGOs (normative isomorphism), though only normative isomorphism has been shown to have a statistically significant relationship with standard adoption (AbouAssi & Bies, 2018; DiMaggio & Powell, 1983). This literature provides

insight into expressed external isomorphic behavior, i.e., whether adoption occurs or not and related nuances.

What about the predicted interactions between internal changes and two or more external sources of pressure? For example, Gugerty et al. (2021) note that resource concerns may affect how standards produced by membership-based organizations describe impact evaluation and present requirements. These concerns may increase or even evolve into explicitly conflicting resource requirements when organizations attempt to comply with multiple standards that have distinct rationalization expectations for impact evaluation, ex. one emphasizing randomized control trials (RCTs) and another emphasizing qualitative or participatory impact evaluations (ex. Johnson & Rasulova, 2016; Van Hemelrijck & Guijt, 2016). I aim to add additional nuance on the range of focuses for expected resource demands that may affect or influence signatories' willingness to commit resources and be in compliance with standards.

Exploring meta-rationalization through standards, in particular those related to and that affect impact evaluation, can help us understand the extent to which standards can produce competing, collaborating, or even synergistic pressures on signatories. For example, the perspective of global accountability communities as driving NGO standards and practices acknowledges the complex interplay between various actors and goes beyond the "singular principal-agent relationship where salient donors and impose financial and legal accountability standards upon NGOs" (Deloffre, 2016, p. 746). Applying this perspective to international development requires us to acknowledge that multiple sources of pressure (or sets of standards) within the same community can drive intra- and inter-organizational changes. Such changes can include the formalization of impact evaluation and similar monitoring, evaluation, accountability, and learning (MEAL) practices that rationalize processes. Meta-rationalization adds an additional analytic layer by exploring how different kinds

of standards in the broader (global) community establish distinct implications for processes within individual organizations.

I explore meta-rationalization by using theory- and literature-derived domains that may influence or be influenced by impact evaluation standards, and thus shape the expressions of meta-rationalization in this space and implications for organizational processes. The four domains include: 1) the type of organization producing standards, 2) the intended scope of focus for the standard (i.e., regional, or national vs. global), 3) the sectoral origin of the producing organization, and 4) whether the standard is self-regulatory or developed by a third party.

Organizational Type. Organizational type captures the role of the standard-producing organization in international development. It also aligns with the organization's placement in the development aid chain, typically consisting of four roles: donors, Global North-based NGOs, local implementing partners, and project beneficiaries (Luchner, 2018). In this paper, I reframe placement in the chain in relation to implementation proximity, e.g., closer to the ground (direct project implementation or interactions with target communities) or further from the ground (less direct implementation, focus on accountability to funding sources such as tax-paying citizens). "Distances" or placements in the chain may contribute to certain perspectives on impact evaluation shared by others at the same placement. Some standard-setting organizations operate outside of the direct chain by providing guidance to the overarching infrastructure of international development from an external or more removed perspective. To capture a producing organization's placement in the aid chain, I use three classifications of organizational type: donor organizations, infrastructure organizations, and associational or membership organizations.

Regional focus. Regional focus can capture shared cultural values that influence organizational practices and responses to country-specific development challenges. I anticipate that the need to

focus on a single political state ("national") as the operating environment or the community to which an organization is accountable will produce different kinds of expectations than a standard that applies to multiple countries or globally ("international, cross-national"). Specifically, standards created by international, cross-national, or transnational organizations will not reflect voices and inputs from national organizations to the same extent as standards created by national organizations. The creation of national organizations' standards often includes solicitations for input from community-based organizations, grassroots efforts, and organizations operating at various levels within a given nation (Kelly, 2021). National standards can reflect nations' beliefs and values such that even a minimal degree of region-specific input will introduce heterogeneity to a globally focused set of standards. This would suggest that there will not be a single dominant expression of meta-rationalization among national standards, and that the distribution of expressions based on their appearances will be more flat or less normally distributed. In contrast, standards from international organizations are more likely to reflect values "averaged out" across countries, those shared by countries with the greatest influence on the organization, or be attempts to go "beyond" country-inspired values. This would suggest that international standards will not reflect signatories' values as consistently as national standards, and that they will be more homogenous with each other. Expressions of meta-rationalization among international standards are thus more likely to be concentrated such that some expressions will be highly prevalent or even dominant, and others not prevalent or even non-existent.

Sectoral origin. The sectoral origin of the standard-setting organization can present an intersection of accountability to the public and pressures from multiple standards. The structural-operational definition of nonprofits suggests that the private nature of nongovernmental organizations (NGOs) and their focus on voluntary engagement will yield differences in impact evaluation expectations

as compared to the public nature and compulsory authority of (inter)governmental organizations (Salamon & Anheier, 1992). NGOs' accountability structures and "to whom" they are accountable vary by the nature of their work; such parties include donors, staff, partners, clients or beneficiaries, and other relevant stakeholders (Brown & Moore, 2001). Standards produced by NGOs may similarly reflect those sometimes-nebulous and context-specific accountability structures, suggesting those standards will present a less consistent need to formalize processes.

(Inter)governmental organizations have more authority than NGOs but experience greater demands for accountability and transparency. At the same time, they may have more diffuse lines of accountability, e.g., not just donors but all taxpayers, either directly (as in a local government that both imposes a tax and implements programs with those funds), more indirectly (a national government whose projects may or may not directly benefit taxed individuals), or most indirectly (a United Nations program whose projects are funded by national governments' revenues from taxing individuals). With the increase in potential stakeholders, (inter)governmental organizations must become more explicit in their actions and programming to be sufficiently transparent and accountable. Such organizations may operate with more formalized processes out of a need to comply with various accountability requirements and regimes. This yields standards that are, in turn, more explicit about compliance requirements – the (inter)governmental organizations transmit their own responses to accountability demands through the standards that they set for recipient NGOs.

Standard Type. The category of a standard implies various aspects of how and why an organization may commit. Standards that are self-regulatory and created by membership organizations act as intended signals of virtue, quality, and ethical actions (Gugerty, 2009). Members may also buy into those standards to gain access to a community of similar organizations or to specialized

knowledge related to their main activities. Standards created by third-party organizations, such as rating agencies, governments, and global initiatives may act as different or signals of quality. Gugerty (2009) notes that "programs that explicitly respond to donor interests tend to use stronger verification mechanisms and to provide higher explicit benefits" (p. 264). Given the influence of donors on both creating and reinforcing the adoption of third-party standards, such standards may use more rigorous verification methods than those of self-regulatory standards, which could increase the perceived legitimacy of third-party standards over self-regulatory standards. This can also contribute to the perception of third-party standards as more universally recognized signals of accountability, even if not as specific or applicable to a particular signatory's activities.

My analysis includes multiple standards to identify trends across the field of international development. As part of this analysis, I identify groups of standards that have common linguistic characteristics, i.e., word constructions, regarding impact evaluation. I explore characteristics within and across categories of standards to understand potential global trends. By focusing on rationalization, I will build on AbouAssi and Bies' (2018) article by exploring isomorphic tendencies within standards and implied post-adoption consequences.

Data & Methods

Dataset

The dataset used in this paper consists of documents created by various organizations that set standards for practices and behaviors related to international development interventions across and within countries. The documents are created by national associational organizations, donors, and infrastructure organizations, i.e., organizations that may be based in one country but establish norms and behaviors across boundaries without the ability to directly use financial resources as

incentives for adherence. Originally compiled by Gugerty et al. (2021), the research team reviewed global inventories of standards (Tremblay-Boire, Prakash, & Gugerty, 2016), members of major international organizations such as CIVICUS and Forus, and additional organizations that may meet the search parameters. The complete list identified by the research team included 193 organizations.

To be included in the dataset, documents had to explicitly address evaluation as defined by the team[6], have the full text publicly available, and be written in English (or have official English translations available). The dataset was expanded until the research team was reasonably confident that it met the principle of data saturation (Saunders et al., 2018), for example, when additional standards repeated or even directly cited identified documents for the vast majority of the eligible text and no potential standards were identified that provided novel information. The final dataset consists of 57 documents produced by 42 organizations, or "sponsors" of standards.

Qualitative Coding

This paper uses an exhaustive coding approach to identify all text in the dataset that contains concepts relevant to the definition of "impact evaluation" established earlier. In using this definition, we can identify explicit expectations imposed through standard-setting documents onto organizations as to how they should perform impact evaluation. The standardized definition allows for identification of comparable text in the dataset documents, regardless of variation among internal definitions of impact evaluation. Analysis of the relevant text will allow me to identify

[6] Gugerty et al. (2021) compiled a codebook of 65 codes, 54 of which captured some element of evaluation. The codes were defined using "a wide array of monitoring and evaluation concepts, including concepts commonly articulated in evaluation guidebooks and textbooks, evaluation courses and trainings, and in our own experience as evaluators and teachers of evaluation" (p. 5). If an identified document did not have at least one of those 54 codes, then it was not included in the project's final dataset and analysis. For additional detail on the codebook, please refer to the article and its Appendix.

expressions of meta-rationalization in the international development sector that may complement each other or may cause friction, with implications for how signatory organizations allocate resources towards core vs. support services. To ensure reproducibility, I used the publicly available copy of the dataset created by Gugerty et al. (2021).

I used the qualitative coding software Atlas.ti 9 to manually review each document and identify relevant text. As noted by Campbell et al. (2013), there is some debate around focusing on sentences, paragraphs, and other "clearly demarcated parts of the text" vs. "units of meaning" that are reliant on coder subjectivity in interpreting and knowledge of the applied code. While this issue is somewhat minimized because the documents are constructed and formatted by the organizations themselves, the question still remains as to focus on established demarcations or units of meaning that could spill across multiple demarcations. In my dataset, there is also inconsistent formatting across documents, as some are composed of single-sentence lists or captions and others of full paragraphs and pages.

I focused on the most common demarcation element, a single sentence, as the unit to be analyzed for the presence of impact evaluation as defined earlier. I define a sentence as a string of text that contains a terminal punctuation or clear completion of thought. In addition, lists where each entry is an incomplete clause were treated as single units, whereas those with complete clauses were treated as consisting of multiple units (i.e., each complete, independent clause was reviewed as a single sentence). Relevant text within the sentence had to contain some combination of key "units of meaning" or concepts from the definition. Table 2 demonstrates how I separated the definition into component concepts and which concepts, on their own, indicate the presence of impact evaluation and which concepts, on their own, are insufficient to do so. If the necessary, minimum combination of concepts was present, I coded that text as having "impact evaluation" present.

Table 2. Comparison of Coded Definition Concepts

Concept	Centrality to the definition	How many other concepts necessary
systematic process	Low	At least one Medium
identify relationship between programs and their effects	Medium	At least one Low or Medium
collecting evidence	Low	At least one Medium
making claims about the relationship between program-specific activities and observed changes in the world	High	None
Programs/program-specific activities and their effects/observed changes in the world	Medium	At least one Low or Medium

The complexity of defining impact evaluation compelled me to account for instances where "units of meaning" is present in or spills across multiple sentences. I did so by implementing a window of +/-1 to capture potential edge-case units: if a sentence partially but insufficiently captured at least one dimension of "impact evaluation" and preceded or followed a coded sentence, then the edge-case sentence was also coded. The requirement for at least one sentence to be coded, on its own, is meant to reduce the potential uncertainty and lack of specificity introduced by allowing multiple sentences to collectively address "impact evaluation" when none can do so independently. The Gugerty et al. (2021) codebook served as a way to identify potential expressions of various dimensions of "impact evaluation." Additional detail on the coding protocol is available upon request and in the online data repository.

Upon review of all documents in the dataset, a new dataset was created consisting of 205 quotations[7] representing 42 documents and 37 document groups. As indicated in Table 3, some

[7] I will later use "quotation" and "document" interchangeably. While the unit of analysis captured during qualitative coding is a quotation from the text, each individual submission in a text corpus is referred to as a document, and traditionally they are composed of multiple sentences, paragraphs, or even pages.

documents and document groups in the source dataset were found to not have text that explicitly captured "impact evaluation." [8] A complete comparison of document groups in the source and final datasets by country and number of quotations is available in Appendix 1B: Data & Analysis.

Table 3. Comparison of source and final dataset.

	Gugerty et al. (2021)	**This paper**
Unit of analysis	By document, aggregated into document group	By sentence or independent clause
Number of documents	57	42
Number of document groups	42	37

Quotations were exported from Atlas.ti, and additional document group-level variables produced by Gugerty et al. (2021) were incorporated: organizational type (donor, infrastructure, or association), regional focus (national vs. international), sectoral origin (governmental vs. nongovernmental), and standard type (self-regulation vs. third party)[9]. The output was checked for spelling and export errors, as well as any duplicate text. For explicit examples and additional information, please refer to Appendix 1B: Data & Analysis, section Preparing Quotation Export.

Structural Topic Models

Analysis was conducted using structural topic models to ensure minimal overlap and maximum conceptual relevance of latent topics, or expressions of meta-rationalization. Traditional expressions of text analysis include document-frequency matrix or term-frequency matrix, in

[8] The document groups that were removed from this dataset are those for CharityNavigator, Cooperation Committee for Cambodia, Council on Foundations, European Foundation Centre, and Imagine Canada.
[9] I retain this ordering of variable values throughout the paper and visualizations to preserve the initial logic rather than impose a semi-arbitrary ordering, such as alphabetical or by word length.

which words that appear across a corpus (or collection) of documents are identified and transformed into matrices whereby each cell reports the presence, count, or ratio of a term within a document, in other words "reduc[ing] each document in the corpus to a vector of real numbers" (Blei et al., 2003, p. 993). Additional techniques look to further reduce the potential noise in such matrices to better identify critical information. One technique is using a probabilistic topic model (Blei, 2012). This unsupervised machine learning approach "uses patterns of word co-occurrences to discover latent themes across documents" (Roberts et al., 2016b, p. 1). Each latent theme, or topic, is "a mixture over words where each word has a probability of belonging to a topic. And a document is a mixture over topics, meaning that a single document can be composed of multiple topics" (Roberts et al., 2019, p. 2).

Latent Dirichlet allocation (LDA) is a common implementation of topic models that leverages statistical methodologies to shed light on statistical relationships, which in turn yields topics that are more meaningful and also allows for multiple topics to be observed in a single document (Blei et al., 2003). Structural topic models further build on topic models by allowing for covariates by document and topic, and they are commonly implemented using the "stm" R package (Roberts et al., 2019).

Topical prevalence covariates are metadata variables that allow researchers to control for and influence the document-topic relationship or the extent to which documents are related to topics, e.g., percent of a document as reflecting a topic. Similarly, topical content covariates influence the topic-word relationship or the extent to which different words are associated with topics. In this way, structural topic models allow for greater discrimination in both sets of relationships than is possible in traditional topic models such as LDA. Figure 2 presents an example, a subset of the output from a structured and unstructured topic model. The only difference in the two implemented

models is that the STM implementation include a topical prevalence covariate for document groups, or to which organization a standard belongs. Both models were implemented with $k = 10$ clusters; the "stm" support documentation notes that using 3 to 10 topics for short, specific corpora or collections of documents is usually sufficient to begin analyses (Roberts et al., 2019). For the full output of the two models, please refer to Figure 6 in Appendix 1B: Data & Analysis.

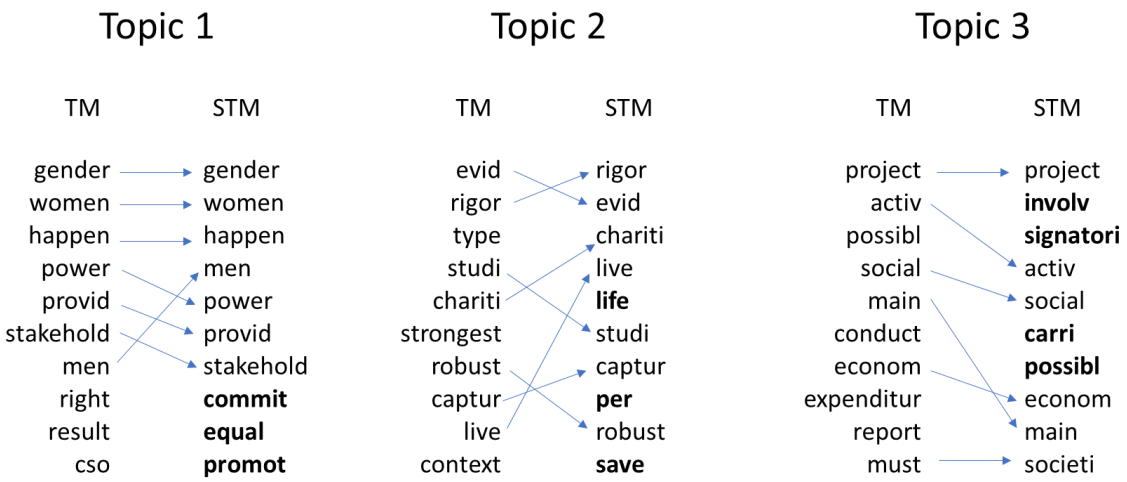

Figure 2. Comparison of two topic models, unstructured (TM) and structured (STM), with identical implementations but for the inclusion of the topical prevalence covariate "Document_groups" in the structured topic model. The model was run for k = 10 topics; only the first three topics are presented here, with a list of the 10 words with the highest score regarding its balance of frequency and exclusivity to that topic (in order from highest to lowest score). The arrows indicate where words changed placement from the TM to the STM. Words in bold only appear in that topic for the STM.

As evidenced in Figure 2, including even a single topical prevalence covariate to help define the document-topic relationship changed the fundamental composition of the three presented topics. For all topics, at least four words changed position in regards to FREX scores, a metric capturing a word's topic-specific frequency and exclusivity to one topic over others (Airoldi & Bischof,

2012; Bischof & Airoldi, 2012). All three topics had at least three novel additions to the topics in the STM implementation. Because the dataset is limited and the content is so specific, the use of topical prevalence and content covariates to further delineate the boundaries between topics becomes even more critical to inductively derive meaningful findings. Too much overlap between topics will reduce relevance, and so I use additional performance metrics to identify appropriate model composition: held-out likelihood, residual analysis, and semantic coherence. Similarly, appropriate model implementation becomes critical to best leverage the available data and identify latent constructs.

These latent constructs within the language used in standard-setting documents to refer to impact evaluation are expressions of meta-rationalization. Because they reflect expectations around the formalization of impact evaluation, and because impact evaluation can, in turn, formalize how NGOs deliver services, the latent constructs or topics generated by the models represent categories of meta-rationalization within the international development sector.

Exploring model parameter *k*

When using topic models, a critical first step is to justify choices for model parameters. The most pressing is that of the number of clusters, *k*, because this determines the number of latent constructs that can be perceived by a given model. Too small of a *k* suggests that we may be omitting constructs, but too many may impose a level of granularity or precision that is inappropriate for our dataset – we may end up with meaningless clusters and proposed constructs that are too narrow, too specific. Roberts et al. (2019) note that possible sizes of *k* will vary by the degree of specificity and size of the text corpus.

Preparing the analysis

I imported my corpus of quotations and their metadata into the statistical software R (R Core Team, 2021). After removing any blank observations from the corpus, I performed standard cleaning of the text data (see Appendix 1B: Data & Analysis for more detail) and converted relevant metadata values to factor values. Because of the small size of the dataset, I did not set a minimum threshold for the removal of documents, words, or tokens, e.g., drop from the dataset words appearing in fewer than 20 documents.

Selecting the number of topics *k*

To determine the appropriate number of *k* topics to specify in the model implementation, I used the "searchK()" function in the "stm" package to test the effects of a range of cluster sizes for four different models (Roberts et al., 2019); see Table 4 for a model overview. The range was chosen to build off of the previous guidance of using 3 to 10 topics for small, specific corpora. To allow for potential complexity resulting from the unique combinations of topic sizes and topical content and prevalence covariates, I utilized a broader range of 2 to 20. The four models varied in complexity: a naïve model with no covariates, i.e., unstructured topic model; a model with the full set of identified topical prevalence covariates; a model with the identified topical content covariate; and a model with both sets of covariates. All models utilized the same initialization type of "spectral," the default value for the "search()" function. The maximum number of expectation-maximization iterations was set to 300 after which, if the model has not converged, then the function will error out[10]. "Spectral" initialization refers to an implementation of using non-negative

[10] In general, the "full" and "prevalence" models (see Table 4) required more iterations to converge at smaller sizes of *k*. With this threshold, only one model failed to converge at 300 iterations, the "naïve" model for $k = 4$ topics. Given the overall performance of the "naïve" model relative to the others, I deemed this as negligible instead of running the models again at a higher threshold.

matrix factorization of word co-occurrence matrices to determine starting values for (or initialize) the model, and this approach has been found to produce reproducible models that do not suffer from issues of local or multiple modes as compared to other popular and effective initialization options (Mimno & Lee, 2014; Roberts et al., 2016b; Roberts et al., 2019).

Table 4. Comparative Summary of the Four Models Compared to Identify Topic Size

Model	**Naive**	**Prevalence**	**Content**	**Full**
Prevalence Variables	None	Document_Groups org_type national governmental std_type	None	Document_Groups org_type national governmental std_type
Content Variable	None	None	org_type	org_type
Topic size (range)	2 to 20			
Initial	Spectral			
Maximum number of iterations	300			

Identifying Topical Prevalence Covariates

The various topical covariates were selected based on the extent to which they would add explanatory power to the relationship between documents and topics (prevalence) and topics and words (content), respectively. Document groups ("Document_Groups"), or the organizations that produced the standards, are expected to control for numerous idiosyncratic organizational characteristics that could affect the extent to which a quotation reflects or contains various topics. I anticipate that other control variables in the dataset created by Gugerty et al. (2021) will have similar influences: organizational type ("org_type"), regional focus ("national"), sectoral origin ("governmental"), and standard type ("std_type").

Identifying Topical Content Covariates

I implement organizational type as a topical content variable because I expect it to influence the kinds of words that are associated with a given cluster. The organizational type indicates an organization (and thus a quotation)'s placement in the aid chain, i.e., "distance" from the ground. This may also align with awareness of local conditions and related factors that can influence the expectations set by the quotation-producing organization around available resources, staffing, and other constraints that may affect impact evaluation. The package "stm" is not able to handle more than one content variable due to computational complexity (Roberts et al., 2019), so the single variable was selected based on its anticipated high degree of explanatory power relative to the other possible metadata variables.

To better understand the potential effects of the sets of covariates as well as the optimal size of *k*, I explored the performance of the four models (see Table 4) for the range of 2 to 20 topics. I use four model performance measures to evaluate performance.

Identifying model performance & implementation

There are four evaluation measures of interest to compare performance within and across models; two that are common across unsupervised machine learning are held-out likelihood and an analysis of the dispersion of residuals. I use these two measures to help me determine which of the four models I should implement. Two commonly used performance measure for structural topic modeling are semantic coherence and frequency-exclusivity (FREX) scores (Roberts et al., 2019). Both of those face some challenges when a topical content covariate is introduced to the model, as discussed in the relevant appendices.

Of the configurations tested for the four models, the "full" model with both topical prevalence covariates and the topical content covariate has superior performance in regards to held-out likelihood and residual analysis. Within the full model, the implementation with $k = 6$ topics has the consistently highest semantic coherence values across topics. With the six topics generated from this model, I can now begin to explore how the impact evaluation standards establish different forms of meta-rationalization that set expectations for signatories' program implementations.

Results

To explore the content of each of the six topics, I focus on values reported by the function "sageLabels()", which sums across covariates to generate marginal scores and performance metrics by word as well as identifying the top n words or kappa for a given topic (Roberts et al., 2019). For each of the three levels of the content covariate "organizational type" (where 1 = donor, 2 = infrastructure, and 3 = association), there are also performance scores and related words available. I refer to them as needed to enhance topic exploration. Before identifying the latent constructs represented by topics, I build my understanding of each topic by identifying their explanatory power and representative documents for each.

How much do topics explain

Returning to the nature of structural topic models, "a topic is defined as a mixture over words where each word has a probability of belonging to a topic. And a document is a mixture over topics, meaning that a single document can be composed of multiple topics" (Roberts et al., 2019, p. 2). I explore the extent to which the six identified topics compose documents, on average, in Figure 3. As a visual reminder, it shows the first five words of the kappa for each topic. The x-axis reports the mean values for each column in theta, a matrix that reports the proportion of each

document *N* composed by topic *K* (Roberts et al., 2019). In other words, the expected proportion is the average proportion of documents captured by that topic. In terms of percentages we see that topic 5 explains, on average, ~30% of documents. Topics 3 and 4 explain ~20%, each, and topics 1, 2, and 6 are closer to ~10%.

To get a sense of how many documents each topic appears in, we can again leverage theta to create a histogram that shows the number of times proportion values for each topic appear (Figure 4). The largest bins containing the most values are for the smaller topic proportions, and the median values tend to be quite close to 0 for all topics except for topic 5. Almost all topics have their second-largest bins in the rightmost side of the histogram, as exemplified with topic 5. When we take into consideration that the largest bins for the other topics indicate proportions between 0 and 0.1 for at least 140 documents, there seems to be a clear emphasis on a few topics strongly representing some key documents, with the dominant topic across documents is topic 5.

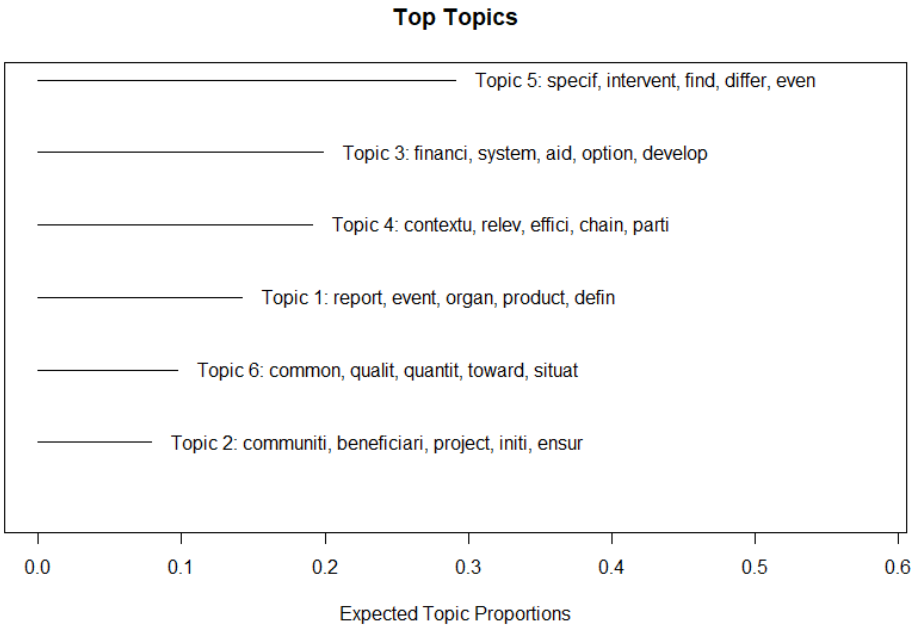

Figure 3. The top 5 words associated with each topic, or the kappa values of each topic for n =5. The x-axis shows the expected topic as the average percentage of documents captured by a given topic.

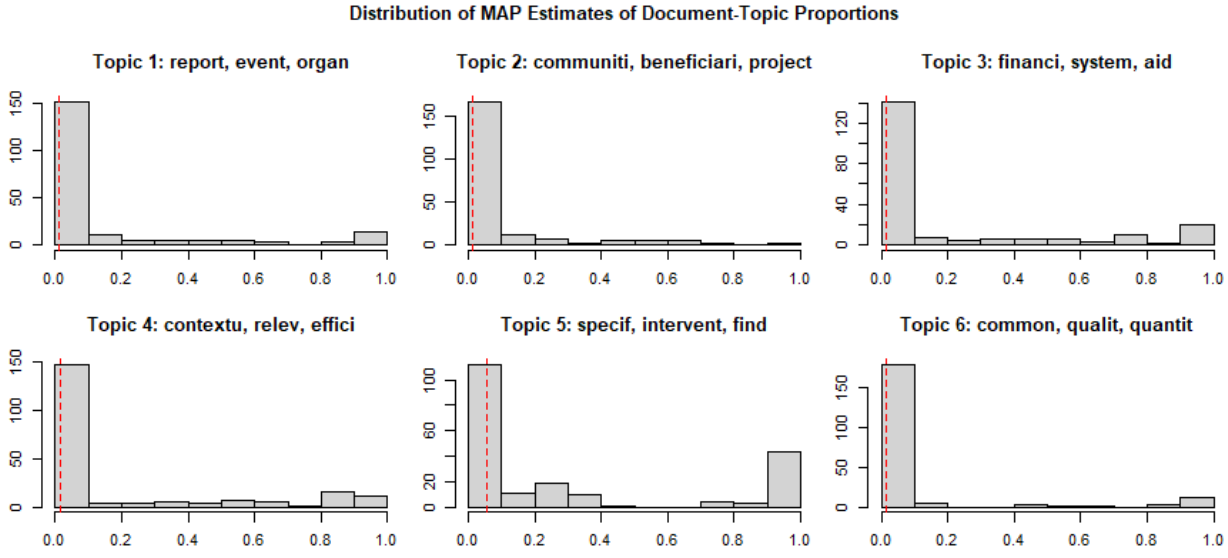

Figure 4. Histogram capturing the number of times a topic is a proportion of a document, for different proportion sizes. The dashed red (vertical) line reflects the median value of proportions for that topic (Roberts et al., 2019).

Representative documents

Table 5 provides summary details on each topic's top five highly representative documents: for a given topic, the five documents which are most dominated by or representative of that topic are identified and summarized. Each row thus reports information on those five highly representative documents in my corpus, which are quotations in the raw standards data. Table 9 in Appendix 1B: Data & Analysis includes quotations and related details for the top two representative documents.

I observe that some topics are dominated by organizations whereas others are more evenly mixed. Topic 6's set of representative documents (set) are dominated by Sphere's entry in its "Unpacked" series called "Sphere for Monitoring and Evaluation." Topic 5's set is dominated by the United Nation's Sustainable Development Goals (UNSDG) evaluation standards. Topics 3 and 4 represent a partial inversion of each other, in that 3's set is dominated 4:1 by the rating organization ImpactMatters over New Zealand's Council for International Development. Topic 4's set is dominated 4:1 by the United Nations Evaluation Group (UNEG) over the Philippine Council for NGO Certification (PCNC). Both sets share the same mix of organizational type, regional focus, and standard type, but are almost opposite when it comes to sectoral origin: Topic 4's set is almost all (inter)governmental whereas Topic 5's set is exclusively nongovernmental. Topics 1 and 2 are the most diverse sets in terms of the number of organizations reflected in the sets. Both topics' sets are predominantly self-regulatory standards and mostly associational organizations. Topic 1's set contains an infrastructure organization and all organizations are nongovernmental. Topic 2 is split 3:2 regarding associations to donors and nongovernmental to (inter)governmental organizations.

Table 5. Summary of top five highly representative documents grouped by topic.

Topic	Producing Organizations	Organizational Type	Regional Focus	Sectoral Origin	Standard Type
1	NACONGO (Tanzania): 2 BBB WGA: 1 InterAction: 2	Association: 4 Infrastructure: 1	National: all 5	Nongovernmental: all 5	Self-regulation: 4 Third party: 1
2	BOCONGO (Botswana): 1 DFID: 1 NDC (Palestine): 2 USAID: 1	Association: 3 Donor: 2	National: all 5	(Inter)Governmental: 2 Nongovernmental: 3	Self-regulation: 4 Third party: 1
3	Impact Matters: 4 CID (New Zealand): 1	Association: 1 Infrastructure: 4	International, cross-national: 4 National: 1	Nongovernmental: all 5	Self-regulation: 1 Third party: 4
4	UNEG: 4 Philippine Council for NGO Certification: 1	Association: 1 Infrastructure: 4	International, cross-national: 4 National: 1	(Inter)Governmental: 4 Nongovernmental: 1	Self-regulation: 1 Third party: 4
5	UN SDG: all 5	Infrastructure: all 5	International, cross-national: all	(Inter)Governmental: all 5	Third party: all 5
6	Sphere: all 5	Infrastructure: all 5	International, cross-national: all	Nongovernmental: all 5	Third party: all 5

Identified latent constructs

In reviewing the kappa for n = 10 of each topic (Table 6) and related metrics, the relatively high exclusivity of terms allows for a straightforward synthesis of labels. For comparison, Table 8 in Appendix 1B: Data & Analysis reports the output for topic 1 that I analyzed; the other five topics had similar outputs. I use the document-topic information, words highly associated with marginal performance metrics, and representative documents in this process.

Table 6. Summary of top (<=10) highly representative words for each topic.

Topic	**1**	**2**	**3**	**4**	**5**	**6**
Proposed Label	*Compliance*	*Community Benefits*	*Establishing Systems*	*Developing Understanding*	*Creating Change*	*Data Engagement*
	report	communiti	financi	contextu	specif	common
	event	beneficiari	system	relev	intervent	qualit
	organ	project	aid	effici	find	quantit
	product	initi	option	chain	differ	toward
	defin	ensur	develop	parti	even	situat
	expenditur		will	result	two	adapt
	reason		activ	critic	theori	intern
	inform		measur	institut	analysi	
	servic		monitor	extent	chang	
	period		whose	recommend	case	

Topic 1. The language focuses on reporting, defining, organizing, and the assertive word "shall." The set of representative documents (set) discussed earlier was highly diverse in regards to contributing organizations, but all focused nationally, all were NGOs, and almost all were from self-regulatory standards. The set consisted of explicit language around establishing systems for communication and producing information, settings goals and objectives, creating reports and budgets, and comparing sets of data. Given the highly representative terms and linguistic construction of the set, I propose the label of "**Compliance**."

Topic 2. The most representative words across metrics focused on community, projects, beneficiaries, needs, and NGOs. The set was highly diverse in regards to contributing organizations and almost split evenly in terms of organizational type and sectoral origin. All focused nationally and almost all were from self-regulatory standards. The set consisted of awareness around moral responsibilities to the community, understanding beneficiaries vs. non-beneficiaries, and community composition, needs, and participation throughout the project

lifecycle. I propose the label of "**Community Benefits**" to capture the linguistic orientation around community needs and benefits that are derived from programming.

Topic 3. The words are technical in nature, primarily focusing on concepts related to MEAL (monitoring, evaluation, accountability, & learning) and similar concepts as noted in the Gugerty et al. (2021) codebook. Example stemmed terms include "measur," "monitor," "evalu" (ex. evaluate, evaluating), and "program." Some of the most frequent terms that appear alongside the MEAL-related words across performance metrics include "develop" (ex. development), "financi" (ex. financial or financing), "system", "activ" (ex. active or activity), and "aid" are. Documents in the set were dominated by an infrastructure organization imposing cross-national third-party norms around the process of impact evaluation. There was also an associational NGO establishing guidance for its member organizations. The set focused on establishing evidentiary support in regards to intervention performance, making progress towards goals & objectives, monitoring the relationship between beneficiaries and traditional logic model elements such as activities, outputs, and outcomes, and using information from MEAL-type processes to improve how aid is delivered. I summarize these various elements as establishing systems for evidence creation and use, which I summarize as the proposed label, "**Establishing Systems**."

Topic 4. Terms that consistently appear across performance metrics include "result," evalu," "chang" (ex. change or changing), "relev" (ex. relevant), "contextu" (ex. contextualize). After reviewing them, the words in the kappa seem easier to synthesize; there is a narrative that can be constructed around the need for context and relevant information or outcomes related to perceptions of efficiency, participation, etc. An inter-governmental organization dominates the set, one that is establishing a globally oriented infrastructure for impact evaluation. Like Topic 3, there was also an associational NGO establishing guidance for its member organizations. The set is

oriented around understanding program results and their broader societal effects through examining intervention steps and criteria, how to do this, the importance of engaging stakeholders, and the importance of time passing to determine if effects were positive or not. Each of these elements underscores a common theme of understanding interventions and their effects, which are affected by or require many of the terms mentioned earlier. Given this information, I propose the label "**Developing Understanding**."

Topic 5. As with topic 4, reviewing the array of terms across performance metrics gives insight into the connective construct underlying the terms in the kappa. Among the most common terms across metrics are "impact," "chang," "intervent" (ex. intervention), and "evalu." The terms in the kappa suggest a focus on understanding how even similar programs and interventions can differ, and doing so via theories and analysis. This set consists of documents from a single organization, an internationally oriented intergovernmental organization that establishes third-party infrastructure norms around impact evaluation. Common themes within the set include understanding the relationship between gender, power, and interventions to create sustainable change, the challenges in understanding the effects of multi-stakeholder interventions, and understanding effects of interventions on stakeholders. What comes across to me is a need for awareness of intersectional elements, such as social dynamics, relationships, and human desires and needs with the intended vs. actual outcomes of interventions. I synthesize this as developing awareness of the extent to which interventions create change in multifaceted ways, which I shorten to the proposed label "**Creating Change**."

Topic 6. The terms listed in the kappa seem to revolve around kinds of data and how it is used. When looking at the performance metrics, common terms include "countri" (ex. countries), "random," "monitor," and "chang." There is a more pronounced difference between the terms in

the kappa and the performance metrics for this topic than for other terms, and I look to the set of representative documents to shed some light. This set consists of documents from a single source, an internationally oriented nongovernmental organization that establishes third-party standards around multiple dimensions of international development work. One such standard relates to impact evaluation, which produced the documents in my corpus; it is worth noting that this standard is one of several geared less to leadership of implementing organizations and more to practitioners, e.g., of monitoring and evaluation. Common themes within the set include measuring change through collecting feedback and using progress indicators to create data, encouraging awareness of limitations of data's explanatory capacity, and understanding how situations change before and after an intervention. What seems consistent across the themes is the importance to not just collect information and data but to be cognizant of limitations when using it. This notion aligns both with the concepts inferred from the kappa and the presence of common terms from the performance metrics. I propose the label "**Data Engagement**."

Interpreting the process and output

The structural topic model that I use equips us with evidence to better understand rationalization and trends in international development that impose structures on and requirements for impact evaluation.

The topical content covariate allows me to consider how the organizational type affects the kind of language that is used to formalize impact evaluation. The topical prevalence covariates provide insight on the relationship between standards and the organizations that produce them in regards to what kind of organization they are and their regional focus, relationship to governmental entities, and relationship with implementing, on-the-ground organizations. The model takes an unsupervised machine learning approach to identify clusters of words that suggest constructs latent

within the standard-setting documents that I coded. I then analyzed each cluster's most representative words along multiple dimensions and quotations, and I proposed labels for each.

The proposed labels of, Compliance, Community Benefits, Establishing Systems, Developing Understanding, Creating Change, and Data Engagement reflect constructs latent in the rationalization[11] of impact evaluation.[12] I previously identified that part of the fundamental intent of impact evaluation is to shape the processes that yield changes in the world. In this way, it acts as a form of rationalization because it formalizes processes in that it reduces, or at least changes, permutations of steps. The six topics are categories of how organizations intend for their standards to formalize impact evaluation processes in other organizations. This means that each of the six topics reflect categories of meta-rationalization in the international development context.

The analytic lens brought to bear on international development can shape the utility, applicability, and generalizability of the implied effects of the six topics for implementing NGOs. Lack of dominant topics within the same organizational type would suggest that an NGO trying to sign on to standards from multiple such organizations would face greater heterogeneity of resource demands, risking their ability to be compliant. Depending on the kind of organization, this risk could have consequences for their ability to acquire future resources and perform their work. Alignment of topic prevalence between national and inter- or cross-national organizations' standards could also indicate whether it is realistic for signatories to strive for compliance with standards from both focuses, or if it would be more cost-effective to limit themselves to only one. The sectoral origin of the standard-setting organization identifies the potential for coercive

[11] Which I define as the formalization of core (service delivery) and support (management) processes within a nonprofit.

[12] Which I define as a systematic process to identify the relationship between programs and their effects through collecting evidence and making claims about the relationship between program-specific activities and observed changes in the world

authority by the standard-setting organization. It also captures whether compliant signatories will be able to perform impact evaluation in alignment with their existing accountabilities as compared to taking on new lines of accountability, noting that each stakeholder can have its own preferred goal for impact evaluation. Similarly, the standard type sheds light on the dominant perspective on impact evaluation for standards that are signed on to solely for signaling purposes and with potentially more rigorous verification methods, as compared to standards that may have less rigorous verification methods but may also provide more to NGOs than solely acting as signaling mechanisms.

The more frequently a construct appears in the data, the greater its prevalence. The more prevalent a topic is within standards, the more it becomes the dominant expression of how impact evaluation should be conducted in international development. Regardless of what a signatory believes is the purpose of impact evaluation, the most prevalent topics of meta-rationalization would have to be addressed and incorporated into its processes. The four analytic perspectives represented by the covariates provide a sense of how that imposition of purpose onto signatories can vary by: standard-setting organizations' location (and thus power and influence) in the development aid chain; a global vs. domestic focus; influences from the standard-setting organizations' sectors; and whether the standard is created by self-regulatory, membership organizations for themselves and future members or by third-party actors who are, to an extent, external and outside of the production cycle.

Ultimately, the prevalence of each of these topics provides insights into the dominance of certain global trends and how different portions of the sector work to replicate and reinforce those priorities. They also shed insights into the explicit and implied requirements established for implementing organizations by the standards.

Sectoral Implications by Topic

When examining expected topic prevalence by covariate, some topics are strongly present and often dominant (Creating Change), whereas others are weakly present or never dominant (Compliance, Community Benefits, Data Engagement). For each topic, in order of most to least prevalent across topics, I explore the insights that the findings provide visually (Figure 5) and looking at the raw data (Appendix, Table 10).

Regardless of covariate, Creating Change (Topic 5) is within the 3 most highly prevalent topics. The only exception is for self-regulatory standards, where Creating Change was slightly more prevalent than Community Benefits, but was still fourth out of six. This suggests that most organizational types and sectors emphasize it similarly and thus ascribe it the same value; I expect little disagreement when considering standards from those perspectives. It is less emphasized among nationally-focused and self-regulatory standards, suggesting that organizations focused on national needs and establishing norms for their members do not put as much emphasis on understanding the various ways that interventions can effect change.

The urge to develop an understanding of interventions and their effects (Topic 4) is consistently in the middle of range, sometimes among the most prevalent topics, but never the least prevalent. Its heavy emphasis by donors (at an expected prevalence of 0.347 across documents) is not nearly as surprising as how rarely prevalent it is among infrastructure organizations. Donor standards act as vehicles for the cultural norms and expectations of the country and community of origin, and they are also accountable to various funders and funding sources. A recipient's capacity regarding Topic 4 is thus beneficial to them, and they often wield a coercive authority stemming from the threat of punishment for poor – or promise of reward for good –understanding of intervention effects. In contrast, the infrastructure organizations seem much less focused on promoting this sort

of development among implementing organizations that they influence. The difference between self-regulatory standards and third-party standards is almost as sharp; that said, it is almost tied at second-most prevalent topic for third party standards (along with Establishing Systems and Data Engagement). This suggest less of a difference in value and importance of Developing Understanding and more that, for third-party standards, topics 3, 4, and 6 are equally important (or, perhaps, unimportant when compared to Creating Change).

Establishing systems for evidence creation and use (Topic 3) is also consistently in the middle range of prevalence among documents. Where infrastructure and association organizations feature it in over 20% (0.208 and 0.227, respectively) of documents, donors only feature it in 8%. It is also sharply de-emphasized by intergovernmental organizations, despite their high emphasis on both Developing Understanding and Creating Change. Another jarring comparison is the high prevalence by self-regulatory standards on establishing systems but low prevalence of data engagement; where the two topics seem clearly connected at first read, Establishing Systems tends to be low for organizations that also have a low prevalence of Data Management.

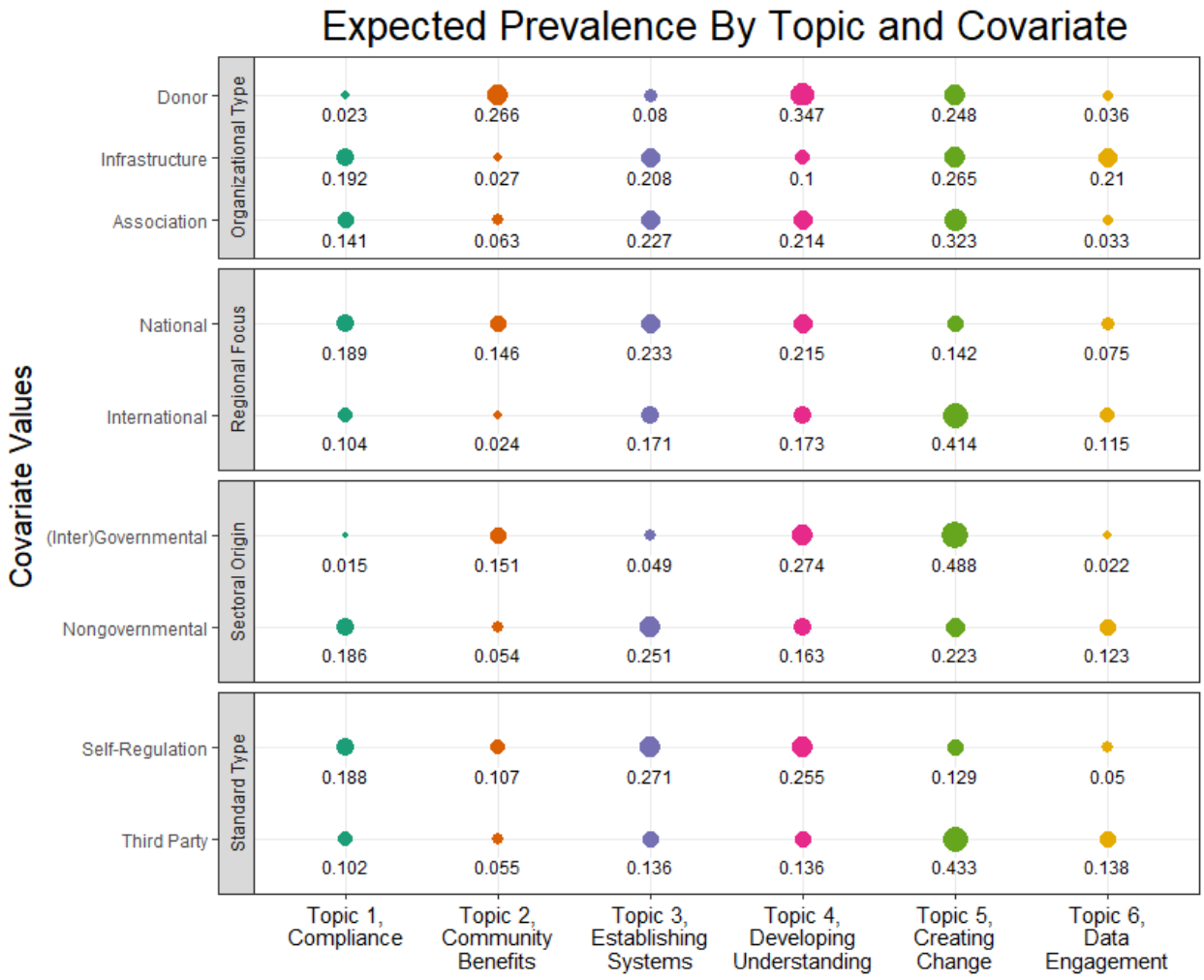

Figure 5. Comparison of expected (mean) topic prevalence, rounded, by topic and by topic prevalence covariate. Color is unique to that column's respective topic (x-axis). Each dot captures the mean of topic prevalence for that covariate (row) and topic (column). Each row's proportions sum to 1. The covariate name is in the vertical row label. Each bubble is labeled with the respective rounded proportion.

Compliance with specified impact evaluation practices (Topic 1) is never highly prevalent within covariate types (at most, third-most prevalent), and it varies noticeably within covariate values. Donors least discuss it out of all topics, in noticeable contrast to infrastructure and associational organizations at a difference of 16.9 and 11.8 percentage points, respectively. The similar sharp

contrast is noticeably between nongovernmental and (inter)governmental with a difference of 0.0171 or 17.1 percentage points. This suggests that donors and intergovernmental organizations are not explicitly stating that compliance must be met, but instead may be offering guidance or best practices that are not explicitly mandated. The difference between regional focus and standard type values was about 8 percentage points; while this notes a clear and explicit difference, it is less than half of the difference observed within the other covariates.

Awareness of community composition, needs, and received/perceived benefits throughout impact evaluation (Topic 2), is consistently among the least prevalent topics. Noticeable exceptions are for donor organizations, where the topic ranks as second most prevalent; the topic is far and away the most prevalent for donors as compared to any other covariate value. This may again relate to the need for donors to create stories for their own funders, constituents, etc., and understanding community benefits from funded interventions may help donors' efforts to both humanize international development and appeal to their public.

Data Engagement (Topic 6) is always within the 3 least prevalent topics, apart from infrastructure organizations and third-party standards, where it was the second-most prevalent topic. The collection of data and understanding its limits varies within covariates, noticeable only in this scope of comparison but not when considering the other topics. The intra-covariate differences are relatively small, with the largest being between donor and association organizations (low value) versus infrastructure organizations, whose emphasis on data engagement may stem from their own need for that data to impose external rankings.

Discussion

Analysis of the six topics as categories of meta-rationalization presents various practical implications for NGOs and theoretical implications for applying STM to this novel theoretical construct. The ways in which standards act as expectation-setting vehicles for NGOs interact for each other in ways that yield important insights as to how much resource diffusion signatories may need to overcome to be compliant with multiple expressions of meta-rationalization.

Generalized findings

In looking at the broad spectrum of findings within and across topics and covariates, some interesting relationships emerge that suggest trends within and disagreement among cross-sections of the sector.

My findings suggest that organizations that do not emphasize a technical topic such as compliance, data management, or establishing systems are likely to not emphasize one or both of the others. While this may seem negative, a lack of explicit requirements may allow organizations the freedom to be creative and resource-aware in addressing those aspects of their impact evaluations, for example by drawing on data sources that a donor or intergovernmental organization may not be aware of or consider as legitimate. On the other hand, a lack of explicit guidance may limit organizations' capacities to effectively achieve the desired outcome, for example, meeting compliance requirements in a way that also allows them to establish effective systems and create and manage data; impact evaluation quality and quantity may vary widely. This may not be an issue for the purposes of internal MEAL practices among implementing organizations, but would present challenges for cross-intervention analyses and developing a broader understanding of the sector.

The topic of Community Benefits is consistently not prevalent: there is little emphasis across covariates on ensuring that impact evaluations yield understanding of community benefits. While this may, again, be positive in that it avoids imposing restrictive and resource-demanding protocols, the sharp contrast between donors and the other configurations of organizations also suggests that there is some sort of institutional pressure regarding intervention benefits for target communities that is affecting donors and is causing the concept to be so highly prevalent across their standards. The topic is also noticeably mentioned by intergovernmental organizations, self-regulatory standards, and nationally focused organizations. Self-regulatory and national orientation makes sense because these standards tend to be produced by organizations "lower" on the development aid chain, i.e., closer to the ground and direct implementation or interaction with target communities. Inter-governmental organizations may emphasize this topic through a combination of pressures from members or key stakeholders as well as practicing MEAL, in particular the "learning" dimension and integrating evaluation findings, academic research, and evolving sectoral practices into their own standards.

Potential implications for on-the-ground direct implementation organizations can be determined through both the presence and the absence of topic prevalence. Imposing and specifying explicit guidance with high degree of prevalence is a signal of what that standard and organization wants signatories to pursue. It gives implementing organizations clear goals to achieve or targets to meet. Lack of specificity, as suggested by the lack of topic prevalence, may be desirable by implementing organizations because it frees them from constraints imposed by standard-setting organizations that are unfamiliar with their specific operating contexts. It also has the potential to reduce accountability because adherence to standards regarding impact evaluation processes becomes that much harder to evaluate across signatories systematically and consistently.

Impact evaluation is, at its core, an effort to understand change. It is, therefore, not surprising that the topic of Creating Change is the most dominant across and within covariates. No matter how I have distinguished organizations and their standards from each other, the need to understand how interventions yield change is consistently more emphasized than any technical or human-experience-oriented topic. This sharp emphasis is to be expected, so when there is a large gap between covariate values (ex. regional focus, standard type), this suggests that there are varying opinions about what impact evaluation is meant to achieve, at its core, and hints at a fundamental disagreement among some cross-sections of the international development sector.

Theoretical Implications

The results demonstrate that meta-rationalization is clearly observable amongst the standards documents. Meta-rationalization is proposed as the rationalization of rationalization, the formalization of core and support processes that, themselves formalize core and support processes. Impact evaluation standards impose formalizing structures on a series of systematic processes. By using structural topic models to account for various document-level characteristics as well as potential influence on topic construction, I generated meaningful clusters of words as topics that provided insights into a sector, international development, that align with and expand on existing literature (ex. Gugerty et al., 2021). The generated topics are within the expected bounds of meta-rationalization expressions in this sector. This suggests that structural topic models are a viable method to use for identifying meta-rationalization expressions and their prevalence across sectors. The use of a sector-specific measure as a topical content covariate to define topic-word relationships contributed to a deeper meaning and sharper delineation between meta-rationalization expressions. Not all expressions of meta-rationalization are equally present. How those expressions are observed and categorized are likely to vary by the analyzed sector and the

documents or standards that rationalize it. An STM implementation with analyzed documents and document-level variables that are relevant to the analyzed sector should yield similar useful results for analysis.

Conclusion

The variation in impact evaluation expectations established by standard-setting documents displays how signatory organizations may face competing resource demands to be compliant. Whereas almost all standards prominently feature the meta-rationalization categories of creating change and understanding interventions and their effects, there is a lack of consistent presence of the other categories across the standards. This suggests that, for a signatory signed on to multiple standards, they would have to either invest in a wide array of support processes around impact evaluation or selectively invest in a way that may limit their compliance, their ability to be accountable, and how they formalize core, mission-related service delivery. Either option suggests sub-optimal compliance and limited utility of the standards to achieve their intended goals. Signatory organizations may be forced to opt out of or hesitate to opt into multiple standards if there is too much variation in the meta-rationalization categories and related resource demands. Alternatively, if an organization perceives a standard as necessary to gaining legitimacy, signaling accountability, or accessing future resources, then they may divert current resources towards compliance and away from service delivery and mission-related processes. Returning to Figure , the organization may not be able to be fully compliant with one or more standards because of conflicts from overlap around which steps are required, which required steps are possible for that organization to perform, and their own internal commitments to maintaining certain orders of steps. Any ensuing noncompliance could have significant internal, relational, and reputational consequences. For this reason, NGOs should take on a critical perspective when considering

standards to adopt, and standard-setting organizations should review their documents to identify potential overlaps or synergies with related standards from other organizations, potential sources of friction and alternative options, or greater flexibility for signatories as to which elements of a standard will be signed on to rather than the seemingly ubiquitous all-or-nothing approach.

These examples highlight the need for research focused on the relationship between impact evaluation, rationalization, and meta-rationalization. This need includes a greater understanding of resource acquisition in response to those pressures, the effect of impact evaluation requirements on organizations with varying degrees of rationalization, and the potential benefits for impact evaluation from meta-rationalization. Motivations for increased rationalization may indicate possible resource diversions and signal trends in the activities around (i.e., processes) and resources required (e.g., staff) for impact evaluation.

Potential limitations

This paper explores meta-rationalization triggered by external forces and does not consider expressions that arise internally to an organization. In a sufficiently complex organization, where various units act with sufficient autonomy and independence to formalize their processes, it may be possible to observe meta-rationalization as imposed across hierarchical levels or between units, rather than solely from external mechanisms. Future studies should work to understand the conflicting dynamics and competitive resource demands that arise from meta-rationalization imposed from multiple external sources, intra-organizational sources, and combinations of the two. In addition, the order in which an organization undergoes or experiences meta-rationalization may have its own resource costs and may be controllable by organizations. Such observations would help to establish the presence or lack of a bi-directional power dynamic between process implementers and meta-rationalization sources. Finally, the relationship between rationalization

and meta-rationalization should be explored and tested across a range of sectors and industries, which will shed further insights on how to mitigate expected resource demands due to both.

With the model, corpus size is a key factor. While a larger corpus of documents is always desirable, intentional model design and thorough testing can still yield meaningful outcomes when analyzing a corpus that is small but specified (ex. international development standard-setting documents). In addition, the corpus is composed of documents that reflect a one-way relationship, of standard-setting entities towards standard-complying entities, i.e., signatory organizations. Future work would benefit from exploring how prevalent the topics identified here are among signatory organizations' standard procedures for impact evaluation, identifying what topics are latent within their documented approaches to impact evaluation, and comparing the degree of intersection between the categories of meta-rationalization and rationalization.

The Council of Finnish Foundations notes that "impact evaluation methods used in one field do not necessarily apply to all fields" (Suvikumpu et al., 2015, p.5). The definition of impact evaluation used in this paper may be biased towards certain types of organizations, e.g., implementing organizations, which may exclude how other organizations think about the formalization of impact evaluation. It may also suffer from generalizability to other sectors that engage in impact evaluation. In addition, it focuses on programs, projects, initiatives, and activities within the organization, not organization-level impact. Disaggregation of impact by program-level (as compared to aggregation at the organizational level) is an explicit component of the proposed definition as it compensates for diminished clarity by organizations with multiple streams of work that may have limited overlap and similarity. In addition, the coding focused on text with explicit alignment to the definition; alignment that is implicitly obvious for those with specialized subject

matter expertise was not considered as grounds for coding. Some scholars may disagree with these approaches; they were chosen to improve replicability through increased specificity.

Of note is that regional distinctions within the same language, such as English from the United Kingdom and from the United States, may affect the performance of topic modeling techniques. “Program” and “programme” were not converted into the same term by the default stemmer in the package “stm,” which relies on the Snowball stemming tool (Porter, 2001). Snowball is now virtually ubiquitous: it has been heavily adopted by and implemented in libraries for the most prevalent scientific and open-source computing tools, including R and Python. What this suggests is a need for further awareness in understanding how regional differences in word choice and construction may not be appropriately addressed by the default implementations of text cleaning tools such as the Snowball stemmer. This need is even more prevalent for languages with more variation in vocabulary and word construction than English.

References

AbouAssi, K., & Bies, A. (2018). Relationships and resources: the isomorphism of nonprofit organizations'(NPO) self-regulation. *Public Management Review*, *20*(11), 1581-1601.

Airoldi, E. M., & Bischof, J. M. (2012). A Poisson convolution model for characterizing topical content with word frequency and exclusivity. *arXiv preprint arXiv:1206.4631*.

Bernard, T., Delarue, J., & Naudet, J. D. (2012). Impact evaluations: a tool for accountability? Lessons from experience at Agence Française de Développement. *Journal of Development Effectiveness*, *4*(2), 314-327.

Bingham, R., & Felbinger, C. (2002). *Evaluation in practice: A methodological approach* (2nd ed.). CQ Press.

Bischof, J., & Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* (pp. 201-208).

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Bromley, P., & Orchard, C. D. (2016). Managed morality: The rise of professional codes of conduct in the US nonprofit sector. *Nonprofit and Voluntary Sector Quarterly*, *45*(2), 351-374.

Brown, L. D., & Moore, M. H. (2001). Accountability, strategy, and international nongovernmental organizations. *Nonprofit and voluntary sector quarterly*, *30*(3), 569-587.

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological methods & research*, *42*(3), 294-320.

Castro, M. P., Fragapane, S., & Rinaldi, F. M. (2016). Professionalization and evaluation: A European analysis in the digital era. *Evaluation*, *22*(4), 489-507.

Cumming, G. D. (2008). French NGOs in the global era: Professionalization "without borders"?. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, *19*(4), 372-394.

Dart, R. (2004). Being "business-like" in a nonprofit organization: A grounded and inductive typology. *Nonprofit and voluntary sector quarterly*, *33*(2), 290-310.

Deloffre, M. Z. (2016). Global accountability communities: NGO self-regulation in the humanitarian sector. *Review of international studies*, *42*(4), 724-747.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American sociological review*, 147-160.

Ebrahim, A. (2003). Accountability in practice: Mechanisms for NGOs. *World development*, *31*(5), 813-829.

Eisenstein, J., Ahmed, A., & Xing, E. P. (2011). Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 1041-1048).

Epstein, D., & Klerman, J. A. (2012). When is a program ready for rigorous impact evaluation? The role of a falsifiable logic model. *Evaluation Review*, *36*(5), 375-401.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2016). *Impact evaluation in practice*. The World Bank.

Gibson, M. & Sautmann, A. (2021, April). *Introduction to randomized evaluations*. J-PAL: Abdul Latif Jameel Poverty Action Lab. https://www.povertyactionlab.org/resource/introduction-randomized-evaluations

Gugerty, M. K. (2009). Signaling virtue: Voluntary accountability programs among nonprofit organizations. *Policy Sciences*, *42*(3), 243-273.

Gugerty, M. K., Mitchell, G. E., & Santamarina, F. J. (2021). Discourses of evaluation: Institutional logics and organizational practices among international development agencies. *World Development*, *146*, 105596.

Hvenmark, J. (2016). Ideology, practice, and process? A review of the concept of managerialism in civil society studies. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, *27*(6), 2833-2859.

Hwang, H., & Powell, W. W. (2009). The rationalization of charity: The influences of professionalism in the nonprofit sector. *Administrative science quarterly*, *54*(2), 268-298.

Innovations for Poverty Action. (n.d.-a) *What Do We Mean By Impact?* https://www.poverty-action.org/impact/what-do-we-mean-impact

Innovations for Poverty Action. (n.d.-b) *What We Do.* https://www.poverty-action.org/about/what-we-do

International Initiative for Impact Evaluation. (n.d.). *Impact evaluation*. https://www.3ieimpact.org/What-we-offer/impact-evaluation

Johnson, S., & Rasulova, S. (2016). Qualitative impact evaluation: Incorporating authenticity into the assessment of rigour.

Kelly, L. M. (2021). *Evaluation in small development non-profits: Deadends, victories, and alternative routes*. Springer International Publishing.

Luchner, C. D. (2018). Contact zones of the aid chain: The multilingual practices of two Swiss development NGOs. *Translation Spaces*, *7*(1), 44-64.

Maier, F., Meyer, M., & Steinbereithner, M. (2016). Nonprofit organizations becoming business-like: A systematic review. *Nonprofit and Voluntary Sector Quarterly*, *45*(1), 64-86.

Mimno, D., & Lee, M. (2014, October). Low-dimensional embeddings for interpretable anchor-based topic inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1319-1328).

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262-272).

Mitchell, G. E. (2018). Modalities of managerialism: The "double bind" of normative and instrumental nonprofit management imperatives. *Administration & Society*, *50*(7), 1037-1068.

Murtaza, N. (2012). Putting the lasts first: The case for community-focused and peer-managed NGO accountability mechanisms. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, *23*(1), 109-125.

Podems, D. (2014). Evaluator competencies and professionalizing the field: Where are we now?. *Canadian Journal of Program Evaluation*, *28*(3).

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

Prakash, A., & Gugerty, M. K. (2010). Trust but verify? Voluntary regulation programs in the nonprofit sector. *Regulation & Governance*, *4*(1), 22-47.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Raina, R. S. (1999). Professionalization and evaluation: The case of Indian agricultural research. *Knowledge, Technology & Policy*, *11*(4), 69-96.

Ravallion, M. (2001). The mystery of the vanishing benefits: An introduction to impact evaluation. *the world bank economic review*, *15*(1), 115-140.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016a). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988-1003.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016b). Navigating the local modes of big data. *Computational social science*, *51*.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, *58*(4), 1064-1082.

Roberts, M.E., Stewart, B.M., & Tingley, D. (2019). "stm: An R Package for Structural Topic Models." *Journal of Statistical Software*, **91**(2), 1–40. doi: 10.18637/jss.v091.i02.

Roberts, S. M., Jones Iii, J. P., & Fröhling, O. (2005). NGOs and the globalization of managerialism: A research framework. *World development*, *33*(11), 1845-1864.

Roche, C. (2000). Impact assessment: Seeing the wood and the trees. *Development in Practice*, *10*(3-4), 543-555.

Rossi, P., Lipsey, Mark W, & Freeman, Howard E. (2004). *Evaluation : A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Salamon, L., & Anheier, H. (1992). In search of the non-profit sector. I: The question of definitions. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *3*(2), 125-151. https://doi.org/10.1007/BF01397770

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., ... Jinks, C. (2018). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity*, 52(4), 1893–1907.

Savedoff, W. D., Levine, R., & Birdsall, N. (2006). *When will we ever learn?: Improving lives through impact evaluation*. Center for Global Development. https://www.cgdev.org/sites/default/files/7973_file_WillWeEverLearn.pdf

Shaffer, P. (2011). Against excessive rhetoric in impact assessment: overstating the case for randomised controlled experiments. *Journal of Development Studies, 47*(11), 1619-1635.

Shah, N. B., Wang, P., Fraker, A., & Gastfriend, D. (2015). Evaluations with impact. *Decision-Focused Impact Evaluation as a Practical Policymaking Tool (3ie Working Paper 25).* International Initiative for Impact Evaluation, New Delhi.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). Broadening the range of designs and methods for impact evaluations. London: DFID. http://www.dfid.gov.uk/R4D/Output/189575/Default.aspx

Suárez, D. F., & Gugerty, M. K. (2016). Funding civil society? Bilateral government support for development NGOs. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, *27*(6), 2617-2640.

Suárez, D. F., & Hwang, H. (2013). Resource constraints or cultural conformity? Nonprofit relationships with businesses. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *24*(3), 581-605.

Suvikumpu, L., Tikka, P., & Saukkonen, P. (2015). *A Foundation With Impact! Principles and practices of evaluating the impact of foundations*. Helsinki, Finland: Council on Finish Foundations.

Taddy, M. (2012, March). On estimation and selection for topic models. In *Artificial Intelligence and Statistics* (pp. 1184-1193). PMLR.

Tremblay-Boire, J., Prakash, A., & Gugerty, M. K. (2016). Regulation by reputation: Monitoring and sanctioning in nonprofit accountability clubs. *Public Administration Review*, *76*(5), 712–722. https://doi.org/10.1111/puar.12539.

United Nations Evaluation Group. (2016). Norms and Standards for Evaluation. New York: UNEG.

Van Hemelrijck, A., & Guijt, I. (2016). Balancing inclusiveness, rigour and feasibility: Insights from participatory impact evaluations in Ghana and Vietnam.

W. K. Kellogg Foundation. (2004). Logic model development guide: Using logic models to bring together planning, evaluation, and action. Battle Creek, MI: Author.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009, June). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1105-1112).

Appendices

Appendix 1A: Impact Evaluation

RCTs and Counterfactuals

"Impact evaluations are designed to answer the question: 'What was the effect of an intervention on an outcome?'" (International Initiative for Impact Evaluation, n.d.). While this question can be answered using a multitude of methods, RCTs and counterfactuals as methods dominate this conceptual space. Randomized evaluations are a type of impact evaluation that uses treatment and control (or comparison) groups to establish counterfactuals for causal attribution of program (intervention) effects and outcomes (Bernard, Delarue, & Naudet, 2012; Gertler et al., 2016; Gibson & Sautmann, 2021; Ravallion, 2001; Shah et al., 2015). Counterfactuals are pursued not solely via RCTs because they provide an observable, measurable, and testable basis for comparing how program recipients (treatment group) and non-recipients (control group) fared during and after implementation, which provides a positivist determination of a program's effect on outcomes, or its impact. Some authors place emphasis on the need for treatment and control groups regardless of whether group assignment is random (see Gertler et al., 2016; Gibson & Sautmann, 2021). Once produced, counterfactuals and causal attributions are used to explore program effectiveness and theories of change (Gibson & Sautmann, 2021).

Across organizations and authors, there tends to be an explicit focus on quantitative data for these kinds of impact evaluations (see Gibson & Sautmann, 2021). That said, mixed-methods studies that use quantitative and qualitative are also seen as viable impact evaluations (Shah et al., 2015). More broadly, Ravallion (2001) notes that "evaluation is essentially a problem of missing data" (p. 137); they and others identify the utility of alternative methods to construct or analyze

counterfactuals that compensate for missing data and allow for causal attribution, such as quasi-experimental designs (International Initiative for Impact Evaluation, n.d.; Ravallion, 2001).

Some see RCTs as but one step of impact evaluation, and impact evaluation as but one part of implementation. Bernard, Delarue, & Naudet (2012) note that impact evaluation can serve to create a feedback learning mechanism among pilot or small-scale programs before scaling up interventions to the level of policies. Similarly, the group Innovations for Poverty Action (IPA) emphasizes tracking how rigorous evidence can influence programs and policies that go beyond a single program's evaluation (Innovations for Poverty Action, n.d.-a). IPA frames RCTs and related analyses as a critical part of impact evaluation and in service to their particular societal goal of reducing poverty (Innovations for Poverty Action, n.d.-b).

Challenging the "When" and "What" of Impact Evaluations

"When" to conduct

Impact evaluations are not always appropriate. For example, it is a resource-intensive process, so it should be applied when there is a need for outcome-related information and the program is well-structured and defined (Rossi et al., 2004). They are best for new(-ly growing) programs and programs with uncertain effectiveness, and should generate generalizable knowledge beyond a single program (Savedoff et al., 2006). Even these authors note the need for counterfactuals; Savedoff et al. (2006) emphasize that construction of the necessary comparison groups should be present in an impact evaluation from the first stages of design, rather than as ad hoc efforts for causal attribution. Stern et al. (2012) present impact evaluation definitions from across organizations and identify three primary elements that align with the definition of random evaluations; while they note a need for a causal link, they caution that "attribution" implies an exclusive relationship between interventions and impact that are unlikely.

"What" & how to define

Similarly, there is no consensus among authors and organizations as to how define impact evaluation. Rossi et al. (2004) broaden the scope to the measurement of programmatic effects on societal conditions and define impact as the net effects of programs beyond what the beneficiary group would receive without the program. Roche (2000) provides a definition produced by a group of NGOs: "impact assessment is the systematic analysis of the lasting or significant changes-positive or negative, intended or not-in people's lives brought about by a given action or series of actions" (p. 546). The author notes that, in this definition, contextual information and judgments would drive what was and was not considered impact.

Logic models and theory of change

There are efforts to ground impact evaluation within the language of theory of change (TOC) and logic models. Schaffer (2011) identifies impact evaluation as focused on the last two elements, outcomes and impacts, referring to "longer-term effects of projects usually on some dimension of well-being" (p. 1621). Stern et al. (2012) explicitly note that impact evaluations should be consistent with TOC logic.

Identifying gaps in definitions

It is also possible to expand on potential gaps in the previously identified definitions by analyzing trends in discourse, i.e., the literature used to train evaluators. In their analysis, Gugerty et al. (2021) identify certain frequent concepts: internal validity, disclosure of methodological limitations, theory-based evaluation using formalized concepts and methods (ex. theory of change, logical frameworks), attempts to establish causal attribution, using both qualitative and quantitative methods (or even explicit "mixed methods"), and the use of a counterfactual. While

not all of the concepts are presented in the cited literature here, their presence in discourse suggests that each of these concepts compose some part of impact evaluation.

Appendix 1B: Data & Analysis

Final Dataset by Country and Quotation Count

The following dataset is organized in alphabetical order, first by "Self-Identified Country of Origin" and second by "Document Group."

Table 7. Reported output for n=10 for Topic 1. Comparison of source and final dataset.

Document Group	Self-Identified Country of Origin	In new dataset?	Number of quotations
Agency Coordinating Body for Afghan Relief and Development (ACBAR)	Afghanistan	Yes	3
Australian Council for International Development (AFCID)	Australia	Yes	5
Botswana Council on NGOs (BOCONGO)	Botswana	Yes	1
Cooperation Committee for Cambodia	Cambodia	No	0
Canadian Council for International Co-operation	Canada	Yes	1
Imagine Canada	Canada	No	0
Consortium of Christian Development and Relief Association (CCRDA)	Ethiopia	Yes	4
EU European Foundation Centre	European Union	No	0
Council of Finish Foundations (COFF)	Finland	Yes	4
Voluntary Action Network India (VANI)	India	Yes	3
Dochas	Ireland	Yes	3
Japan NGO Center for International Cooperation (JANIC)	Japan	Yes	1
Viwango	Kenya	Yes	4
Korean NGO Council for Overseas Cooperation (KCOC)	Korea	Yes	1
Council for International Development (CID)	New Zealand	Yes	2
Norwegian Refugee Council	Norway	Yes	16

Document Group	Self-Identified Country of Origin	In new dataset?	Number of quotations
Pakistan NGOs Forum	Pakistan	Yes	2
NGO Development Center (NDC)	Palestine	Yes	3
Philippine Council for NGO Certification	Philippines	Yes	1
National Council on NGOs (NACONGO)	Tanzania	Yes	2
Quality Assurance Mechanism (QuAM)	Uganda	Yes	5
Bond	United Kingdom	Yes	6
Department for International Development (DFID)	United Kingdom	Yes	8
Better Business Bureau-Wise Giving Alliance	United States	Yes	1
Candid (formerly GuideStar)	United States	Yes	4
Charity Navigator	United States	No	0
Council on Foundations	United States	No	0
GiveWell	United States	Yes	10
Impact Matters	United States	Yes	4
InterAction	United States	Yes	4
United States Agency for International Development (USAID)	United States	Yes	7
Accountable Now (formerly INGO Accountability Charter)	N/A	Yes	6
Active Learning Network for Accountability and Performance in Humanitarian Action (ALNAP)	N/A	Yes	25
Core Humanitarian Standard (CHS)	N/A	Yes	3
Global Reporting Initiative (GRI)	N/A	Yes	10
Global Standard for CSO Accountability	N/A	Yes	4
OECD Development Assistance Committee (DAC)	N/A	Yes	9
Sphere	N/A	Yes	10
United Nations Development Program (UNDP)	N/A	Yes	5
United Nations Environmental Group (UNEG)	N/A	Yes	4

Document Group	Self-Identified Country of Origin	In new dataset?	Number of quotations
United Nations Sustainable Development Goals (SDG)	N/A	Yes	19
World Association of NGOs (WANGO)	N/A	Yes	2

Preparing the data for analysis within R

Any observations that did not have quotation content were removed as a check. The quotations were cleaned and transformed into a "corpus" object using the provided functions in the R package "stm" (Roberts et al., 2019). The steps include converting all text to lowercase, stemming text (i.e., removing ending characters that cause "professor" and "professing" to be treated as distinct from "profess", where "profess" can be seen as the trunk for both terms), removing words of less than three characters, replacing non-alphanumeric character values, and removing strings with low explanatory utility, such as stop words (e.g., "a" and "the"), numbers, and punctuation. Finally, all relevant metadata values were converted to factor values.

Preparing Quotation Export

Quotations were exported from Atlas.ti as a quotation report, where each row is an observed, coded unit. Columns reflect quotation-level characteristics. These variables are included because of their demonstrated utility in demarcating and differentiating among clusters within the dataset, as evidenced in Table 6 of Gugerty et al. (2021, p. 9). Only three quotations appeared more than once. The standard written by ALNAP had the following language reported both in the main body and in the glossary section:

> An evaluation that focuses on the wider effects of the humanitarian programme, including intended and unintended impact, positive and negative impact, macro (sector) and micro (household, individual) impact.

> Analysing contribution in evaluation refers to finding credible ways of showing that an intervention played some part in bringing about results.

ImpactMatters used the same language when describing their audit standard and when providing an overview of their methodology:

> Quality of Impact Evidence captures how confident we are that the nonprofit's program is leading to impact on its primary outcomes.

While the differing locations in the text and intentional choice by ALNAP and ImpactMatters to explicitly retain and reinforce that language may be an argument to retain all three sets of quotations in the corpus, the small and specific nature of the corpus suggests that the duplication of 3 entries out of 205, or 1.5% of the dataset, may affect the output. For example, a previous version of the final model featured the duplicated quotations as among the most representative documents for two different topics, so the effect is clearly observable. To prevent such issues, I have removed the duplicated text from the output. Finally, to account for transcription issues when exporting the quotations from ATLAS.ti, I ran the basic spell check feature in Excel for both "English (United Kingdom)" and "English (United States)" and compared each identified issue to its coded sentence in ATLAS.ti to ensure fidelity to the source data. I also manually entered some sentences that were inappropriately captured by ATLAS.ti, as well as made edits to words that were broken across lines in the source data due to formatting constraints (ex. "im- pact" to "impact").

Comparison example outputs of TM and STM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
TM	gender	evid	project	will	programm	approach	target	long	humanitarian	effici
	women	rigor	activ	identifi	effect	case	beneficiari	shall	action	relev
	happen	type	possibl	allow	part	theori	system	term	impact	communiti
	power	studi	social	improv	caus	intend	monitor	work	given	posit
	provid	chariti	main	past	chain	method	polici	particip	interest	sustain
	stakehold	strongest	conduct	organis	causal	unintend	understand	collect	want	negat
	men	robust	econom	carri	examin	test	within	among	popul	need
	right	captur	expenditur	report	focus	observ	design	govern	strategi	oper
	result	live	report	environment	nrc	can	take	improv	affect	address
	cso	context	must	process	result	explain	potenti	much	wider	whether

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
STM	gender	rigor	project	well	difficult	observ	system	long	humanitarian	effici
	women	evid	involv	improv	part	theori	target	shall	action	communiti
	happen	chariti	signatori	report	usual	multipl	nonprofit	term	affect	relev
	men	live	activ	allow	intend	case	set	work	given	sustain
	power	life	social	past	caus	method	beneficiari	organiz	wider	address
	provid	studi	carri	identifi	might	approach	deliv	support	popul	oper
	stakehold	captur	possibl	process	examin	attribut	harm	among	peopl	program
	commit	per	econom	scale	causal	test	potenti	improv	last	servic
	equal	robust	main	estim	effect	complex	within	make	done	object
	promot	save	societi	order	defin	group	take	made	household	posit

Figure 6. Comparison of two topic models, unstructured (TM) and structured (STM), with identical implementations but for the inclusion of the topical prevalence covariate "Document_groups" in the structured topic model. The model was run for k = 10 topics. For each topic, there is a list of the 10 words with the highest score regarding its balance of frequency and exclusivity to that topic (in order from highest to lowest score).

Reported outputs for Topic 1

Table 8. Reported output for n=10 for Topic 1.

Topic 1	Marginal Highest Prob.	impact, shall, report, defin, program, evalu, organ, effect, outcom, inform
	Marginal FREX	shall, report, defin, impact, program, organ, mission, evalu, effect, outcom
	Marginal Lift	defin, shall, report, government, event, product, govern, annual, organ, avail
	Marginal Score	shall, report, defin, program, organ, impact, inform, mission, event, govern
	Topic Kappa	report, event, organ, product, defin, expenditur, reason, inform, servic, period
	Kappa with Baseline	report, program, organ, inform
Covariate 1	Marginal Highest Prob.	defin, evalu, impact, effect, inform, develop, outcom, assess, intervent, organ
	Marginal FREX	defin, organ, report, program, evalu, effect, inform, outcom, assess, use
	Marginal Lift	defin, event, product, one, pilot, basic, dfid, partner, expenditur, reason
	Marginal Score	defin, report, organ, inform, event, product, servic, reason, period, expenditure
Covariate 2	Marginal Highest Prob.	impact, report, organ, program, evalu, measur, assess, effect, inform, signific
	Marginal FREX	report, impact, organ, program, measur, estim, signific, activ, evalu, assess
	Marginal Lift	estim, report, reflect, mitig, organ, event, reason, product, defin, environment
	Marginal Score	report, impact, organ, program, estim, measur, signific, mitig, defin, reason
Covariate 3	Marginal Highest Prob.	shall, program, impact, evalu, report, outcom, effect, inform, organ, particip
	Marginal FREX	shall, program, report, outcom, organ, mission, particip, govern, plan, defin
	Marginal Lift	shall, government, govern, annual, avail, incom, regard, plan, report, organiz
	Marginal Score	shall, report, program, organ, outcom, mission, govern, particip, plan, inform

Marginal performance metrics are calculated using sageTopics() in the R package "stm."

Selection of Highly Representative Documents

Table 9. Selection of documents highly ranked (first & second) as representative of topics.

Topic	Quotation (Raw text)	Producing Organization	Organizational Type	Regional Focus	Sectoral Origin	Standard Type
1	Each Non Governmental Organization shall: (a) regularly communicate, in clear and accessible manner its values, governance structure, mission objectives and approaches and progress made in its work (shall share the vision and mission to stakeholders and new members as often as possible); (b) develop reasonable budgets that clearly correspond with its programs and plans; (c) systematically monitor, evaluate, document and report on the progress of its programs and plans; (d) conduct periodic independent evaluations that shall examine, among other aspects, the quality of results, effectiveness and impact of its work; and I compile and make available to stakeholders an annual report that shall state the governing structures; the main achievements, challenges and lessons learned in the course of implementation, as well as the annual incomes, expenditures and balances.	NACONGO (Tanzania)	Association	National	Nongovernmental	Self-regulation
	Each Non Governmental Organization shall ensure that there is effective Management and Information System (MIS) which provides frameworks and guidelines of how stakeholders shall be informed of the organizational functions and outcomes and vice versa.	NACONGO (Tanzania)	Association	National	Nongovernmental	Self-regulation
2	NGOs have a moral responsibility to ensure that projects they initiate are sustainable and economically viable, and in particular such projects will: i. Be responsive to community needs and aspirations and contribute to their overall development directly or indirectly. Such projects should not be donor driven. ii. Not be detrimental to the well being of the communities. iii. Promote and support effective community participation by empowering communities to take responsibility and ownership. iv. Provide enough political and social space for communities to determine the modes of implementation and project management relevant to them.	BOCONGO (Botswana)	Association	National	Nongovernmental	Self-regulation

Topic	Quotation (Raw text)	Producing Organization	Organizational Type	Regional Focus	Sectoral Origin	Standard Type
	They are also useful and feasible where a programme has a phased roll-out or where there is expected over-subscription of beneficiaries (and hence not all will receive the programme) as an opportunity is presented to establish a comparison group of non-beneficiaries.	DFID	Donor	National	(Inter)Governmental	Self-regulation
3	We rate the quality of the following systems used to monitor delivery of the intervention: Activity: track program activities and outputs delivered Targeting: identify beneficiaries to receive the program Engagement: track if participants are taking up the program and meeting targets Feedback: understand how participants view the program Outcomes: measure changes in beneficiary outcomes	Impact Matters	Infrastructure	International, cross-national	Nongovernmental	Third Party
	Nonprofits with strong monitoring systems can credibly show that their programs are reaching the claimed number of beneficiaries, and that nonprofit staff have the data and systems to identify problems in implementation and take action to correct those problems.	Impact Matters	Infrastructure	International, cross-national	Nongovernmental	Third Party
4	It analyses the level of achievement of both expected and unexpected results by examining the results chain, processes, contextual factors and causality using appropriate criteria such as relevance, effectiveness, efficiency, impact and sustainability.	UNEG	Infrastructure	International, cross-national	(Inter)Governmental	Third Party
	Evaluation aims to understand why — and to what extent — intended and unintended results were achieved and to analyse the implications of the results.	UNEG	Infrastructure	International, cross-national	(Inter)Governmental	Third Party
5	The goal of gender-responsive evaluation is to: 1. Assess the degree to which gender and power relationships— including structural and other causes that give rise to inequities, discrimination and unfair power relations—change as a result of an intervention using a process that is inclusive, participatory and respectful of all stakeholders (rights holders and duty bearers) 2. Provide information on the way in which development programmes are affecting women and men differently and contributing towards achievement of these commitments 3. Help promote social change by using the knowledge produced from an evaluation for better development programming that promotes gender equality, women's empowerment and human rights in a sustainable manner	UN SDG	Infrastructure	International, cross-national	(Inter)Governmental	Third Party

Topic	Quotation (Raw text)	Producing Organization	Organizational Type	Regional Focus	Sectoral Origin	Standard Type
	Gender-responsive evaluation assesses the degree to which gender and power relationships—including structural and other causes that give rise to inequalities, discrimination and unfair power relations, change as a result of an intervention using a process that is inclusive, participatory and respectful of all stakeholders (right holders and duty bearers).	UN SDG	Infrastructure	International, cross-national	(Inter)Governmental	Third Party
6	The affected people are the best judges of changes in their lives; hence outcome and impact assessment must include people's feedback, open-ended listening and other participatory qualitative approaches, as well as quantitative approaches.	Sphere	Infrastructure	International, cross-national	Nongovernmental	Third Party
	In order to monitor the results of a project: You need to measure a change in an indicator, and It must be possible to attribute this change to the project activities, in part or in full.	Sphere	Infrastructure	International, cross-national	Nongovernmental	Third Party

Comparing Expected Topic Prevalence

Table 10. Comparison of Expected (Mean) Topic Prevalence, rounded, by Topic and Prevalence Covariate.

Covariate Value	**Topic 1** *Compliance*	**Topic 2** *Community Benefits*	**Topic 3** *Establishing Systems*	**Topic 4** *Developing Understanding*	**Topic 5** *Creating Change*	**Topic 6** *Data Engagement*	Row Sums
Organizational Type							
Donor	0.023	0.266	0.080	0.347	0.248	0.036	1
Infrastructure	0.192	0.027	0.208	0.100	0.265	0.210	1
Association	0.141	0.063	0.227	0.214	0.323	0.033	1
Regional Focus							
National	0.189	0.146	0.233	0.215	0.142	0.075	1
International, Cross-national	0.104	0.024	0.171	0.173	0.414	0.115	1
Sectoral Origin							
(Inter) Governmental	0.015	0.151	0.049	0.274	0.488	0.022	1
Nongovernmental	0.186	0.054	0.251	0.163	0.223	0.123	1
Standard Type							
Self-regulation	0.188	0.107	0.271	0.255	0.129	0.050	1
Third party	0.102	0.055	0.136	0.136	0.433	0.138	1

Appendix 1C: Held-out Likelihood

Held-out log likelihood is a metric proposed by Wallach et al. (2009) that is designed to compensate for numerous approaches to evaluating topic models at the time of writing. In particular, the document completion method (implemented in the "stm" package) uses a cross-validation-style approach in which a subset of the data is allocated for training the model, a second portion is used for testing, and a third, smaller subset is created to test for and validate optimal model parameters before running the model on the "testing" subset. Document completion divides a document's content into two subsets: the first is used to train the model and the second (or held-out portion) is used to evaluate the model in terms of its predictive capacity (Roberts et al., 2019; Wallach et al., 2009). Higher numbers, i.e., values that are closest to 0, indicate better model performance for the given data (Roberts et al., 2016a). When I identify solely the top-performing cluster values for each of the four models (see Appendix: Held-out Likelihood), I find that there is not a clear trend towards smaller or larger sizes for *k*. I do observe that the "full" model consistently outperforms the other three in terms of predictive capacity. However, as Roberts et al. (2016b) note, "in the case of topic models, prediction is not the only relevant standard" (p. 11).

Table 11. Comparison of top held-out likelihoods by topic and model

Model	**Held-Out Likelihood (Rounded)**	**Corresponding *k* Topic Size**
Naïve	-5.596	3
Prevalence	-5.511	3
Content	-5.510	12
Full	-5.375	20

It is not obvious which of these values represent a universal (i.e., across all topic sizes and models) peak and which may represent more of a potential outlier due to fluctuations in performance. As presented in Figure 7 and Figure 8, we can visually explore the held-out likelihood by number of topics for each of the four models. This allows us to gain insights into which models typically perform better than the others. Via the fixed y-axes in Figure 8, we can observe that the "full" and "content" models consistently outperform the other two models in terms of held-out likelihood metrics across the range of *k* topic sizes. The "prevalence" model has its best performance for 2, 3, and 6 topics, but performance drops significantly with increasing topic sizes. The "naïve" model similarly experiences a general reduction with increasing topic sizes.

In contrast, the "content" and "full" models report generally equivalent performance across the range of *k* values. Both models show a performance reduction in the range of $k = [15,20]$, but even then, the "full" model's performance is better than the "content" model. Both models' performance is more desirable to that of the "naïve" and "prevalence" models, with the "full" model consistently outperforming the "content" model in terms of predictive capacity. However, as Roberts et al. (2016b) note, "in the case of topic models, prediction is not the only relevant standard" (p. 11).

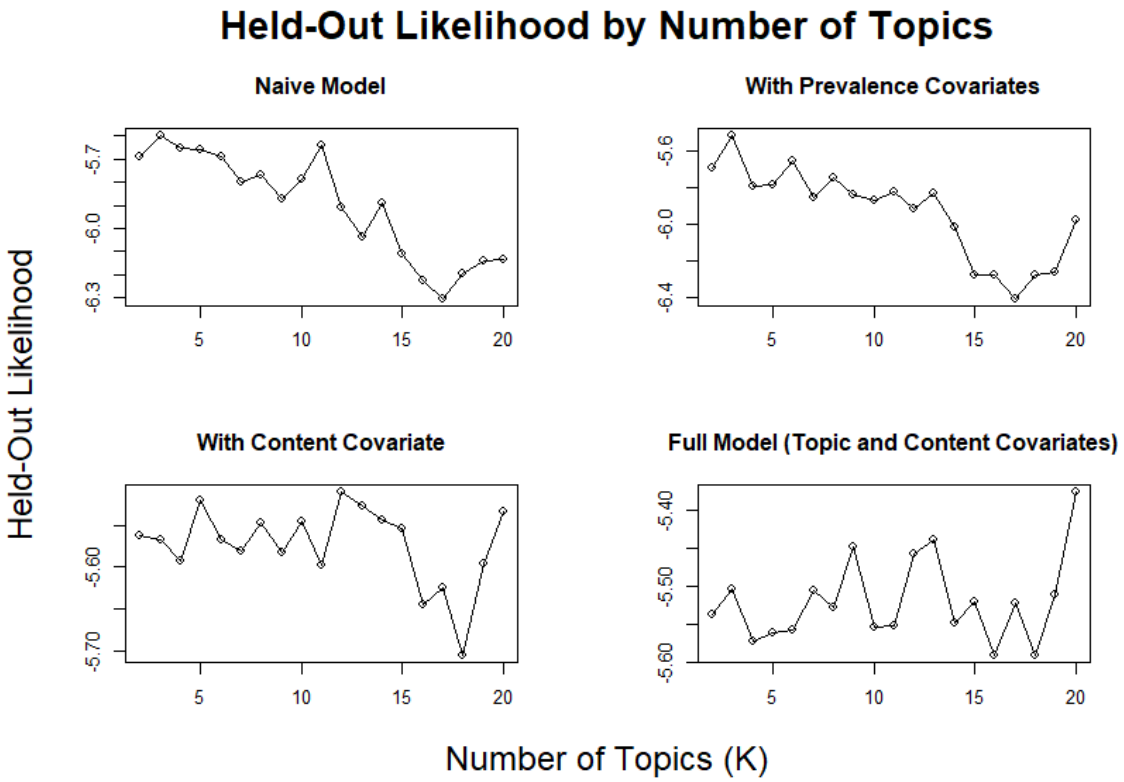

Figure 7. Comparison of the held-out likelihood values by number of topics for each of the four models. Higher values (closer to 0) are desirable.

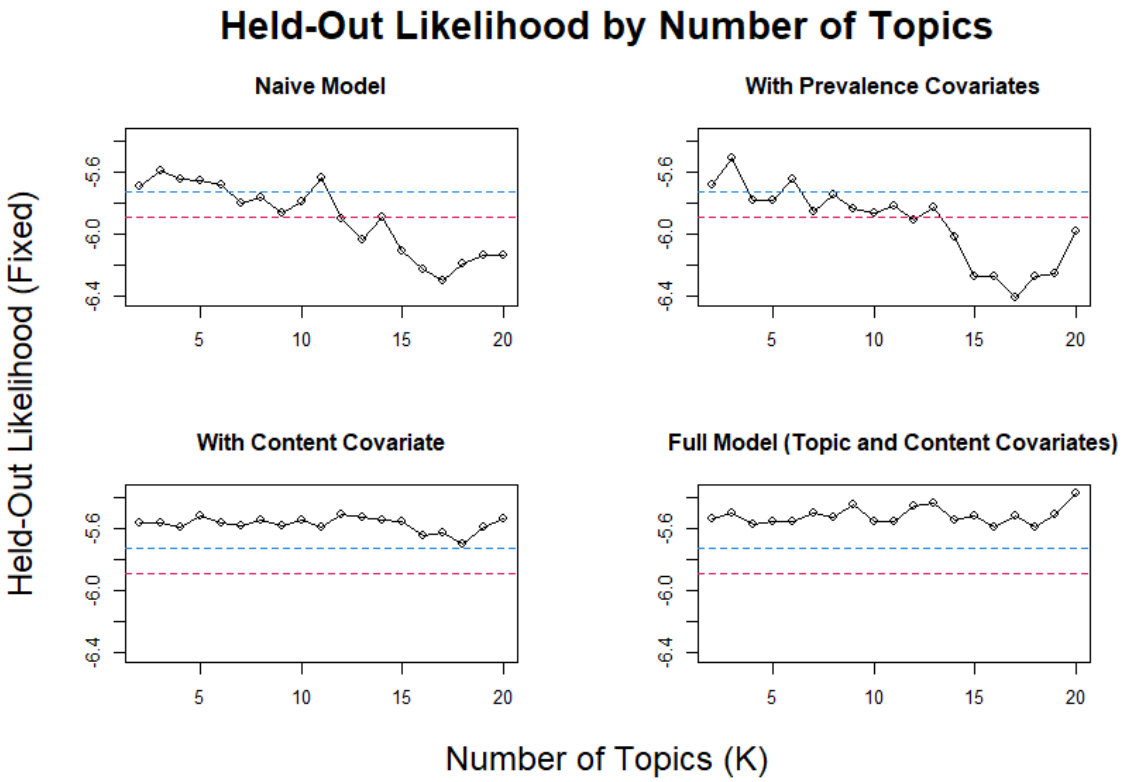

Figure 8. Alternative version of Figure 7, with y-axis minimum and maximum values shared across all four graphs. The red dashed (lower) line indicates the mean of the selected y minimum and maximum values. The blue dashed (upped) line indicates the mean of all held-out likelihood values across combinations of model parameters and k values.

Appendix 1D: Residuals Analysis

As a companion measure of performance, I also looked at the reported values for analysis of residuals. Residual analysis is defined by Taddy (2012) as "the simple connection between number of topics and model fit...[where] any fitted overdispersion $\hat{\sigma}^2 > 1$ indicates a true *K* that is larger than the number of estimated topics" (p. 1188), where instances where the measurement of dispersion $\sigma^2 = 1$ indicate that the observed and theorized number of clusters *k* are equivalent. In other words, values closer to 1 are preferable because they indicate less distance between the observed and true *k* number of topics. Roberts et al. (2019) provide a slightly alternate framing in the "stm" support documentation, noting that the implied or target theoretical value of 1 indicates when the identified number of latent topics accounts for the observed dispersion.

Given the held-out likelihood and residual analysis performance of the four models by topic size, I will use the full model to generate topics and analyze the quotations. Because I still have some questions around the topic size *k*, I can turn to an alternative performance measure that considers the content of the topics.

As we did with the held-out likelihood values, we can identify the optimal residual values and *k* topic size for each of our four models (see Table 12). As compared to the held-out likelihood metric, focusing solely on the residuals would have us leaning primarily towards larger topic sizes. Of note is that the "content" and "full" models have smaller residual values than either the "naïve" or the "prevalence" models.

Table 12. Comparison of top residual analysis values by topic and model

Model	**Residuals (Rounded)**	**Corresponding *k* Topic Size**
Naïve	1.115	8
Prevalence	1.134	7
Content	0.940	17
Full	0.969	17

We can again gain insight into the structure of the data and trends by looking at visualizations of the residuals (Figure 9 and Figure 10). Both the "naïve" and "prevalence" models exhibit the concave curves typically evidenced for this analysis. For those models, the minimum residual values reflect global minima for our range of *k* values, after which point residuals begin to increase as *k* increases. This is not the case for the "content" and "full" models. Increasing the upper range of *k* values may demonstrate that the identified minima are merely local or indeed global. There is another concern, that of overfitting the data or when a model has been trained to perfectly predict only on the provided data and suffers when applied to other datasets. Because of this concern, a model that exactly aligns with a theoretical performance metric is not actually desirable. When looking at the comparison graphs of Figure 10, the "content" model seems to be consistently overfitting the data at *k* values of 15 and larger and the "full" model at *k* values of 17 and larger.

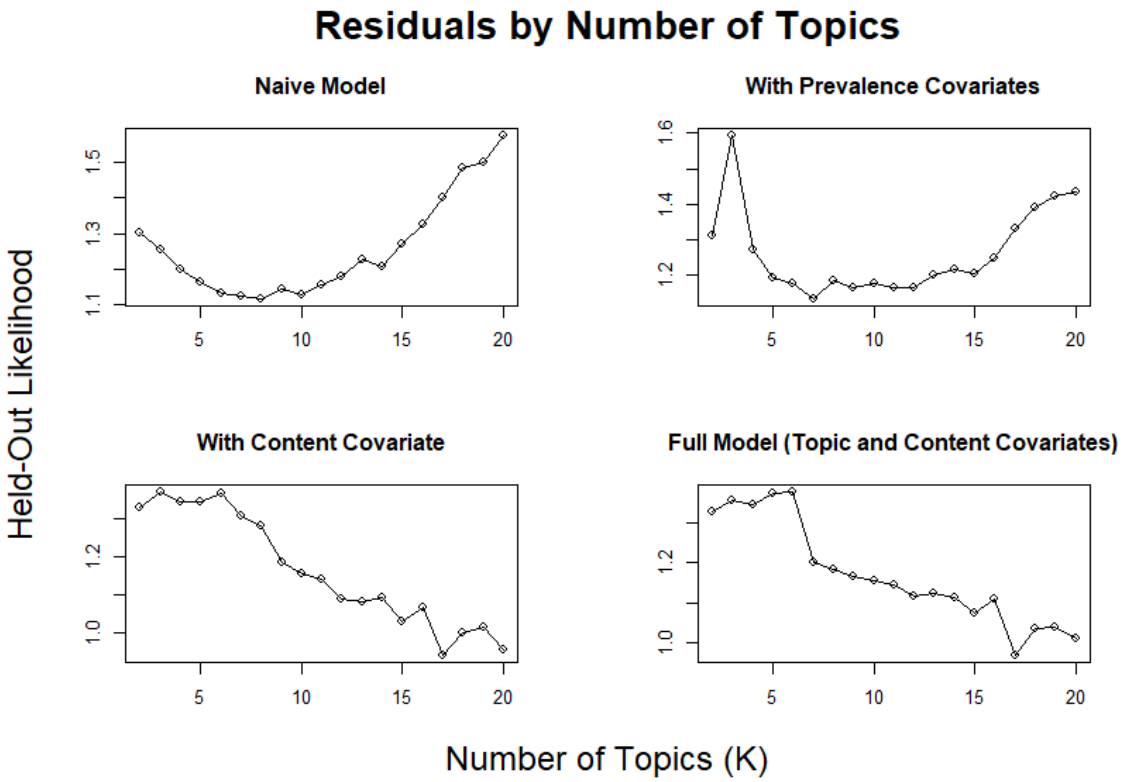

Figure 9. Comparison of the residual analysis values by number of topics for each of the four models. Values closer to 1 are desirable.

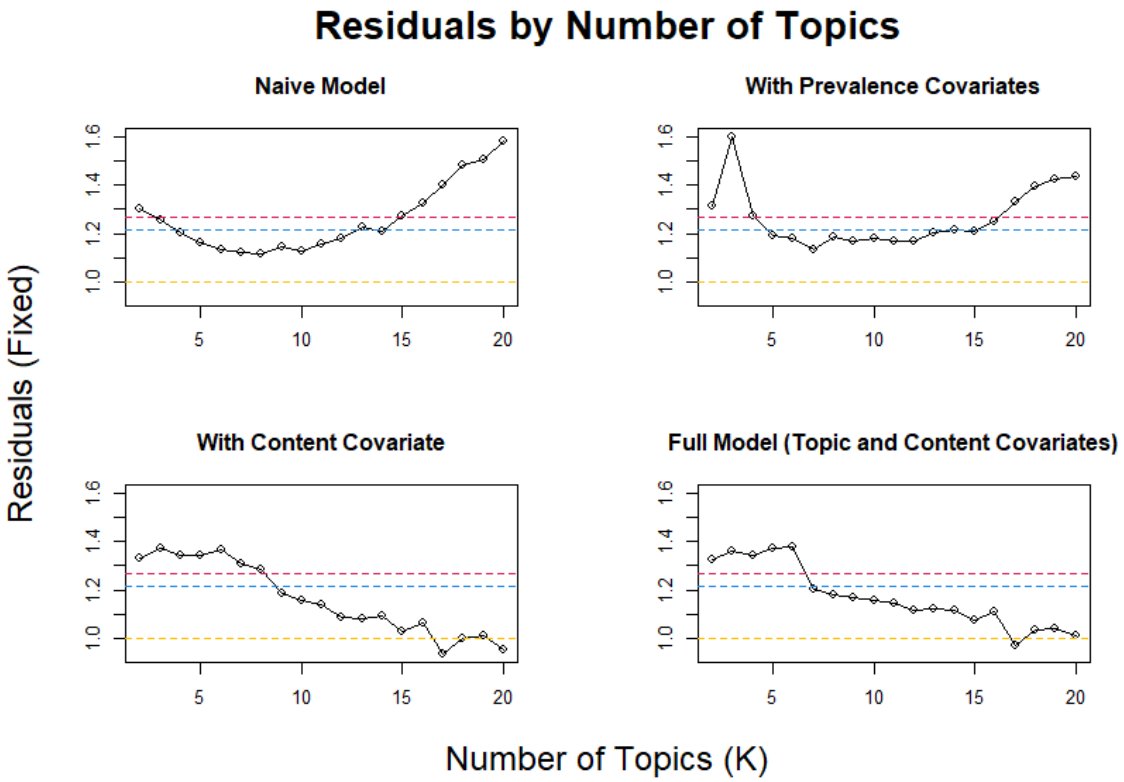

Figure 10. Alternative version of Figure 9, with y-axis minimum and maximum values shared across all four graphs. The red dashed (upper) line indicates the mean of the selected y minimum and maximum values. The blue dashed (lower-center) line indicates the mean of all residual analysis values across combinations of model parameters and k values. The yellow dash (bottom) line indicates the target residual dispersion value of 1.

Appendix 1E: Semantic Coherence

A substantive challenge of topic modeling, in general, comes from the push-and-pull of statistical validity and model accuracy against the need to make sense of seemingly random collections of words. "There is a strong relationship between the size of topics and the probability of topics being nonsensical as judged by domain experts: as the number of topics increases, the smallest topics (number of word tokens assigned to each topic) are almost always poor quality" (Mimno et al., 2011, p. 262). While I have a sense of which topic model sizes to consider for the full model, gaining insight into which size yields meaningful semantic coherence can help me narrow it down. Semantic coherence is a measure initially proposed by Mimno et al. (2011) in their efforts to create a metric that could approximate the capacity of experts to determine topic quality. The authors focus on word co-occurrence within documents and found that it performed better than comparable metrics, such as pointwise mutual information. They created a metric that does so by taking the log of the co-document frequency of pairs of the most probable words for a given topic divided by the document frequency of just one of those words for that topic (Mimno et al., 2011). High semantic value is thus achieved when words appear more commonly together than on their own. There is the risk that this measure of internal consistency can be warped by how common some words are, creating a high semantic coherence for topics that suffer from poor exclusivity (Roberts et al., 2014, 2019). I can still use semantic coherence to approximate which value of *k* will yield a model whose topics will consistently be of least poor quality, as indicated by which set of topics have the highest or least-negative semantic coherence values.

I ran the full model with topical prevalence and content covariates for a range of *k* values of 2 to 20 to allow for comparative performance of potential numbers of topics *k* relative to neighbors. The utility of the metric may be limited by the introduction of the content covariate (which

introduces versions of each topic; B. Stewart, personal communication, March 20, 2022). With that in mind, I compared the reported values to scores in the published literature (ex. Mimno et al., 2011); as they were in line, the measure seems appropriate for use. I explored the descriptive statistics for each model-*k* combination (see Appendix: Semantic Coherence). *k* = 6 has the best semantic coherence values, in terms of the highest minima, median, and mean, the smallest range, and the third-highest maxima, suggesting that the structural topic model with *k* = 6 contains topics that consistently seem to score among the best quality.

We can visually explore the data via Figure 11. The lowest semantic coherence value ranges and among the smallest values can be found among two of the smallest *k* values, 3 and 4. Other poorly performing topics can be found at *k* = 9, 13, and 18. While these may seem like extreme values, there is an observable deterioration in performance of the ranges as *k* values increase when one observes the shift in the mean and the overall distribution. Some of the most semantically coherent topics can be found for *k* = 18, 19, and 20. But, the mean and minima of those ranges do not compare favorably to those of *k* = 5 and 6. Finally, when comparing those two topic sizes, we see that the semantic coherence value for the lowest-performing topic of *k* = 6 is almost equivalent to the mean of the range for *k* = 5. In other words, *k* = 6 contains topics that consistently seem to score among the best quality.

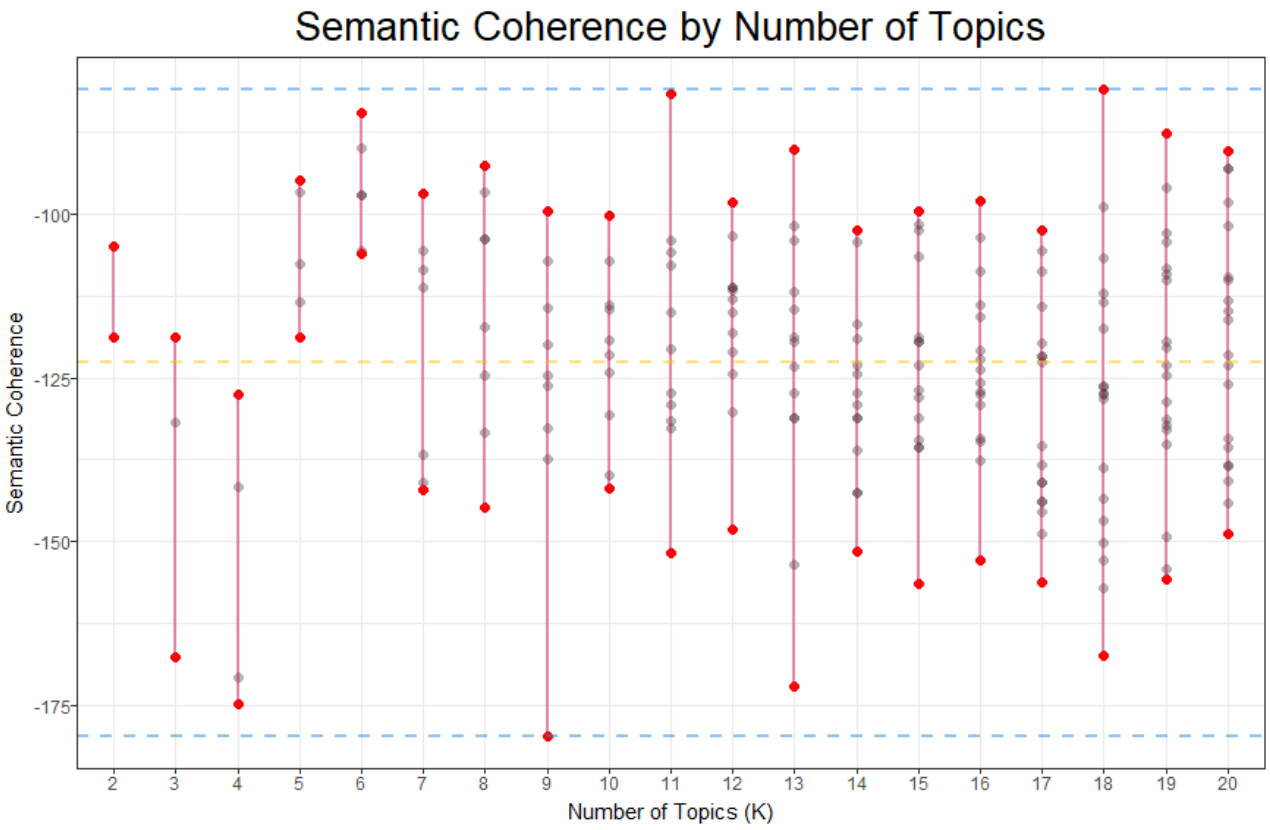

Figure 11. Comparison of semantic coherence by number of topics for the model using topical prevalence and content covariates. Red dots indicate the highest and lowest values for a given topic size. The red (vertical) line indicates the range of semantic coherence values for that number of topics. The blue dashed (lower and upper horizontal) lines indicate the minimum and maximum semantic coherence values across all combinations of models and number of topics. The yellow dash (center horizontal) line indicates the mean semantic coherence value across all models and number of topics.

Table 13. Descriptive statistics for semantic coherence values (rounded) by number of topics (k)

k	**Min**	**Max**	**Range**	**Median**	**Mean**
2	-118.878	-104.854	14.024	-111.866	-111.866
3	-167.632	-118.816	48.816	-131.793	-139.414
4	-174.782	-127.491	47.291	-156.206	-153.671
5	-118.882	-94.737	24.145	-107.621	-106.299
6	**-106.083**	-84.462	**21.621**	**-97.001**	**-96.683**
7	-142.079	-96.828	45.251	-111.094	-120.25
8	-144.884	-92.568	52.316	-110.53	-114.603
9	-179.826	-99.606	80.22	-124.687	-126.883
10	-141.935	-100.145	41.79	-120.309	-121.288
11	-151.867	-81.703	70.164	-120.532	-118.836
12	-148.068	-98.095	49.973	-113.977	-117.101
13	-172.061	-90.041	82.02	-119.428	-123.014
14	-151.531	-102.373	49.158	-128.238	-127.247
15	-156.449	-99.589	56.86	-123.006	-122.612
16	-152.992	-97.926	55.066	-124.705	-123.462
17	-156.342	-102.545	53.797	-135.409	-130.052
18	-167.548	**-80.963**	86.585	-127.512	-129.04
19	-155.827	-87.601	68.227	-123.077	-122.378
20	-148.871	-90.3	58.571	-118.82	-119.571

Added bold emphasis is used to indicate the maximum value in a given column. In the column "Range", the value in bold indicates the smallest range.

Appendix 1F: FREX

As noted earlier, there are some potential issues with semantic coherence. One method to counter-balance the potential for a few common words that appear frequently together to undermine the utility of semantic coherence is to pair that metric with a FREX score, a method that calculates the harmonic mean of a word in regards to its topic-specific frequency and its exclusivity to a topic relative to the others (Airoldi & Bischof, 2012; Bischof & Airoldi, 2012; Roberts et al., 2014; Roberts et al., 2019). For a topic model with no content covariates, we can use the relationship between semantic coherence and FREX scores to identify a model with topics at the frontier of the relationship, and a balance between the two performance metrics (Roberts et al., 2014). In Appendix: FREX, Table 14 presents descriptive statistics and Figure 12 visualizes the ranges for FREX scores by combination of model and number of topics *k*.

Those findings have limited utility because the full model I propose includes a content covariate, which introduces complexities to the calculation of FREX scores. Because each of the *k* topics has a different version for each level of the content covariate and the FREX score treats each version as separate topics, it would then be calculating FREX scores across the full universe of topics – different versions of the same topic (by content covariate level) would be evaluated against each other in regards to frequency and exclusivity of words (B. Stewart, personal communication, March 20, 2022). In addition, including a content covariate automatically enables covariate-topic interactions and the creation of topics via the Sparse Additive Generative (SAGE) model, which yields terms with greater saliency for a given topic through summing the total variation of words within each topic label, thus "focus[ing] on high frequency terms with accurate counts" (Einstein et al., 2011, p. 1046; Roberts et al., 2019). In other words, the defaulted inclusion of the SAGE model should address whatever goal we have of achieving desirable FREX scores within our

topics. In addition, since the FREX scores between topics is what is of interest rather than within versions of the same topic, current FREX calculations will not prove helpful in identifying optimal *k* sizes.

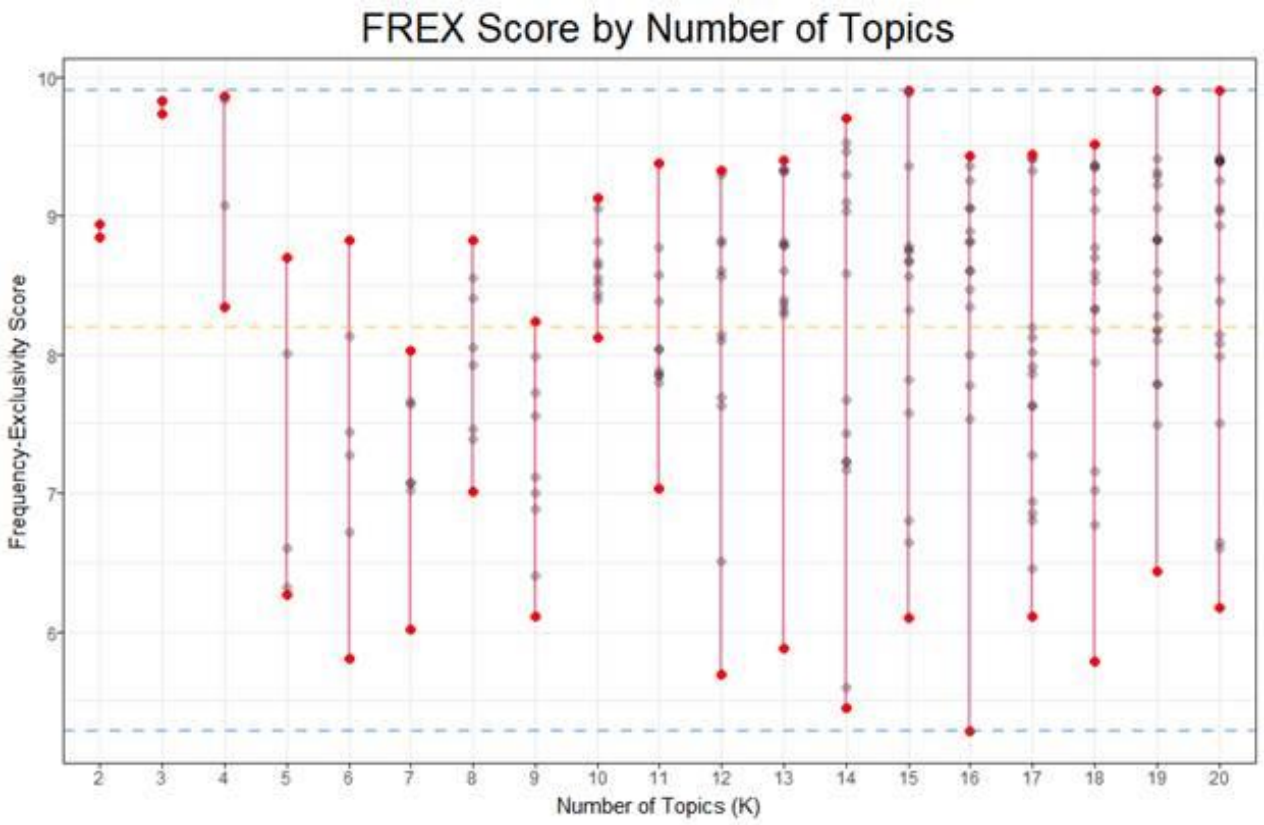

Figure 12. Comparison of FREX frequency-exclusivity score by number of topics for the model using topical prevalence and content covariates. Red dots indicate the highest and lowest values for a given topic size. The red (vertical) line indicates the range of FREX scores for that number of topics. The blue dashed (lower and upper horizontal) lines indicate the minimum and maximum FREX scores across all combinations of models and number of topics. The yellow dash (center horizontal) line indicates the mean FREX score across all models and number of topics. These values have limited utility because they ineffectively address the intra-topic levels introduced by the content covariate.

Table 14. Descriptive statistics for FREX scores (rounded) by number of topics (k)

k	**Min**	**Max**	**Range**	**Median**	**Mean**
2	8.843	8.942	0.099	8.892	8.892
3	**9.738**	9.825	0.086	**9.817**	**9.793**
4	8.336	9.855	1.519	9.453	9.274
5	6.264	8.702	2.438	6.599	7.177
6	5.809	8.82	3.011	7.353	7.363
7	6.015	8.029	2.014	7.076	7.216
8	7.015	8.823	1.807	7.98	7.951
9	6.106	8.235	2.129	7.115	7.222
10	8.118	9.122	**1.004**	8.592	8.629
11	7.037	9.382	2.345	8.037	8.145
12	5.696	9.329	3.632	8.351	8.098
13	5.881	9.397	3.517	8.782	8.585
14	5.453	9.701	4.247	8.13	8.034
15	6.098	9.896	3.798	8.667	8.305
16	5.281	9.431	4.15	8.707	8.454
17	6.106	9.437	3.331	7.855	7.845
18	5.785	9.512	3.727	8.554	8.326
19	6.43	**9.899**	3.469	8.595	8.523
20	6.169	9.897	3.728	8.737	8.368

Added bold emphasis is used to indicate the maximum value in a given column. In the column "Range", the value in bold indicates the smallest range. These values have limited utility because they ineffectively address the levels introduced by the content covariate.

Paper 2. Rationalizing the Valuation of Goods and Services

Introduction

The gap in understanding of programmatic efficiency in the nonprofit sector is hampered by an incomplete awareness and categorization of resources required to deliver those programs. Perhaps the least understood of these resources are in-kind contributions[1] of goods and services[2] ("in-kind contributions"), such as donations of canned food or helping at an animal shelter. For many organizations, these contributions are an important part of nonprofits' resource portfolios, yet their valuation is inconsistent across most nonprofits. While some organizations develop their own valuation methods (ex. Feeding America), in-kind contributions are to be financially valued and reported in different ways in their annual financial statements relative to their tax forms, further limiting consistency in valuations and data used for analysis (FASB, 2008; USDT, 2018). Inconsistencies even among this minimal, financial valuation suggest that this commonly used

[1] Financial Accounting Standards Board (FASB) guidance and IRS Form 990 instructions vary in their interpretation of "contribution" vs. "donation." The IRS instructions use the words interchangeably, whereas "contribution" is used almost exclusively in the FASB sources cited herein. For consistency, I default to using "contributions" from here on except when a cited or quoted source uses "donations." Various websites provide guidance on how the two words can be used differently, such as: https://subfictional.com/donation-vs-contribution-payment-processors-the-irs-and-your-organization/

[2] Examples of in-kind contributions include donations of: equipment and supplies, such as food and supplies from pet stores, canned and non-perishable food drives, and new or used fridges or televisions; volunteers helping to pick up trash, plant trees, prepare food, or perform other forms of labor that an average person can perform with no to minimal training or prior experience; donated professional skills that require training and prior experience, such as event planning, website design, or preparing specialized financial documents. As used in this paper, I consider contributions of services by board members to be a part of a board member's duties to the organization, such as a board member who is a practicing lawyer also providing pro bono legal services to the organization, and so I do not anticipate such contributions to receive a valuation. I do not consider contributions of goods from a board member to be part of their duties, and thus I anticipate such contributions to receive a valuation. I acknowledge that both of these situations may vary by context, e.g., legal, organizational.

data has limited reliability, and it is not obvious that financial valuation is the sole useful valuation for nonprofits and for those who want to improve organizational efficiency.

Without an understanding of the complete portfolio or set of resources, critical information is lost that is necessary to appropriately evaluate programs and identify how to replicate, reproduce, and deliver them efficiently and at scale. The lack of data on resources "consumed" by programs in turn implies numerous analytic challenges. Evaluation efforts and studies suffer from biased predictions because incomplete data can yield inaccurate estimates and even scales of analysis, e.g., showing a negative effect per individual when the effect is actually positive at a scale of ten individuals. Limited generalizability and comparative ability across organizations affects considerations for replicating programs and understanding endogenous and exogenous characteristics necessary to achieve maximum program efficiency and possible social impacts. Proposed causal mechanisms such as theories of change or logic models are incomplete in ways that may question their validity and utility in program design, monitoring, evaluation, accountability, learning, and other uses. Analyses are frequently limited to a microscopic pool of organizational outcomes, i.e., financial, which have issues in their utility for nonprofits.

I focus on two expressions of organizational efficiency, (1) the efficient allocation of resources to maximize their utility in achieving outputs and (2) the technical efficiency in transforming inputs into outputs such that "it is technologically impossible to increase any output and/or reduce any input without simultaneously reducing another output and/or increasing one other input" (Ruggiero, 1996, p. 553; Brueckner, 1982; Coupet & Berrett, 2019). Both of these perspectives have been applied to discussions of public sector and nonprofit efficiency for decades, from Samuelson's (1954) exploration of minimizing government inputs to maximize collective outputs/benefits to explorations within the nonprofit sector (ex. Callen, 1994; Coupet & Berrett,

2019; Fernández-Blanco & Rodríguez-Álvarez, 2018), and they can be seen reflected in popular tools used by nonprofits. For example, a logic model is a tool commonly used by nonprofits to identify the intended relationship between available resources, planned actions, and desired outcomes; the first step, inputs, requires nonprofits to identify the "human, financial, organizational, and community resources" (W. K. Kellogg Foundation, 2004, p. 2) that the organization believes are necessary to achieve outcomes. Nonprofits and scholars alike rely on information about resources consumed during program implementation to evaluate the extent to which that program achieved the desired outcomes (Bingham & Felbinger, 2002).

The lack of data on in-kind contributions used to deliver programs present an analytic challenge and implies that organizations likely overconsume or underreport necessary resources, thus failing to capture true costs for program-related benefits. For example, studies of nonprofit revenue stream stability and effects on organizational health are likely to produce biased estimates because they fail to capture the relationships of underreported assets with other revenue streams. Callen (1994) finds that a nonprofit's technical efficiency and ability to attract donations of time (volunteering) both have positive effects on attracting donations of money. As financial donations are among the most inconsistent revenues for nonprofits, improved data on underreported assets (e.g., donations of time) can help scholars and nonprofits alike identify how to further stabilize revenue streams and improve organizational health by creating more accurate estimates and evaluations (Tuckman & Chang, 1991).

Recent scholarship has also begun to question the reliance of overhead or program expense ratios when exploring nonprofit efficiency (Coupet & Berrett, 2019). These ratios capture the relationship between programmatic expenses, non-programmatic expenses related to general management (and, sometimes, fundraising), and total expenses, where 0.8 is the minimum

recommended value for a nonprofit to be considered efficient or effective (Kioko & Marlowe, 2016). Because the ratio is relatively simplistic, nonprofits can meet or surpass that threshold depending on how they present or adjust specific financial line items, especially in-kind contributions and other resource inputs. The ratio imposes a generalized understanding of expenses or consumed resources by implying that financial valuations can sufficiently distinguish between the contributions of different programmatic inputs towards achieving observed outputs. In reality, financial valuations may not accurately affect the contributions of inputs towards those outputs, i.e., a more expensive input may contribute less to the output than does a less expensive one. In addition, the ratio is highly reliant on financial valuations as proxies for performance without actually capturing or addressing how a nonprofit would itself consider its performance (Coupet & Berrett, 2019). Such critiques suggest that reliance on financial data alone to understand nonprofit performance imposes unintended limitations, and more meaningful analyses would benefit from inclusion of non-financial valuations of organizational inputs and outputs that reflect the nonprofit's own values.

For these reasons, in-kind contributions as organizational assets[3] are inconsistently valued, underrepresented, and underreported in current nonprofit data sources, and compose an *excluded class of data* in the field (hereafter referred to as "underreported assets"). Efforts to address the data gap face challenges such as conflicting reporting and valuation guidance identified earlier (Brown & Zahrly, 1989; Cordery & Narraway, 2010; Cordery et al., 2013; Einolf & Yung, 2018; Gordon et al., 2010). Such efforts must integrate existing theories and test data on in-kind donations to assess their explanatory power to studies of nonprofits.

[3] Assets and resources will be used interchangeably in this manuscript. Distinctions made by authors will be identified as needed. Overall, assets will be considered a type of resource controlled by the firm, in line with Barney's (1991) framing.

I frame this idea of valuing underreported assets as a process of formalization in nonprofits. Despite the rich literature on institutionalization and formalization among nonprofits, underreported asset valuation presents an underexplored context where those two phenomena have a diminished presence – these is questionable institutionalization and variable to minimal formalization. While there is limited research on the formalization of valuing underreported assets, numerous organizational and management theories study, explore, or assume the effect of resources on organizational outcomes and competitive advantage. Scholars often apply such theories in combination to better explore recent advancements in firm strategy, such as relationships between human resource management systems and firm resources (ex. Andersén, 2019; Chadwick et al., 2015; Chapman et al., 2018). Of these theories, resource-based theory (RBT) is among the most generalized and flexible. RBT provides numerous approaches to analyzing and understanding the effect of internal organizational characteristics on organizations' competitive advantages, which include tangible and intangible assets, e.g., as the underreported assets of in-kind contributions of goods (tangible) and services (tangible or intangible; Barney et al., 2011; Hoskisson et al., 1999; Wernerfelt, 1984). Whereas the financial valuation for tangible goods can be obtained from a store or online retailer, intangible assets may prove more difficult to value due to inconsistent market prices, lack of transparency, and variations in individuals' willingness to pay for or accept a donated service. Two approaches from RBT provide additional insights around understanding intangible assets: first, RBT provides an explicit framework linking firms' production functionalities (capabilities) and intangible assets; second: provides a multidisciplinary approach that emphasizes embedding clear definitions of intangible assets within theory (Coyne, 1986; Hall, 1993; Molloy et al., 2011).

This paper presents an effort to remedy the lack of formalization in valuation of underreported assets by introducing a taxonomy to categorize in-kind contributions. The taxonomy draws upon RBT alongside relevant theories in nonprofit research as well as professional and legal requirements for the valuation of goods and services. By having a tool that sheds light on donated goods and services and on a nonprofit's full resource portfolio, organizations and researchers can gain superior data to understand programmatic effects and the necessary components for replication. For example, complete awareness of program implementation costs provides an opportunity to reevaluate the theory of change and other core elements of program design that factor into future implementations, scaling efforts, and the like; in this way, improving cost evaluations can enhance program evaluations and organizational learning. These findings can also yield better insight into topics such as market complexity, social capital production, and organizational health. The taxonomy can thus expand empirical studies while also testing and advancing theories such as transaction-cost economics and resource-dependence theory, as well as enhance the programmatic efficiency of all nonprofits that implement it.

Rationale for and challenges in valuing underreported assets

The lack of information on the effects of underreported assets on nonprofit performance (ex. building organizational capacity, reducing process-related costs) imposes a fundamental limit on understanding production processes among nonprofits. The ensuing limits are vast, ranging from true costs for producing social capital in communities to anticipating consequences from shocks to resource streams to producing limited information for replication programs at scale. Nonprofits that do not value underreported assets will face challenges in evaluation, learning, and maintaining organizational and programmatic performance, let alone in improving performance. Examples of how underreported assets present both challenges and opportunities for insights include how

nonprofits value volunteers, understand their revenue mixes and consequences for longevity, and analyze true programmatic expenses and cost efficiencies. Further detail on each example is available in Appendix: Current Challenges & Opportunities.

Underreported assets limit understanding of how different organizational characteristics such as the resource portfolio affect organizational and programmatic performance. As a result, they also limit the ability to apply theory to understanding intra- and inter-organizational relationships. I identify two primary challenges in addressing this issue: (1) understanding what resources are received & consumed and (2) assigning value & meaning to resources for the analysis of organizational efficiency & efficacy.

Challenge 1: defining resources

Challenge 1, understanding resources, more accurately refers to the challenges in understanding how different organizational characteristics, in particular resource portfolios, affect organizational performance. Part of this challenge rests on inconsistencies in defining resources, resulting from differences between and gaps among the organizational theories that dominate nonprofit studies. In particular, they do not provide:

- Consistent definitions of resources (institutional entrepreneurship: Hardy and Maguire, 2008; resource dependence theory: Malatesta & Smith, 2014, and Pfeffer & Salancik, 1978),
- Treatment of resources as multidimensional variables for analysis (organizational ecology, compare: Carroll & Swaminathan, 2000, and Dobrev et al., 2001),
- Or detailed conceptual constructions (transaction cost economics in Malatesta & Smith, 2014, and Williamson, 1981).

Lack of consistency hinders analyses, classification, and understanding of the elements within the universe of underreported assets. For example, institutional isomorphism's foundational article explicitly identifies the importance of resource supply and sources in theories of organizational and field-level predictors of isomorphic change (DiMaggio & Powell, 1983). Yet, nonprofit studies that refer to institutional isomorphism consistently look to other theories, such as resource dependence, for the analysis of resource-related variables (e.g., AbouAssi & Bies, 2018; Schmid et al., 2008; Verbruggen et al., 2011). This practice limits the study of resources because it requires an external definition, and those conceptual constructs may not align with core elements of a given theory and have their own limitations (as in the case of resource dependence theory). Resource acquisition and related variables thus become a target of study while overlooking resources as a unit of analysis.

Other approaches that are more explicitly oriented towards resources rather than how they are acquired still struggle with this question of definition. Writing about resource-based theory, Priem and Butler (2001) note that "virtually *anything* associated with the firm can be a resource" (p. 32, emphasis in original). This contributes to validity issues and challenges results across RBT studies, particularly when definitions are not consistent among authors (Molloy et al., 2011). Efforts to enhance testing of RBT note the importance of clearly defining and establishing constructs before engaging in experiments, but eschews an explicit definition in favor of providing guidance on what a definition should encompass (Molloy et al., 2011). In this regard, analysis and valuation of underreported assets may be easier because one can fall back to definitions provided in legal or formatting guidance (ex. IRS, FASB). Unfortunately, such definitions may not always overlap, and their valuation methods can even be contradictory to or exclude others.

Challenge 2: assigning value to resources

Even when a consistent definition is determined, challenge 2 remains: the process of assigning meaning and value to non-financial resources is itself inconsistent. This results in the questionable inclusion and frequent exclusion of underreported assets from common sources of nonprofit data.

Challenge 2 is complicated by competing requirements for nonprofits as to whether they should even attempt assigning value to in-kind contributions or not. The typical, annual presentations of nonprofit financial and programmatic information are financial statements, IRS tax filings, and annual reports. Table summarizes the requirements for all three. For each, I explore the ways in which one common underreported asset, volunteering, is presented. The valuation of volunteer time has questionable utility. Cordery et al. (2013) identifies arguments such as their contribution "is beyond measure and valuing is unhelpful" (p. 3). Others note that current valuation methods face implementation barriers, and data constraints coupled with rejection of external guidance create additional impediments (Callen, 1994; Cordery et al., 2013). At the same time, volunteers are a highly visible underreported asset, which is why I use them as an example.

Table 1. Comparative summary of annual report requirements for in-kind contributions

Source of Information	**Formalized valuation process**	**Are in-kind contributions required to include?**
Annual Report	No	No
IRS 990 Tax Filing	Yes	Partial/some
Annual Financial Statement	Yes	Yes

Annual Reports

Organizations' annual reports may identify in-kind contributions, but their inclusion is neither standardized across organizations nor follows fair market valuation methods, and so annual reports are not a strong source of nonfinancial information (Gordon et al., 2010). After reviewing numerous annual reports, I have observed that volunteer time may be presented as total number of received volunteer hours, number of people who volunteered, both, or not be presented at all. Instead of using a standardized set of metrics across all nonprofits, annual reports are also more likely to use metrics that the nonprofit identifies as useful, metrics that it adopts due to isomorphic pressures such as donor-required data (coercive), metrics included as a part of dominant practices in its service area (ex. industry and geography; normative), and metrics identified through efforts to emulate other organizations perceived as top performers (mimetic).

The inclusion of financial values for underreported assets can also depend on whether the organization has implemented a methodology in the style of Feeding America or look to sources for pre-established calculations, such as the calculations for a given dollar value of volunteer time provided by the Urban Institute (ex. see table 6 in McKeever & Pettijohn, 2015) or by Independent Sector & the Do Good Institute (Independent Sector, 2022). Once again, while these methodologies provide a broad and highly generalized valuation approach, they fail to recognize that not all volunteers' contributions to the organization or program are equal. Even when focusing solely on financial valuations and not a nonprofit's own subjective valuations, it may be inappropriate to set one value for any volunteer's donated hour instead of distinguishing by what was achieved or accomplished during that hour.

Annual Financial Statements

The Financial Accounting Standards Board (FASB) sets the standards for generally accepted accounting principles, which are adhered to when creating financial statements. FASB defines in-kind contributions[4] as unconditional, voluntary transfers of assets from one entity to another, and includes materials, supplies, intangible assets, and services (FASB, 2008, section 5). These contributions must be reported at their fair market value, consisting of "quoted market prices" for those or similar assets (FASB, 2008, section 19). Any nonprofit that produces an audited financial statement should thus be providing the statement preparation team with the information necessary to assign a fair market value to in-kind contributions during the period in question.

Reporting failures in financial statements of just one underreported asset, volunteer time, have been explored in multiple countries. Only 3% to 8% of respondents recognize donated time in their annual financial statements, in studies conducted in New Zealand, Canada, and the U.S. (Adams et al., 1989; Mook et al. 2005; both as cited in Cordery & Narraway, 2010).

Tax Filings

Nonprofits submit annual IRS tax filings via IRS Form 990 as part of the requirements to maintain exemption from federal income tax; versions of the form vary according to organizations' gross receipts and total assets (USDT, 2019). The IRS Form 990 instructions explicitly state that any reported values for contributions should include "neither donations of services…nor donations of use of materials, equipment, or facilities" (USDT, 2019, p. 57). The form instructions indicate that

[4] Financial Accounting Standards Board (FASB) guidance and IRS Form 990 instructions vary in their interpretation of "contribution" vs. "donation." The IRS instructions use the words interchangeably, whereas "contribution" is used almost exclusively in the FASB sources cited herein. For consistency, I default to using "contributions" from here on except when a cited or quoted source uses "donations." Various websites provide guidance on how the two words can be used differently, such as: https://subfictional.com/donation-vs-contribution-payment-processors-the-irs-and-your-organization/

it is optional to report such values in Part III's narrative fields but to not report them in the financial fields of that section, "even if prepared according to generally accepted accounting principles" (USDT, 2019, p. 12). Noncash contributions of goods are defined in broad terms as "contributions of property, tangible or intangible, other than money" (USDT, 2019, p. 68). The form instructions clarify that services and materials, equipment, and facilities are not considered to be noncash contributions of goods. In regards to volunteers, the IRS Form 990 Instructions only request the number of volunteers; how organizations track this number, the number of volunteer hours served, and the services provided by volunteers are all optional to report (USDT, 2019, p. 10).

Implications

The lack of consistent expectations for documenting and communicating the value of in-kind contributions yields at best a patchwork of solutions among different relational networks of nonprofits and at worse idiosyncratic methods to assign value or none at all. Rationalizing the value assignation process is a solution to this uncertainty because it formalizes a particular process within a nonprofit, the support (management) process of assigning value and perceived utility to in-kind contributions.

Proposed Solution: Taxonomy

I reframe the inconsistent valuation methodologies for underreported assets as a question of limited process formalization. Formalizing the support or managerial process of assigning value and utility to in-kind contributions of goods and services will standardize information such that data used to understand programmatic efficiency within and among nonprofits will be more consistent, more comparable across organizations and sectors, and perhaps even less resource intensive to produce in the long run. In turn, efforts to replicate program results and scale them up will face less barriers

due to informational asymmetries by allowing implementers and designers to make more appropriate and effective modifications. I define formalization as the reduction of permutations of a process that yield the same output, with the goal of codifying and standardizing the steps performed to achieve that output. I propose that this challenge is one of limited rationalization, *the formalization of core (service delivery) and support (management) processes within a nonprofit* (Dart, 2004; Hwang & Powell, 2009; Suárez & Hwang, 2013; Maier et al., 2016).

To resolve the situation, I present a taxonomy that rationalizes the valuation method by coupling existing approaches with numerous theoretical perspectives. This approach allows both for standardization across nonprofits as well as customization within nonprofits to best support their unique situations and resource portfolio mixes. I draw heavily from resource-based theory (RBT) to inform the overall taxonomy and its dimensions; for more information, refer to Appendix: Applying Resource-Based Theory. The output is intended to represent a resource inventory or profile of underreported assets (Hofer & Schendel, 1978). The inventory can be developed at the organizational level or for individual units or programs. Each approach can contribute to organizational learning and academic studies, and programmatic inventories can be aggregated to allow for cross-organizational comparisons.

The presented taxonomy is meant to be a modular analytic tool, where researchers and practitioners can implement just the dimensions that are relevant, applicable, and not cost prohibitive. The taxonomy is not meant to be a policy recommendation for how all underreported assets should be valued. Instead, it is meant to empower individuals and organizations to explore what matters and to move forward efforts to identify the true costs of program implementation and effectiveness.

I recommend two dimensions, at minimum, for practical application. Because fair market value is an existing data point and required by at least some oversight organizations, i.e., FASB, it should

always be populated. Practitioners and organizations should select at least one more dimension that is meaningful to answer specific questions or performance concerns. A comprehensive application of all eight dimensions might be time and resource intensive in a way that reduces its utility for many practitioners, especially ones working with small nonprofits. Therefore, it would be best suited for broad efforts, such as academics or researchers looking to develop holistic understanding and comparisons across nonprofits.

Overview of Taxonomy Dimensions

The proposed taxonomy links current methods, practical needs, and academic research to fill the gaps in the data and the literature.

Identified Concepts

Fair market value is a key dimension because it is the most common, existing data point related to underreported assets. I include it to enhance the utility of the taxonomy: additional dimensions add on to existing valuation processes, rather than requiring totally new processes that cannot be compared to historical data or data from other organizations. The additional concepts that compose the dimensions are chosen to reflect the broad spectrum of categories of underreported assets, their possible compositions, utility, and concerns related to underreported assets that may be part of why they are underreported, e.g., how do I categorize a broad-reaching service donation or object used by multiple programs? These concepts reflect core thinking in nonprofit studies and resource-based theory, as well as desired categories shared with me in conversation practitioners and academics.

Dimension Design

In response to Hall's (1993) framework, dimensions are not mutually exclusive. They are designed to "enable construct clarity" and to be valid, reliable, and allow for practical measures (in Molloy et al.'s MAP, 2011, pp. 1507 and 1510). At the same time, each dimension contributes to an understanding of how a resource contributes to a nonprofit's competitive advantage as indicated by the VRIN/VRIO framework values: value, rareness, degree of imitability, non-substitutability, and organization, i.e., a firm's ability to take advantage of the resource (Barney, 1991, 1995). Some resources, recommendations, and measurement approaches may be more difficult for practitioners to measure, manipulate, and implement (Priem and Butler, 2001). For example, a community food bank that is heavily reliant on donations may not be able to conduct internal studies for valuations, like Feeding America does. Similarly, not all organizations may not agree on volunteer valuation (Cordery et al., 2013). To allow for these anticipated measurement issues, and in accordance with the non-exclusive goal of the taxonomy, dimensions must be sufficiently broad to provide useful information to researchers and practitioners even when some values are not ascertainable. At the same time, each dimension draws from existing research and standards. This is intentionally done to contribute to the cross-disciplinary and -context comparative utility of the taxonomy.

When possible, the dimension should be applied to the donated underreported asset. While this is the default unit of analysis, there may be situations or contexts in which it is more appropriate to parse the asset into distinct subunits, e.g., a donation consisting of multiple kinds of kitchen equipment such as stoves, ovens, and fridges. I do not recommend a particular approach for such parsing, but rather leave it to the discretion of those implementing the taxonomy. Instead, I emphasize that all analyses can be aggregated into a shared and consistent understanding, at the organizational level, of what an asset unit is. In such cases, as non-numeric dimensional values

cannot be directly summed, the taxonomy implementer should pre-determine which approach to take. Some possible approaches include: taking the average (ex. two "high priority" and two "low priority" yields "low priority"); choosing which value is most appropriate given the organization's values; selecting the most frequent value as the overall value; assigning numeric values to the values and creating a scale (ex. "tangible" = 0 and "intangible" = 1), then finding the average, median, or modal value, etc. What is critical is that the implementer maintain a consistent approach to preserve the valuations' internal validity and not compromise the utility for comparisons. Table 2 presents the dimensions and their values; the order of dimensions does not connote importance.

Table 2. Overview of Taxonomy Dimensions & Values

Dimension	**Value(s)**	**Dimension**	**Value(s)**
Fair Market Value	Financial	Temporal	Time bound Repeating time bound Ongoing
Tangibility	Tangible Intangible	Organizational Need	High priority Low priority No priority
Justification	Instrumental Expressive	Complexity	High Medium Low
Structural Level	Program Organization	Related Expenses	High Medium Low

Values are meant to be flexible and allow for modifications as needed while retaining the overarching structure to allow for broader comparisons. For example, an organization may find that High-Medium-Low is too restrictive, and fails to capture the full range of perceived related expenses. In that case, they may choose to convert the values to an interval scale, such as low = 0,

medium = 0.5, and high = 1 or introduce additional ordered entries, such as medium-high and low-medium. What must be consistent and preserved within the organization's implementation of the taxonomy is the order of values, such that the scale (ordinal, interval, or ratio) can be converted back or mapped onto the pre-assigned, existing values for each dimension (Stevens, 1946).

Market Value

Possible values: financial. This dimension captures the current fair market valuation method for in-kind contributions. It has utility for both theory and practice by connecting to existing efforts, i.e., financial valuation of contributions. It allows the organization to understand how the resource can affect its net cash flow during a given period, and thus the organization's growth overall, e.g., by reducing certain expenses or allowing it to reinvest certain funds (Hofer and Schendel, 1978). Regarding the VRIN/O frameworks, market value identifies value, both in the literal (i.e., financial) sense and in how the organization perceives an increase in its ability "to conceive of or implement strategies that improve its efficiency and effectiveness" (Barney, 1991, p. 106). This secondary notion of value is indicated by the valuation approach used by the organization, in alignment with the flexibility prescribed by FASB's guidance.

Guidance on the valuation of in-kind contributions vary by the standard-setting organization, e.g., Internal Revenue Service (IRS), Financial Accounting Standards Board (FASB). Nonprofits that have attempted to value their in-kind contributions thus have historical data regarding underreported assets. Acknowledging and including those organizations' valuations, methods, and historical data is critical to establishing longitudinal datasets despite their potential scarcity, as indicated by the lack of recognition of volunteer time in audited financial statements (Cordery & Narraway, 2010).

Fair [market] value is defined in FAS No. 157[5] as "the price that would be received to sell an asset or paid to transfer a liability in an orderly transaction between market participants at the measurement date" (FASB, 2010, section 5). In other words, the fair value of a contribution should reflect what the organization would have to pay for it at the time of contribution.

Per FAS No. 116 (FASB, 2008),

- Section 8: non-monetary and non-service contributions are to be measured at fair value;
- Section 19: organizations can report either the fair value of the service received, the asset resulting from the service, or the asset enhancement resulting from the service.

FAS No. 157 contains additional guidance for how to measure fair value, emphasizing that it should reflect market values and the price at which someone would want to sell the item (referred to as the exit price; FASB, 2010, FAS157-2 & FAS157-3). FAS No. 157 includes an acknowledgment that such approaches may not always be feasible in the nonprofit sector: "an exemption to the requirement to measure fair value if fair value cannot be measured with sufficient reliability" (FASB, 2010, FAS157–39, C21.d).

In summary, assigning a financial value to the in-kind contribution should capture market trends at the time of contribution, such as how much the average person would be willing to spend to acquire it; in cases where this is not possible to determine reliably, then a fair market value is not required. With that said, determining some sort of value floor for certain classes of contributions received by a nonprofit would provide a basis of comparison. Using values below what you anticipate a market rate would be is related to the notion of a "balanced budget," in that

[5] Please note that FASB issued an update in 2018 to Topic 820 that took effect after December 15, 2019 (FASB, 2018).

underestimating can yield surpluses in revenue or asset calculation (Kioko & Marlowe, 2016). If the organization's assets are worth more than the floor values, it is in a stronger, fiscally responsible position. On the other hand, overestimation can inflate revenue or asset calculations in a way that leaves the organization in a weaker position by acting as though it is more financial health than the reality. To reflect appropriate fiscal stewardship of organizations, unreliable valuations of contributions should be underestimates, if included at all. Because valuations will be numerical, subunits should be summed when aggregated to the level of one asset unit.

This dimension allows for increased distinction between resources as compared to solely relying on overhead or program expense ratios. But, on its own, it does not resolve the earlier critiques of how pure financial valuations fail to reflect a nonprofit's subjective characterization of the contribution of a resource to a given programmatic output or outcome.

Tangibility

Possible values: tangible, intangible. When discussing in-kind contributions, goods are inherently easier to understand than services because they are physical objects, perceivable by one's five senses. This dimension has utility primarily for theory, as distinctions between goods and services or reputation will be unimportant for most practical analyses of efficiency.

Tangible assets can represent their own category of resources, as compared to the categories of human, organizational, technical, and reputational resources (Hofer & Schendel, 1978). These latter categories are usually more appropriate to consider as intangible assets, in alignment with the definition of invisible assets as containing elements such as organizational culture, management skills, and trust (Itami, 1984/1991), people-based skills (Grant, 1991), tacit firm knowledge (Chapman et al., 2018), and intellectual capital (Kong, 2008). Hall's (1993) separate

of intangible resources into assets and competencies can be helpful here, by expanding the definition to include conceptual products such as patents and databases. Finally, Molloy et al.'s (2011) three distinctions between tangible and intangible assets can serve as the ultimate test of which value is appropriate for a contribution. The extraction distinction discussed by the authors relates to Barney's (1991) framework: a resource that cannot be easily exchanged with or extracted from its owner will not be easily imitable or substitutable. Whereas tangible assets can also be imperfectly imitable and non-substitutable, those classifications are more inherent to intangible assets as an overall group.

It is important to note that the (in)tangibility of a contribution is affected by how the firm chooses to recognize what is being donated and its effect on the organization, as recognizing either the service or the asset resulting from/enhanced by said service are acceptable points at which to determine the contribution's value (FASB, 2008). Inconsistency in how an organization recognizes contributions may introduce measurement bias, so deciding beforehand what to recognize is important.

Tangible. A physical good or measurable asset is the explicit focus or overwhelming majority of the contribution. A contribution of canned food, for example, may require transportation of the cans to the contribution site. If the main focus of the contribution is on the canned food rather than transportation or pre-sorting of the cans, then the contribution would still be considered as tangible. This dimension also includes physical objects that may not be housed at the organization's property, acknowledging that some contributions may not be a physical object in a purely physical sense. For example, a contribution of online storage space is intangible in that it consists of access or permission to a distributed resource that is virtual. At the same time, the data is stored in a physical place, and effectively replaces physical assets such as a filing cabinet or a storage room.

In this sense, the online storage is a measurable asset. Other characteristics of a tangible item may include that: it can be consumed, wear down over time or need to be renewed; it is limited in the number of people that can use it at a time; it can be readily distinguished (e.g., conceived of as separate) from whoever owns it (Molloy et al., 2011).

Intangible. An intellectual item, such as a service or skill or conceptual product, is the focus or majority of the contribution. This may require or produce a tangible item, but the main focus of the contribution is not of said tangible item. FASB 116 defines a contribution of services as one that may "(a) create or enhance nonfinancial assets or (b) require specialized skills, are provided by individuals possessing those skills, and would typically need to be purchased if not provided by donation" (FASB, 2008, section 9). Intangible assets can also include conceptual, intellectual, or virtual products, such as a database, online storage space, and patents. Ultimately, the contribution can be considered intangible if: it cannot be consumed or does not depreciate with use and over time; it is not limited in the number of people that can use it at a given time; it cannot be readily distinguished or conceived of as separate from whoever owns it (Molloy et al., 2011).

Justification

Possible values: instrumental, expressive, both (see Figure). The fundamental nature of the work of nonprofits can be justified in terms of achieving public purposes (instrumental) and expressing the values of the community (expressive; Frumkin, 2012). This dimension applies the two concepts to the level of contributions by acknowledging that a contribution can be perceived by a nonprofit as contributing to different aspects of its fundamental nature and whether it will be used towards expressive ends, instrumental ends, or a combination of the two. It has utility primarily for theory, as how a nonprofit perceives a contribution to be expressive, instrumental, or both will be unimportant for most practical analyses of efficiency. That said, it may be useful for practitioners

looking to understand the kinds of contributions generated from outreach and fundraising campaigns.

The notion of justification echoes reputational resources (Grant, 1991) or high brand loyalty as a type of resource (Hofer & Schendel, 1978). This also relates to the value perceived by the recipient organization of the contribution's ability to improve competitive advantage and reduce or neutralize threats (Barney, 1991, 1995). An organization may perceive the contribution as doing so through better accomplishing its mission or indicating values. While the donor and the recipient nonprofit may differ or disagree on which aspect of justification a contribution is meant to contribute to, this taxonomy value focuses on the recipient nonprofit's intended use.

Prakash and Gugerty (2010) note that nonprofits can be both expressive and instrumental. Accordingly, this taxonomy acknowledges that an asset can be used to further an expressive aspect of the recipient nonprofit, instrumental, or both, as seen in Figure . Tension seems to exist in the literature between these two concepts as to whether they are opposite ends of a single continuum or on distinct continua. I follow the approach of Frumkin (2012), who introduces them as "two *different* ideas about what justifies and gives meaning to the work that is carried out in the sector" (p. 26, emphasis added). Because this concept is applied at the level of individual contributions, its analysis for all contributions may shed light on how the organization identifies with each of these concepts and which is dominant.

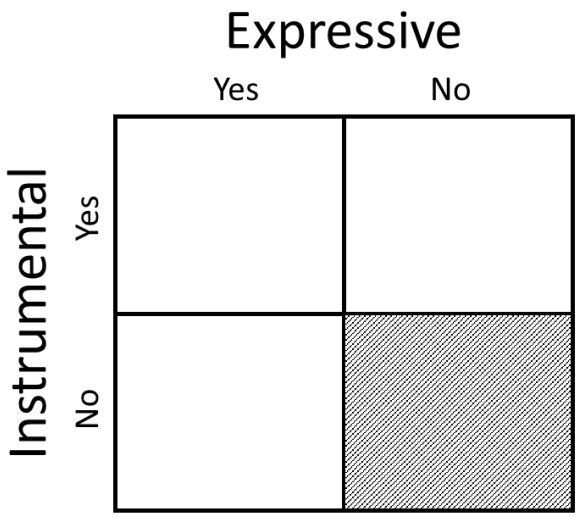

Figure 1. Proposed interactions for characteristics of justification. A contribution is assumed to always contribute or used towards an expressive purpose of the recipient nonprofit, instrumental purpose, or both.

Instrumental. Frumkin (2012) identifies the instrumental justification for nonprofits as focused on "the outcomes that are generated for society" (p. 25). To borrow from the author, the organization's actions are instruments to accomplish tasks that the entity offering the contribution views as important (Frumkin, 2012, p. 26). This notion can be operationalized in regards to contributions by considering the implications of this concept: the organization is a tool, and the contribution is meant to make that tool better. This is evaluated through concepts such as effectiveness, and an instrumental contribution is thus one that affects the organization's ability to provide services, accomplish tasks, or make progress towards some empirical, clear goal. Consider the function that served by second-hand organizations such as Goodwill or St. Vincent de Paul. Those that contribute clothing, household items, or the like to those organizations could sell them or throw them away. Instead, they provide them to this organization because they believe that this organization will appropriately allocate and make available such items to those who need them.

Expressive. Frumkin (2012) suggests that nonprofits are vehicles for the expression of people's "values and commitment through work, volunteer activities, and donations." This expressive justification for nonprofits explicitly identifies contributions as an action that individuals can take to signal values and commitment to them. To apply this concept to our taxonomy, expressive contributions signal alignment with the organization's values and a commitment to them. People who contribute to civil liberties organizations do so because they want to promote and support such values.

To allow for the treatment of both justifications as separate but interacting, valuations along this dimension should be captured in a way that captures them both, for example:

- Instrumental, expressive, instrumental & expressive
- Y-Y, Y-N, N-Y where the output reports [Instrumental]-[Expressive]
- 1-1, 1-0, 0-1 where 1 indicates the concept is met and 0 indicates it is not and the output reports [Instrumental]-[Expressive]

Structural Level

Possible values: program, organization. The scope of the contribution's intended recipient, either a part of the organization (i.e., program) or the whole, relates to how well that structure is able to leverage the potential of the resource to assist the organization in establishing competitive advantage (Barney, 1995). Scope can also affect how firms can use contributed resources, most explicitly financial, human, and organizational system resources (Hofer & Schendel, 1978).

Donor restrictions that stipulate how contributions are used may limit them based on structural level (e.g., to a specific program), and these restrictions can be temporary or permanent (Kioko & Marlowe, 2016). In an analysis of 637 RBT-related articles, 76% of them only understood

intangible assets as operating or existing at the firm level, providing evidence of a "macro bias" in RBT that ignores how intangible assets are developed and used at different levels within an organization (Molloy et al., 2011, p. 1505).

This dimension of structural level captures how nonprofit leadership can choose or are able to structure resources and generate competitive firm advantage, while also anticipating potential macro bias by acknowledging that resources restricted to programs could be structured, bundled, and leveraged by managers at that level or depth of the nonprofit, i.e., program managers instead of overall executive directors (Sirmon et al., 2011).

Any contribution that benefits a program will likely benefit the organization as a whole. This dimension is meant to consider the immediate target of the contribution, rather than the effects of the contribution over time on the organization. For this reason, a contribution that a donor restricts to a program for a limited window of time upon receipt should still be classified at the program level.

This dimension can contribute to practice and, to an extent, theory. The practical applications include, among others, identifying what resources are used by and in demand across levels of the organization and whether a contribution should be secured consistently, e.g., purchasing or contracting instead of relying on donations. Theoretical applications include, at minimum, allowing researchers to filter out contributions used by multiple programs and compare efficiencies for single programs.

Program. The contribution is intended for or benefits one specific program offered by or part of the organization. This may be explicit in the form of a restriction on the contribution, or some kind of highly targeted unrestricted net asset that, for all practical purposes, is restricted to use by one

program. Some organizations may only have one program. Nonetheless, if the target of the contribution is that program and not the organization as a whole, then it receives this classification.

Organization. The contribution is intended for or benefits the organization, either as a whole or as an umbrella entity overseeing one or more programs. The contribution is not specific to one program, and is intended for or may benefit more than one program. Most unrestricted contributions will likely receive this value, but not all, for example a stove being offered to a child-focused organization that has two programs, one for preparing meals and one for providing literacy lessons. Consider also a skilled professional who offers their knowledge to the organization, but it is only relevant to one program; the benefit is only received by the one program.

There is the possibility that an unrestricted contribution to a single program may have spillover effects on multiple programs. Because this possibility increases as the number of programs-as-intended-recipients increases, one may observe a contribution to multiple programs affecting other programs within the organization. For this reason and for the sakes of both simplicity and parsimony, the structural level of "multiple programs" is collapsed into "organization."

If the intended beneficiary is a program run by multiple organizations in partnership, it is still at the program level. If the intended beneficiary is multiple organizations, the value of "organization" should be used.

Temporal

Possible values: time bound, repeating time bound, ongoing. The importance and effect of time on organizations and their resources is noted in the literature. Understanding a firm's history and its interactions over time is a potentially fruitful yet challenging avenue of exploration for RBT scholars (Priem & Butler, 2001). It is present in resource orchestration's consideration of how "the

resource orchestration actions needed to not only survive but thrive in each stage [of an organization's life cycle] must be prioritized" (Sirmon et al., 2011, p. 1400), and thus manager's actions are affected by the life-cycle stage of the organization. This dimension provides a way to explore the relationship between time and an organization at the level of individual contributions. The style and duration of the act of contributing require different kinds of actions on the part of nonprofit managers. Furthermore, there is evidence that suggests that donating and volunteering have a complementary relationship (Yao, 2015). In other words, the temporal component of a contribution of services may be related to the likelihood of future contributions that are financial, goods, or intangible, e.g., services. This dimension has utility primarily for theory; meaningful integration of this data into practical analyses of efficiency may not be value add and could be costly or intensive. That said, temporal components of contributions can help inform practitioners and nonprofit staff, such as executive directors and heads of development, as to how to structure solicitations for certain contributions.

Whereas intangible assets are "expected to confer benefits for an undefined time frame" (Molloy et al., 2011, p. 1498) and, in this way, are not depreciable, this dimension does not focus on how long the asset is intended to last but rather the style and duration of the act of contributing. Questions of depreciation of the contribution are thus outside the scope of this dimension. The notion of imitability (Barney, 1991) is loosely found in this dimension, in that contributions that occur repeatedly or without a determined end may not be imitable and can generate sustained competitive advantage; a time-bound contribution may only generate competitive advantage for the duration of the contribution, and is all the easier to replicate or imitate.

Time bound. A time-bound contribution is discrete in that it has a clearly recognizable beginning and end. There is some aspect of finality or termination that is fundamental to the contribution, at

least at the initial time of offering. Examples may include a single instance of a contribution, a project with a clear end product, or a service provided for a predetermined period of time.

Repeating time bound. A contribution may be time bound but occur more than once or multiple times. If the repetition of the action is established at the time of offering as how the contribution will be delivered, then each unique delivery event is a time-bound contribution occurring within a larger temporal setting specified at the time of offering. This could be considered a special instance of time bound or of ongoing, but is substantively different enough to merit consideration as a separate category. The repetition or delivery cycle may be time bound or ongoing. Examples may include multiple instances of a good or service being donated or a project composed of multiple projects, each of which could be considered a unique contribution otherwise.

Ongoing. An ongoing contribution is continuous in that it does not have a clearly recognizable end or termination. There is a distinct lack of finality to the contribution, at least at the initial time of offering. Examples may include a commitment to provide a contribution over time, a project that does not have a deadline or end product that would indicate its end, or a service that is provided on a continuing basis such as maintenance work.

Organizational Need

Possible values: high priority, low priority, no priority. Nonprofits are mission oriented, yet a contribution that the organization deems to be of high priority or essential may not necessarily be tied to its core mission. This is apparent when considering the notion of a resource's value (Barney 1991, 1995). If competitive advantage of a nonprofit is defined as ability to pursue mission or be financially viable as compared to its peers, then a contribution that clearly does so will be highly prioritized by the organization. For this reason, this dimension has utility primarily for practice, as

the subjective valuation is likely to be used for internal analyses. I anticipate difficulties when replicating how one nonprofit implements this dimension at another nonprofit, and the potential conflict between external and internal validity will undermine its use in some theoretical studies. With that said, some researchers may be interested in understanding this dimension.

Organizational need relates to whether a resource can be substituted. A resource that the organization prioritizes highly, is not easily accessible or possessed, and provides an idiosyncratic type of benefit for the organization, cannot be easily substituted (Barney, 1991). This could be a resource that competitors have access to but the organization does not (the inverse of imitability), or, if it has few barriers to access for both the organization and its competitors, then it is imitable (Barney, 1991, 1995). Resources that are not commonly accessible to or controlled few competitors are, by definition, rare (Barney, 1995). Contributions of resources that are in high demand for the organization are likely to be valuable and may also be rare, imperfectly imitable, and non-substitutable.

That said, nonprofits do have an obligation to pursue activities related to their missions, and more broadly, their tax-exempt purposes. Per the Internal Revenue Code's (I.R.C.) regulations, organizations with 501(c)(3) status must meet both the organizational test and the operational test, or else it may lose its federal tax-exempt status (I.R.C.Reg. Sec. 1.501(c)(3)-1(a)(1)). Within the organizational test, the nonprofit's articles of organization should not allow for the organization to engage in "activities which in themselves are not in furtherance of one or more [tax-]exempt purposes" in a substantial way (I.R.C.Reg. Sec. 1.501(c)(3)-1(b)(1)(i)(b)). This is reiterated within the operational test, which similarly specifies that the nonprofit will lose its status "if more than an insubstantial part of its activities is not in furtherance of an exempt purpose" (I.R.C.Reg. Sec. 1.501(c)(3)-1(c)(1)). An organization's drift from its mission is made all the more problematic by

this threat to its tax-exempt status. For these reasons, the notion of mission is centered in the descriptions of values for this dimension, but the values are not limited to how well a contribution aligns with the organization's mission.

High priority. The contribution immediately relates to the organization's ability to achieve its mission or accomplish the intended goals of the program. If the contribution was not received, then the organization would have to find some other way of receiving or source for providing it. The contribution is a must-have or core item for the given organizational level.

Low priority. The contribution relates to the organization's ability to achieve its mission or accomplish intended program goals in a broad sense, at a high level, or indirectly. If the contribution was not received, then the organization may try to secure it or something analogous in the future. At the present moment, it is not high priority. This may be an item that is not deemed as mission-critical because of limited organizational or programmatic capacity to provide it, but it is known of and desired in some way by the recipient. This can include things like future programming that were waiting on possible grants. The contribution is a nice-to-have item for the given organizational level.

No priority. The contribution does not clearly relate to the organization's ability to achieve its mission or accomplish intended program goals. This may be something that the organization or program did not consider previously to act upon, act towards, or request, or an item that the organization has not anticipated being offered. The item is outside of the (previous) scope for the given organizational level.

Complexity

Possible values: high, medium, low. This dimension synthesizes elements of the classifications of rare, imperfectly imitable, and non-substitutable resources. Contributions of resources that cannot be replicated, produced, or accessed by the average person are likely to be rare if this holds for the nonprofit's competitors (Barney, 1995). If such resources cannot be easily accessed, then they are imitable (Barney, 1995). Finally, if those resources or the benefit that they provide cannot be replaced by something else accessible by the nonprofit, they are non-substitutable (Barney, 1991). Regarding skills, given the difficulty in extracting intangible assets from their owner (Molloy et al., 2011), higher levels of complexity associated with the contribution would suggest that the skill is less general or accessible. This is extrapolated and applied to tangible donations of goods in the values below, such that complexity of contributions is a component in valuing goods as well as services. Contributions which are people dependent and protected in law (Hall, 1993) are likely to be classified as more complex in this taxonomy.

Complexity also ties into the notion of professionalization, which encapsulates efforts by members of a profession to define their work, place controls on its production, and legitimate it (DiMaggio & Powell, 1983). In the nonprofit sector, professionalization places experts in roles of positional authority, increase the qualifications for volunteers, results in more paid employees, and increases the importance of formal education credentials (Maier et al., 2016, p. 71, citing in order: Salamon, 1999; Lundström, 2001). Goods or services produced by professionalized roles would therefore be complex because they require a high degree of training to produce or advanced qualification or education to deliver, with years of experience substituting for education in some cases.

This dimension has utility for both theory and practice. It allows practitioners to identify a less-quantified element that affects inputs and potential threats to operations, e.g., a program that

requires multiple complex resources could be more at risk to supply chain shocks. Researchers and practitioners alike could analyze variations across programs with similar outputs but differing levels of resource complexity as part of efforts to scale impact.

High. Barriers to accessing the requisite knowledge, skills, or abilities necessary to produce or deliver the contribution are high. The contribution consists of a specialized skill (FAS No. 116: see FASB, 2008) or a non-standard type of service for the organization, such as leadership or professional skills (Einolf & Yung, 2018). Goods or physical assets that are part of the contribution are not reasonably or readily accessible via commercial or market sources. The custom production of a tangible or intangible item (physical object, website), a high-quality handmade good, or an experienced tradesperson providing work to the organization are just a few examples of a skilled contribution.

Medium. Barriers to accessing the requisite knowledge, skills, or abilities necessary to produce or deliver the contribution are low. The contribution requires some degree of specialization to produce or deliver. The average person can access that specialization, but is not necessarily guaranteed to have it. For example, most people are familiar with the core functionalities of digitized spreadsheets such as Microsoft Excel or Google Sheets. Not all people are familiar with how to use functions within them to produce financial estimates, but most people could learn with training.

Low. Barriers to accessing the requisite knowledge, skills, or abilities necessary to produce or deliver the contribution are negligible or nonexistent. The contribution consists of a general skill, standard type of service, or a good or physical asset that is reasonably or readily accessible via markets. With minimal training, the average person off the street would be able to demonstrate competency at the skill, such as organizing documents or preparing food. The good can be acquired

or produced with minimal effort or cost (relative to similar goods) by the average person. For example, the primary barrier to purchasing items for a canned food or toy drive is financial and not tied to education, professional experience, credential, or the like.

Related Expenses

Possible values: high, medium, low. An organization must "be organized to exploit its resources and capabilities" (Barney, 1995, p. 56). Organizational structure can thus yield varied costs related to the nonprofit's ability to exploit, incorporate, and use the contribution. How the organization considers and recognizes anticipated expenses related to accepting a contribution indicates how well the organizational structure is designed to exploit that particular contribution. For example, costs related to depreciation and maintenance of an asset, how to control access if it is rivalrous, and difficulty with which it can be extracted from or exchanged with its owner (Molloy et al., 2011) should be considered for tangible and intangible contributions. This dimension allows for the introduction of a cost dimension to resource mobilization and deployment, contributing to institutional entrepreneurship and to RBV and resource orchestration discussions of how the capacity or capability of a firm to deploy resources is itself dependent on resources. Unlike fair market valuation, this dimension captures a financial dimension of resources themselves, rather than identifying that a resource is financial, as in Hofer and Schendel's (1978) classification of resources. It has utility for both theoretical and practical analyses of efficiency and cost effectiveness, as it is somewhat comparable to overhead costs (at the individual resource level) and enhances existing financial valuation efforts.

High. The expenses that will likely be incurred if the organization accepts the contribution are large in value or quantity. They may be significant to or prohibitive for the organization and cause serious concern about the organization's ability to accept the contribution.

Medium. There are some likely expenses associated with the contribution, and the value or quantity are neither negligible nor generate serious concern for the organization's ability to accept the contribution. These are costs may or may not be manageable, depending on the other expenses or budgeted costs that the organizational level is committed to and facing.

Low. The likely expenses are small in value, low in quantity, or negligible to the organization. These are expenses that may already occur as a part of normal business operations.

This dimension refers to expenses, or the consumption of the organization's internal resources, related to the contribution, such as financial, physical, or nonphysical assets. Such expenses or costs may be related to personnel (e.g., a number of staff members, dedicated staff members, some amount of a staff member's time), tangible or physical assets (e.g., a computer, paint, clothing), or intangible or nonphysical assets (e.g., money, access to computer systems or online storage, digital files or information). Components of this dimension include (but are not limited to):

- Costs to review or evaluate the contribution before incorporating it into the organization or program.
- Costs associated with incorporating and/or maintaining the contribution as a part of the organization or program.
- Costs associated with the consumption or use of the contribution as a part of furthering the organization's mission or accomplishing intended program goals.

These costs may be directly or indirectly related to the organizational level. Direct costs are when the consumption of resources can be explicitly tied to the "production of a good or service" (Finkler et al., 2016, pg. 127) or traced to part of the organization, such as a program (Kioko & Marlowe, 2016, pg. 155). Indirect costs are when resources are consumed, not to directly produce a

good/service but rather to generate internal services or as a part of overhead costs (Finkler et al., 2016); they may also "apply to, are incurred by, or shared across more than one part of the organization" (Kioko & Marlowe, 2016, pg. 155).

For example, a contribution that requires access to computers or files may have low related expenses if it just requires copying a file and emailing it: a staff member can accomplish this with ease. Depending on the complexity of the task, the organization's cybersecurity, and privacy or confidentiality concerns with the data in the file, there could be large expenses associated with accepting this contribution. Clothing donated to an organization such as Goodwill or St. Vincent de Paul generate direct and indirect costs, which could include: reviewing the item, cleaning the item, adding the item into inventory, pricing the item, maintaining the item, selling the item. Each of those steps may incur their own expenses. Staff and managers may need to be paid. The organization has to acquire or maintain any equipment, systems, or tools necessary to complete a step.

What Is Now Possible

The taxonomy introduces two new ways of understanding resource-oriented questions, a practical and a theoretical perspective. A more robust and accurate resource portfolio has numerous implications for understanding true resource-related expenses and costs associated with delivering programs and an organization's capacity for or progress towards mission achievement. The dimensions provide us with proposed relationships, individually and in aggregate, that can be used to further theory and test causal relationships towards diverse outcomes. I explore both dimensions in the following sections.

Applications to Practice and Research

A practitioner or researcher that implements the taxonomy within a nonprofit will be, as a result, inventorying its resources such that a complete resource portfolio will eventually be possible for specific time periods (ex. a fiscal year or quarter). In conjunction with data about organizational and programmatic activities, the nonprofit can then use the full suite of data on costs, or resources (including financial) consumed in the process of delivering various services, to improve on current and future production processes. In regards to the nonprofit's current activities, this will increase understanding and evaluation of programmatic efficiency, enhance organizational learning and planning future modifications to activities, and make true replications easier to achieve and implement. If the taxonomy is implemented during the lifecycle of a program implementation, then a program-specific resource portfolio will be constructed that will contribute to formative and summative evaluations in addition to cost tracking. This will also strengthen the claims of causal relationships between particular resources as inputs, produced outputs, and observed outcomes. The increased reliability of such calculations could enable nonprofits to generate summary performance metrics that are more meaningful and accurate than overhead or program expense ratios.

For a nonprofit looking to the future to reproduce programs or activities at scale, having a complete understanding of resource costs will allow it to identify potential challenges or limits and, in turn, potential resolutions. This can include sourcing alternatives to assets that are geographically limited, working with suppliers to stabilize supply chains and increase production or access to necessary resources, and increasing fundraising and development efforts to anticipate new costs associated with the increased scope of service delivery. A complete resource portfolio can be used to explore substitution effects from replacing expensive inputs with more affordable ones, using

less rare and more common resources, and other possible modifications to the service delivery and production lifecycle of the nonprofit and its programs.

In response to earlier critiques of purely financial valuations, the taxonomy allows nonprofits to determine the perceived value of a contributed asset. The assignment of values according to each dimension results in a multi-faceted valuation of an underreported asset that reflects how the organization perceives the utility of and potential contributions from that asset. At the same time, the taxonomy facilitates comparisons using historical data of financial valuations, so that longitudinal studies within and among nonprofits are possible even when the taxonomy cannot be applied to previously received in-kind contributions. Similarly, not all dimensions may be applicable in all situations, e.g., due to resource constraints such as staffing limitations or uncertainty about how to value an asset according to a particular dimension. If enough assets are coded using the taxonomy, then an organization or scholars could predict likely values to fill in blanks in the data using the other dimensions, escaping the limits of solely using market valuations to determine an asset's missing worth. The multi-dimensional predictive utility could then be expanded upon via communal datasets shared among nonprofits, although these would admittedly diminish the nonprofit-specific subjective valuation by standardizing explicit values of underreported assets.

Analyses and calculations generated using the taxonomy can also expand existing theories. For example, in what is termed the "nonprofit starvation cycle," potential donors use overhead or program expense ratios to approximate what percentage of donations go to programming and (presumably) yield social impact (Lecy & Searing, 2015). Because a nonprofit's spending on overhead is equated with a lack of spending on producing social impact, overhead expenditures have been shown to be steadily declining over time in an effort to appeal to and appease donors

and watchdog agencies, despite the resulting negative outcomes on organizational development (Krause et al., 2019; Lecy & Searing, 2015). As noted by Coupet & Berrett (2019), these ratios are ultimately proxy measures of the returns to social impact from a donor's contribution. Taxonomy-derived analyses can yield more accurate performance measures that are not proxies while increasing the validity with which a nonprofit can attribute a donor's contribution to an observed impact, and the reliability of such efforts across programs and organizations.

Applications to Theory & Propositions

In a separate analysis of impact evaluation and rationalization, I propose a definition of impact evaluation that addresses both common and salient elements found across published definitions:

> Impact evaluation is a systematic process to identify the relationship between programs and their effects through collecting evidence and making claims about the relationship between program-specific activities and observed changes in the world.

Implicit in this and many other definitions is that impact evaluation aims to understand how process *inputs* are consumed through activities to yield effects and observable changes. Such inputs include the entire conceivable universe of tangible and intangible resources used to deliver services across sectors, not just cash and financial resources but also food and consumable supplies, staff capacity, organizational reputation and advocacy, equipment for delivering and producing supplies, and more. As defined here, impact evaluation is not sector specific. Rather, it is a process exploring programs and possible effects, where programs are ways to transform inputs and resources into outputs and outcomes. Impact evaluation, thus, is directly affected by an organization's understanding of the resources it consumed through activities to produce outputs and deliver services and programs. I refer to the focus on how well programs turn inputs into

outcomes as efficacy and outcome-oriented impact evaluation. I refer to the focus on how much it costs programs, in financial terms, to produce outcomes as (cost) efficiency and cost-oriented impact evaluation. Increasing the accuracy of data on resources used for inputs would improve impact evaluations and could lead to greater success in achieving missions and on understanding programmatic and organizational (in)efficiencies.

For this reason, I present impact evaluation as a prime example of how the dimensions of the taxonomy can enhance theoretical understandings and generate new proposed relationships, both individually and all together. The following propositions assume a state of *ceteris paribus*, or, all else being equal, such that they hold constant variables such as organizational size, age, observed outcomes from programs and activities, etc. I present each of them in relation to the effects of the taxonomy on the overall resource portfolio and relationships between dimensions and impact evaluations for efficacy and for efficiency. For sample propositions related to the taxonomy dimensions not presented here, refer to Appendix: Additional Sample Propositions.

Sample Proposition: Market Value

> Higher fair market valuations of resources will not affect outcome-oriented impact evaluations, and they will reduce the expected "return on investment" from cost-oriented impact evaluations.

Because outcome-oriented impact evaluations look to connect inputs to observed outcomes and effects, or program efficacy, the value of those inputs should not affect the strength of their relationship to the evidence. In contrast, cost-oriented impact evaluations explore program cost efficiency, e.g., how much effect can be attributed to a single dollar consumed or a dollar's worth of input. Higher fair market valuations of underreported assets would seem to make a program more expensive and less cost effective, e.g., achieving a desired outcome requires more total input

value, alternatively, a single dollar of input will achieve less of the desired outcome. Improving fair market valuations of all inputs would benefit organizations and donors alike through identifying truly cost-effective programs as compared to programs with costs hidden within hard-to-value inputs.

Sample Proposition: Complexity

> Resource complexity will not affect the strength of causal arguments of outcome-oriented impact evaluations, and it will reduce the expected "return on investment" from cost-oriented impact evaluations.

Outcome-oriented impact evaluations focused on creating causal arguments between inputs and outcomes. The level of complexity required to produce and deliver inputs should have no effect on the relationship between those inputs and outcomes. That said, empirical studies may find that complexity of inputs relates to related factors such as project sustainability and longevity. Increasing complexity is often associated with higher acquisition and/or implementation expenses. As the complexity of resources increases, related expenses and overall program costs are likely to increase and cost effectiveness would decrease.

Sample Proposition: Related Expenses

> As related expenses for resources increase, the strength of causal arguments of outcome-oriented impact evaluations will decrease and the expected "return on investment" from cost-oriented impact evaluations will decrease.

Expenses associated with the implementation and use of resources may not always be explicit, obvious, or clearly noted. Those expenses may also be inconsistently observable and measurable, meaning that the nonprofit would know that the resource is expensive for multiple reasons but

without concrete evidence for each. As the percentage of those resources increases, the understanding of how to transform inputs into outcomes becomes more and more opaque. The requisite resource portfolio becomes both larger and harder to fully capture, reducing the strength of causal associations between or attributions of inputs to observed outcomes. On the other hand, because overall input requirements, i.e., expenses, are increasing while observed outcomes remain the same, a single dollar of inputs yields less effect, reducing cost effectiveness.

Sample Proposition: Market Value, Complexity, Related Expenses

I anticipate that fair market valuation will be correlated with complexity and related expenses, but complexity and related expenses may not be correlated with each other. I also expect inputs of high complexity and high related expenses to be less common (ex. a credential or training required to operate complex machinery), but that those inputs will have a high fair market value. Furthermore, as the set of inputs increases across all three dimensions, total costs associated with the input will increase. This will drive down the "returns" produced by a single dollar of input, even if the original calculation was artificially small because of incomplete information. Synthesizing across the three sample propositions here yields a new proposition,

> The expected "return on investment" from cost-oriented impact evaluations will decrease as inputs increase in their fair market valuations, complexity, and related expenses.

As noted in the previous sample propositions, I do not anticipate changes in a resource's fair market valuation or complexity to alter the strength of a proposed causal relationship between that resource-as-input and the intended outcome. Related expenses are more likely to moderate the strength of such relationships and reduce confidence in their validity. I synthesize this information and across the three sample propositions to suggest that

The strength of causal arguments of outcome-oriented impact evaluations will decrease as inputs increase in related expenses, regardless of their fair market valuations or complexity.

Conclusion

Understanding nonprofit program efficiency and performance currently relies on financial valuations of consumed resources, with few alternative approaches that establish a basis for comparison across programs as well as organizations. Existing approaches often exclude underreported assets, and their presence and presentation within common sources of nonprofit data are vastly inconsistent. As a result, the relationship between programmatic costs and outcomes is incorrectly understood and faces serious concerns of bias in findings. The proposed taxonomy combines elements of strategic management and nonprofit management research to create a resource-oriented approach for valuation that incorporates existing methods, allows for subjective valuations by nonprofits, and creates a multi-dimensional understanding of a nonprofit's contribution and utility. Taxonomy dimensions can be applied in a modular fashion and on an as-needed basis to allow nonprofits to analyze and determine utility in accordance with their own organizational capacity, creates a standard platform for comparison, and provides both practitioners and scholars with data for evaluation and learning that is meaningful within and across nonprofits. In doing so, it creates shared foundations of understanding and definitions that can contribute to cross-theoretical studies, empowers organizations to make more strategic and informed decisions, and contributes meaningfully to increasing awareness of the true costs of social services and social change.

Future studies could apply the taxonomy in numerous settings. For example, particular subsectors of nonprofits may have consistent valuations because of legal or professional requirements as to

how to deliver a service. The taxonomy should also be used to generate data for expanding upon existing resource-oriented studies of the nonprofit sector, such as those reliant on transaction cost economics or resource dependence theory. The relationships between and across dimensions should also be understood. Exploring such relationships will yield important insights that will contribute to creating predictive approaches to compensate for blanks or missing data when valuing underreported assets. In doing so, the overall utility of the taxonomy and related data will benefit practitioners and scholars alike.

References

AbouAssi, K., & Bies, A. (2018). Relationships and resources: the isomorphism of nonprofit organizations'(NPO) self-regulation. *Public Management Review*, *20*(11), 1581-1601.

Adams, J. B., Bossio, R. J., & Rohan, P. (1989). *Accounting for contributed services: survey of preparers and users of financial statements of not-for-profit organizations*. Financial Accounting Standards Board of the Financial Accounting Foundation.

Andersén, J. (2019). Resource orchestration of firm-specific human capital and firm performance—the role of collaborative human resource management and entrepreneurial orientation. *The International Journal of Human Resource Management*, 1-33.

Barney, J. B. (1991). Firm Resources and Sustained Competitive Advantage. *Journal of Management*, *17*(1), 99-120.

Barney, J. B. (1995). Looking inside for competitive advantage. *Academy of Management Perspectives*, *9*(4), 49-61.

Barney, J. B., Ketchen Jr, D. J., & Wright, M. (2011). The future of resource-based theory: revitalization or decline?. *Journal of management*, *37*(5), 1299-1315.

Bingham, R., & Felbinger, C. (2002). *Evaluation in practice: A methodological approach* (2nd ed.). CQ Press.

Brown, E. P., & Zahrly, J. (1989). Nonmonetary rewards for skilled volunteer labor: A look at crisis intervention volunteers. *Nonprofit and Voluntary Sector Quarterly*, *18*(2), 167-177.

Brueckner, J. K. (1982). A test for allocative efficiency in the local public sector. *Journal of public Economics*, *19*(3), 311-331.

Callen, J. L. (1994). Money donations, volunteering and organizational efficiency. *Journal of Productivity Analysis*, *5*(3), 215-228.

Carroll, D. A., & Stater, K. J. (2009). Revenue diversification in nonprofit organizations: Does it lead to financial stability?. *Journal of public administration research and theory*, *19*(4), 947-966.

Carroll, G. R., & Swaminathan, A. (2000). Why the microbrewery movement? Organizational dynamics of resource partitioning in the US brewing industry. *American journal of sociology*, *106*(3), 715-762.

Chadwick, C., Super, J. F., & Kwon, K. (2015). Resource orchestration in practice: CEO emphasis on SHRM, commitment-based HR systems, and firm performance. *Strategic Management Journal*, *36*(3), 360-376.

Chapman, E. F., Sisk, F. A., Schatten, J., & Miles, E. W. (2018). Human resource development and human resource management levers for sustained competitive advantage: Combining isomorphism and differentiation. *Journal of Management & Organization*, *24*(4), 533-550.

Cordery, C. J., Proctor-Thomson, S. B., & Smith, K. A. (2013). Towards communicating the value of volunteers: lessons from the field. *Public Money & Management*, *33*(1), 47-54.

Cordery, C., & Narraway, G. (2010). Valuing volunteers: expanding the relevance and reliability debate. *Australian Accounting Review*, *20*(4), 334-342.

Coupet, J., & Berrett, J. L. (2019). Toward a valid approach to nonprofit efficiency measurement. *Nonprofit Management and Leadership*, *29*(3), 299-320.

Coyne, K. P. (1986). Sustainable competitive advantage—What it is, what it isn't. *Business horizons*, *29*(1), 54-61.

Dart, R. (2004). Being "business-like" in a nonprofit organization: A grounded and inductive typology. *Nonprofit and voluntary sector quarterly*, *33*(2), 290-310.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American sociological review*, 147-160.

Dobrev, S. D., Kim, T. Y., & Hannan, M. T. (2001). Dynamics of niche width and resource partitioning. *American Journal of Sociology*, *106*(5), 1299-1337.

Einolf, C. J., & Yung, C. (2018). Super-Volunteers: Who Are They and How Do We Get One?. *Nonprofit and Voluntary Sector Quarterly*, *47*(4), 789-812.

Feeding America. (2012) 2012 Annual Report. Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2013a) 2013 Annual Report. Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2013b). Financial Statements: June 30, 2013 and 2012 (With Independent Auditors' Report Thereon). Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2014a) 2014 Annual Report. Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2014b). Financial Statements: June 30, 2014 and 2013 (With Independent Auditors' Report Thereon). Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2015a) 2015 Annual Report. Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2015b). Financial Statements: June 30, 2015 and 2014 (With Independent Auditors' Report Thereon). Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2016a) 2016 Annual Report. Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2016b). Financial Statements: June 30, 2016 and 2015 (With Independent Auditors' Report Thereon). Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2017a) 2017 Feeding America Annual Report. Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2017b). Financial Statements: June 30, 2017 and 2016 (With Independent Auditors' Report Thereon). Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2018a) 2018 Annual Report. Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2018b). Financial Statements: June 30, 2018 and 2017 (With Independent Auditors' Report Thereon). Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2019a) 2019 Annual Report. Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (2019b). Financial Report: June 30, 2019. Retrieved August 19, 2020 from https://www.feedingamerica.org/about-us/financials

Feeding America. (n.d.). *Home Page*. https://www.feedingamerica.org/

Fernández-Blanco, V., & Rodríguez-Álvarez, A. (2018). Measuring allocative efficiency in cultural economics: The case of "Fundación Princesa de Asturias" (The Princess of Asturias Foundation). *Journal of Cultural Economics*, *42*(1), 91-110.

Financial Accounting Standards Board (FASB). (2008). *Statement of Financial Accounting Standards No.116: Accounting for contributions received and contributions made.* Original Pronouncements as Amended. Retrieved from https://www.fasb.org/jsp/FASB/Document_C/DocumentPage?cid=1218220128831&acceptedDisclaimer=true

Financial Accounting Standards Board (FASB). (2010). *Statement of Financial Accounting Standards No.157: Fair value measurements.* Original Pronouncements as Amended. Retrieved from https://www.fasb.org/jsp/FASB/Document_C/DocumentPage?cid=1218220130001&acceptedDisclaimer=true

Financial Accounting Standards Board (FASB). (2018). *Accounting Standards Update* (ASU 2018-13). Fair Value Measurement (Topic 820). Retrieved from https://asc.fasb.org/imageRoot/81/118196181.pdf

Finkler, S. A., Smith, D. L., Calabrese, T., & Purtell, R. (2016). Financial Management for public, health, and not-for-profit organizations (5th ed.). Prentice Hall.

Frumkin, P. (2012). The Idea of a Nonprofit and Voluntary Sector. In J. S. Ott and L. A. Dicke (Eds.), *The nature of the nonprofit sector* (pp. 17-30). Boulder, CO: Westview Press.

Gordon, T. P., Khumawala, S. B., Kraut, M., & Neely, D. G. (2010). Five dimensions of effectiveness for nonprofit annual reports. *Nonprofit Management and Leadership*, *21*(2), 209-228.

Grant, R. M. (1991). The resource-based theory of competitive advantage: implications for strategy formulation. *California management review*, *33*(3), 114-135.

Hall, R. (1993). A framework linking intangible resources and capabilities to sustainable competitive advantage. *Strategic management journal*, *14*(8), 607-618.

Hardy, C., & Maguire, S. (2008). Institutional entrepreneurship. *The Sage handbook of organizational institutionalism*, *1*, 198-217.

Hofer, C. W., & Schendel, D. (1978). *Strategy Formulation: Analytical concepts*. West Pub. Co. Retrieved from https://archive.org/

Hoskisson, R. E., Wan, W. P., Yiu, D., & Hitt, M. A. (1999). Theory and research in strategic management: Swings of a pendulum. *Journal of management*, *25*(3), 417-456.

Hwang, H., & Powell, W. W. (2009). The rationalization of charity: The influences of professionalism in the nonprofit sector. *Administrative science quarterly*, *54*(2), 268-298.

I.R.C. Reg. Sec. 1.501(c)(3)-1(a)(1)

I.R.C. Reg. Sec. 1.501(c)(3)-1(b)(1)(i)(b)

I.R.C.Reg. Sec. 1.501(c)(3)-1(c)(1)

Independent Sector. (2022, April 18). *Value of Volunteer Time*. https://independentsector.org/resource/value-of-volunteer-time/

Itami, H. (1991). Invisible assets. In *Mobilizing invisible assets* (T.W. Roehl, Trans.). Harvard University Press. (Original work published 1984)

Kioko, S., & Marlowe, J. (2016). Financial Strategy for Public Managers. Rebus.

Kong, E. (2008). The development of strategic management in the non-profit context: Intellectual capital in social service non-profit organizations. *International Journal of Management Reviews*, *10*(3), 281-299.

Krause, R., Wu, Z., Bruton, G. D., & Carter, S. M. (2019). The coercive isomorphism ripple effect: An investigation of nonprofit interlocks on corporate boards. *Academy of Management Journal*, *62*(1), 283-308.

Lecy, J. D., & Searing, E. A. (2015). Anatomy of the nonprofit starvation cycle: An analysis of falling overhead ratios in the nonprofit sector. *Nonprofit and Voluntary Sector Quarterly*, *44*(3), 539-563.

Lundström, T. (2001). Child protection, voluntary organizations, and the public sector in Sweden. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *12*(4), 355-371.

Maier, F., Meyer, M., & Steinbereithner, M. (2016). Nonprofit organizations becoming business-like: A systematic review. *Nonprofit and Voluntary Sector Quarterly*, *45*(1), 64-86.

Malatesta, D., & Smith, C. R. (2014). Lessons from resource dependence theory for contemporary public and nonprofit management. *Public Administration Review*, *74*(1), 14-25.

McKeever, B. S., & Pettijohn, S. L. (2015). The nonprofit sector in brief 2015. *Public Charities, Giving and*.

Molloy, J. C., Chadwick, C., Ployhart, R. E., & Golden, S. J. (2011). Making intangibles "tangible" in tests of resource-based theory: A multidisciplinary construct validation approach. *Journal of Management*, *37*(5), 1496-1518.

Mook, L., Sousa, J., Elgie, S., & Quarter, J. (2005). Accounting for the value of volunteer contributions. *Nonprofit Management and Leadership*, *15*(4), 401-415.

Nag, R., Hambrick, D. C., & Chen, M. J. (2007). What is strategic management, really? Inductive derivation of a consensus definition of the field. *Strategic management journal*, *28*(9), 935-955.

Pfeffer, J., & Salancik, G. R. (1978). Chapter Three. Social Control of Organizations. In *The External Control of Organizations: A Resource Dependence Perspective* (pp.39-61). New York: Harper & Row.

Prakash, A., & Gugerty, M.K. (2010). Introduction. In Prakash, A., & Gugerty, M. K. (Eds.). (2010). *Advocacy organizations and collective action* (pp. 1-28). Cambridge University Press.

Priem, R. L., & Butler, J. E. (2001). Is the resource-based "view" a useful perspective for strategic management research?. *Academy of management review*, *26*(1), 22-40.

Ruggiero, J. (1996). On the measurement of technical efficiency in the public sector. *European journal of operational research*, *90*(3), 553-565.

Salamon, L. M. (1999). The nonprofit sector at a crossroads: The case of America. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *10*(1), 5-23.

Samuelson, P. A. (1954). The pure theory of public expenditure. *The review of economics and statistics*, 387-389.

Schmid, H., Bar, M., & Nirel, R. (2008). Advocacy activities in nonprofit human service organizations: Implications for policy. *Nonprofit and Voluntary Sector Quarterly*, *37*(4), 581-602.

Sirmon, D. G., Hitt, M. A., Ireland, R. D., & Gilbert, B. A. (2011). Resource orchestration to create competitive advantage: Breadth, depth, and life cycle effects. *Journal of management*, *37*(5), 1390-1412.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Suárez, D. F., & Hwang, H. (2013). Resource constraints or cultural conformity? Nonprofit relationships with businesses. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *24*(3), 581-605.

Tuckman, H. P., & Chang, C. F. (1991). A methodology for measuring the financial vulnerability of charitable nonprofit organizations. *Nonprofit and voluntary sector quarterly*, *20*(4), 445-460.

U.S. Department of the Treasury, Internal Revenue Service. (2018). *Instructions for Form 990 Return of Organization Exempt From Income Tax* (Cat. No. 11283J). Retrieved from https://www.irs.gov/pub/irs-pdf/i990.pdf

U.S. Department of the Treasury, Internal Revenue Service. (2019). *Instructions for Form 990 Return of Organization Exempt From Income Tax* (Cat. No. 11283J). Retrieved from https://www.irs.gov/pub/irs-pdf/i990.pdf

Verbruggen, S., Christiaens, J., & Milis, K. (2011). Can Resource Dependence and Coercive Isomorphism Explain Nonprofit Organizations' Compliance With Reporting Standards? Nonprofit and Voluntary Sector Quarterly, 40(1), 5–32. https://doi.org/10.1177/0899764009355061

W. K. Kellogg Foundation. (2004). Logic model development guide: Using logic models to bring together planning, evaluation, and action. Battle Creek, MI: Author.

Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic management journal*, *5*(2), 171-180.

Williamson, O. E. (1981). The economics of organization: The transaction cost approach. *American journal of sociology*, *87*(3), 548-577.

Yao, K. (2015). Who gives? The determinants of charitable giving, volunteering, and their relationship. *Wharton Research Scholars, 126*.

Appendices

Appendix 2A: Current Challenges & Opportunities

I briefly present three examples, two highlighting the importance of valuation and one highlighting how valuation of underreported assets can enhance analytical insights.

Volunteer Valuation

The valuation of volunteers in the U.S. nonprofit sector has implications for research, practice, and the applicability of theory. There were approximately 1.44 million nonprofits in the U.S. registered with the IRS in 2012, involving an estimated 8.5 billion hours of volunteer time valued at $168.3 billion (McKeever and Pettijohn, 2015). If those volunteer hours were distributed equally across nonprofits, then each registered nonprofit in the United States would have received almost $117,000 in volunteer work, which is more than the $100,000 or less in total expenses identified by 29.6% of reporting nonprofits in 2012 (McKeever & Pettijohn, 2015, p. 4). In other words, almost one third of reporting nonprofits in 2012 would have received more in volunteer support alone than the total amount of their expenses for the year. Donations of volunteer labor may explain variation in donations of money, as volunteer labor and donated money are complementary (Callen, 1994). Insufficient reporting of volunteer labor can thus create an inaccurate portrayal of the resources available to the organization. Given that donors can restrict their contributions to certain programs of a recipient nonprofit (Kioko & Marlowe, 2016), complete data on nonprofit resource portfolios are essential to prevent over- and underestimation of the effects of different donations and revenue sources on organization-level outcomes. This ties into discussions of revenue diversification, and how vulnerable a nonprofit may be to changes in their revenue streams.

Revenue Analyses

A nonprofit's longevity and success in mission achievement are highly reliant on revenue generation. Revenue is primarily generated from a combination of donations, grants, contracts, fees for services, membership dues, sales, and investments; data on these distinct streams is found in IRS Form 990s filed by nonprofits (Carroll & Stater, 2009; Tuckman & Chang, 1991). Each revenue stream carries its own risk and variation in possible returns, leading some organizations to move towards diversifying revenue streams to have more predictable and stable returns over time (Carroll & Stater, 2009). In other words, "the larger the number of revenue sources a nonprofit has and the more equally divided its share of revenues from each source is, the less vulnerable it tends to be" (Tuckman & Chang, 1991, p. 453). Donations as a revenue source can be highly unstable due to changes in donor preferences, the economy, and the tax code (Tuckman & Chang, 1991). Given the evidence for the complementary nature of volunteering and donations (Callen, 1994), stability in underreported asset streams may affect other revenue (or received resource) streams. The predicted health and longevity of a nonprofit may be biased due to the lack of data on underreported assets, i.e., producing estimates that are too optimistic or pessimistic. Even more broadly, these discussions replicate the earlier critiques of imposing solely a financial valuation on volunteers' donated time that may fail to reflect the nonprofit's subjective valuations as well as distinguish between the value from different volunteers, e.g., equating the completion of general tasks to pro bono legal work.

Enhancing Analytical Insights

When present, information on underreported assets can yield novel understandings of nonprofits, even when looking at those resources in the aggregate. Discussions of organizational and programmatic efficiency shift in scale entirely, and the resources costs necessary to deliver

services can become significantly higher. The nonprofit Feeding America oversees a national network of hunger relief distribution programs such as food banks, food pantries, and food meal programs (Feeding America, n.d.). Its revenues for 2012 through 2019 range from $1.6B to almost $2.9B and are largely driven by donated goods and services (90.9% to 92.3% of total revenues).[18] Given the volume of donation revenues for Feeding America, excluding them changes the scale of discussion of its revenues from billions of dollars to hundreds of millions (Figure 2). In other words, the resource costs necessary to implement Feeding America's programs and deliver services could be understood to be less than 10% of the reality if the nonprofit did not have a valuation method for underreported assets. Studies of programmatic efficiency and organizational performance would be strongly biased and extremely unreliable.

It is important to note that Feeding America does not include valuations for all kinds of underreported assets: values for media support or "services that do not require special expertise" are not presented in the audited financial statements (identified in Note 2 of the financial statements for Feeding America, 2013b, 2014b, 2015b, 2016b, 2017b, 2018b, 2019b). If such underreported assets are nonnegligible in size, then the total resources necessary to deliver the same level of services each year could be even greater. As a result, replication of those services by other organizations might not be possible at the same scale or to the same degree.

[18] See Notes to Financial Statements, Note 2, g. Donated Goods and Services[18], in Feeding America, 2013b, 2014b, 2015b, 2016b, 2017b, 2018b, 2019b.

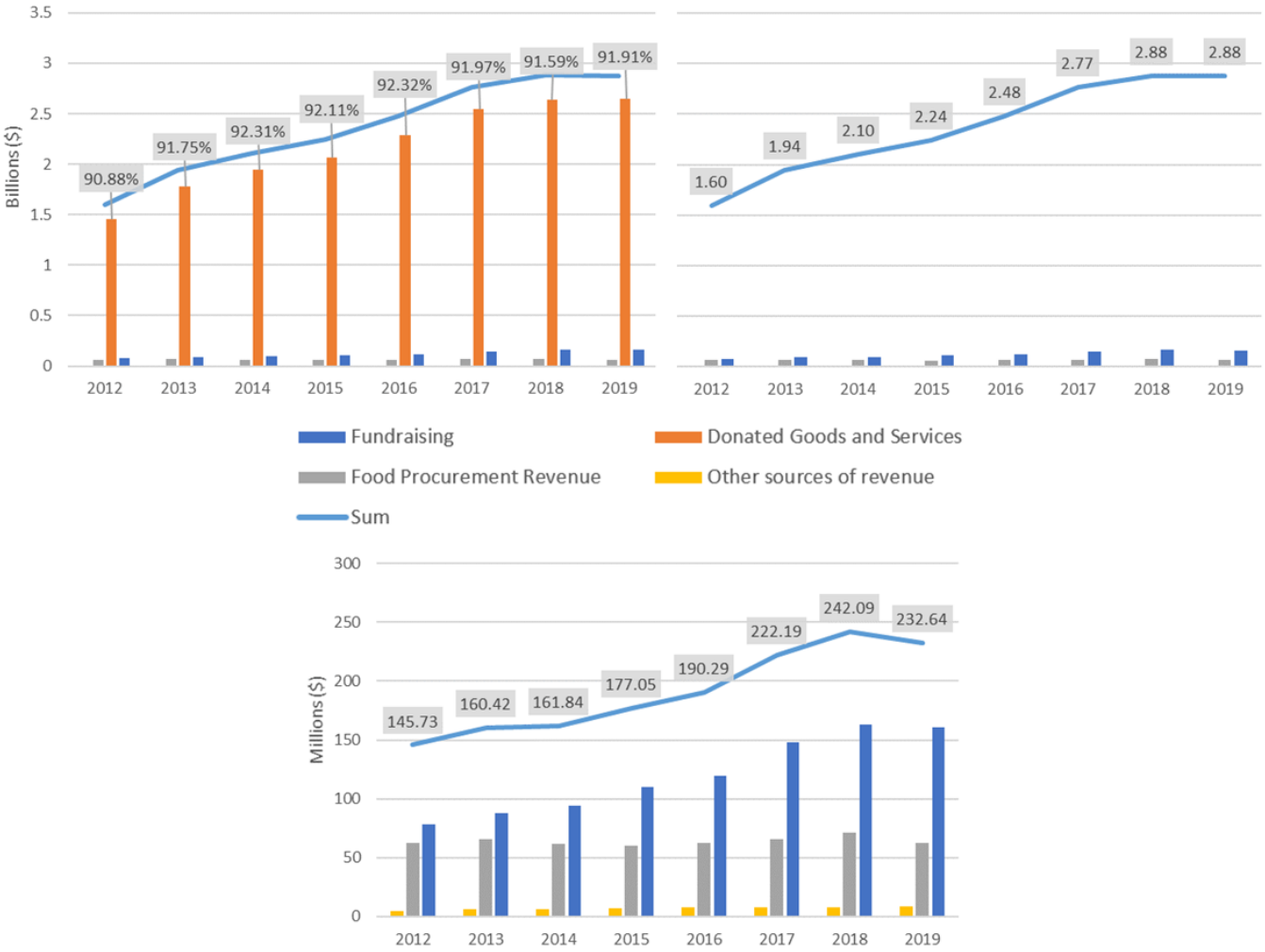

Figure 2. Feeding America's revenues, as reported in the financials sections of their annual reports for fiscal years 2012 through 2019 (Feeding America, 2012, 2013a, 2014a, 2015a, 2016a, 2017a, 2018a, 2019a). The plots in the top line have shared axes and trend lines, the difference being the inclusion versus removal of revenue from donated goods and services. The plot in the bottom line has had its vertical axis and trend line adjusted to not account for donated goods and services. The data labels indicate, in clockwise order and starting in the top left: what percentage of total revenues comes from donated goods and services; the total revenue reported for each year, in billions of dollars; the total revenue reported for each year, in millions of dollars. Despite the general increase in total revenue over time, donated goods and services continues to represent roughly the same amount (between 90.88% and 92.32%) of total annual revenue. Without this line of revenue, the organization has significantly smaller total revenues.

It is also worth noting that Feeding America may be an extreme case. It is a national organization that makes this information available in its annual reports and to the public. The existence of a valuation method already denotes it as an outlier and may be compelled by its dependence on and use of in-kind donations as part of its business model and potential in advertising for future donations. At the same time, these arguments present further justification for why valuation of underreported assets can be useful for nonprofits, e.g., it could increase future donations and strengthen the organization's model or approach. Organizations that do not have the resources or reputation of Feeding America but want to begin valuations of underreported assets face some challenges in developing their own methodologies.

Appendix 2B: Applying Resource-Based Theory

The field of strategic management centers resource use and capabilities in discussions of management and decision making (Hoskisson et al., 1999; Nag et al., 2007). Resource-based theory (RBT) provides a particularly rich analytical lens for the assessment of resources as units. It holds as a tenet that a firm's heterogenous bundle of resources "gives [it] its unique character" and sheds insight into a firm's position in its market or environment (Hoskisson et al., 1999, p. 438). As noted earlier, some scholars argue that "virtually *anything* associated with the firm can be a resource" (Priem and Butler, 2001, p. 32, emphasis in original); this introduces validity issues within and across studies that are further enhanced by a lack of consistent definitions (Molloy et al., 2011). Resource-based theory is inherently oriented to the for-profit sector and suffers from implementation challenges in nonprofit studies. Yet, its in-depth analyses provide a more robust foundation for analyzing resources than other theories commonly used in nonprofit studies.

Donated services and intangible assets are more difficult to classify than in-kind donations of goods or tangible assets. There are two noteworthy attempts to classify intangible assets and information-based resources (Itami, 1984/1991).[19] Hall (1993)'s framework consists of intangible asset classifications mapped onto organizational capabilities. Unfortunately, the framework's dimensions are mutually exclusive at the level of individual resources in a way that is incompatible with nonprofit sector accounting standards (FASB, 2008). Molloy et al. (2011) proposed a "multidisciplinary assessment process" or MAP to help scholars create appropriate constructs to measure and capture intangible assets for research drawing on RBT (Figure 3). Some steps using the MAP process are applicable to my research question, such as using "definitions [to] enable

[19] Itami (1984/1991) originally refers to such assets as "invisible," rather than "intangible," but the terms are used interchangeably in the literature.

construct clarity" (step 1; p. 1507) and using measurements that are valid, reliable, and practical (step 3; p. 1510). The fourth step of the MAP, the validity argument, can be explored and tested through the application of the taxonomy and comparing the outcomes of models.

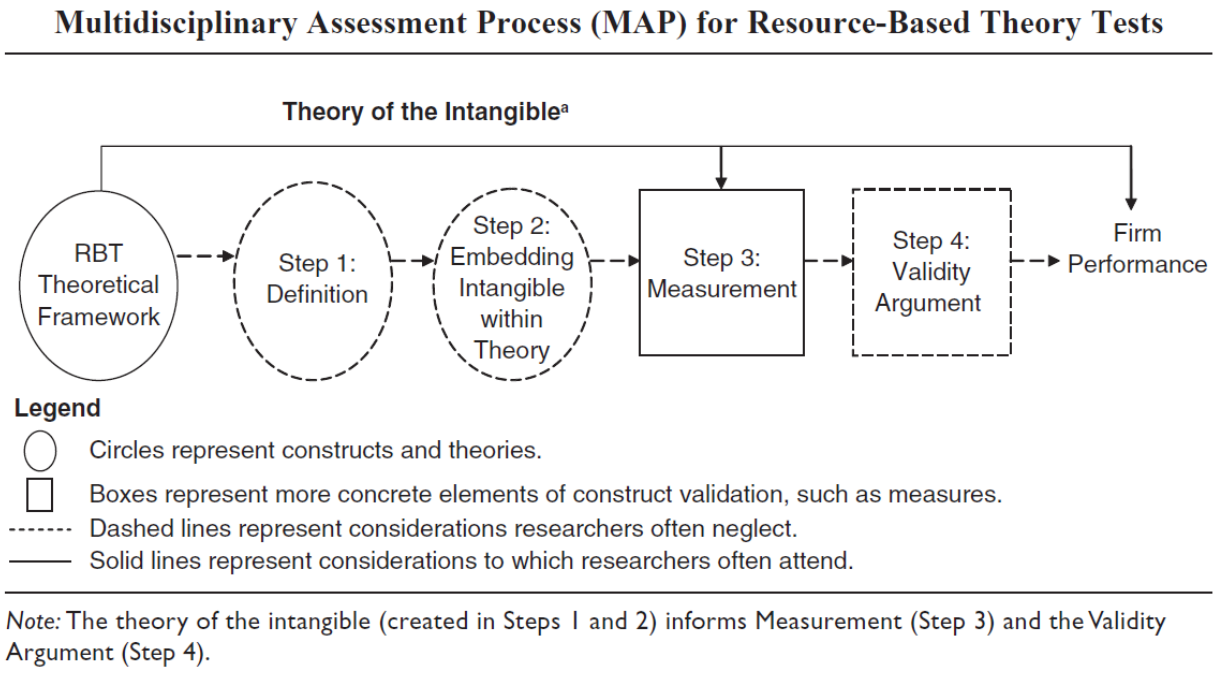

Figure 3. Molloy et al.'s (2011) process to create constructs to measure and capture intangible assets, p. 1508.

Steps 2 and 3 are firmly rooted steps in RBT and profit-maximization literature in ways that conflict with nonprofit, which have both "a legal and ethical obligation…to further their charitable mission" (Gordon et al., 2010, p. 209) Molloy et al. (2011) explicitly identify RBT in the name for step 2 instead of "Theory" used in Figure 3. The indicators proposed by the authors for the measurement step 3 are focused on RBT to the potential exclusion of other resource-aware theories that are not oriented towards profit maximization. The MAP presents a potential framework that can be adapted to provide a functional definition for resources that addresses nonprofits' demands

and requirements. Unfortunately, idiosyncratic or bespoke definitions could introduce unintended heterogeneity into comparisons when applied by various scholars not working in concert. In addition, it does not provide a consistent basis for valuation alongside definitions. The taxonomy is designed to allow both flexibility and customization by scholars and practitioners while retaining a consistent basis for valuation and for comparison across organizational contexts.

Appendix 2C: Additional Sample Propositions

Sample Proposition: Tangibility

> A. Higher percentages[20] of tangible resources will increase the strength of causal arguments of outcome-oriented impact evaluations, and they will increase confidence in the calculated "return on investment" from cost-oriented impact evaluations.

Tangible goods are explicitly measurable and observable. Because inventories of tangible resources can be maintained in real time using common, existing structures and tools, creating a causal argument linking tangible resources to evidence of observed outcomes will be more straightforward to make using existing, dominant econometric and qualitative approaches. For example, including an interview question on how a new community construction affects a relevant aspect of living conditions is a reasonable question to ask program beneficiaries. At the same time, because tangible resources are clearly measurable and observable, financial calculations can be performed that explicitly tie them to each outcome. As the percentage of tangible resources in the resource portfolio increases, the confidence in the accuracy and precision of cost calculations will increase.

> B. Higher percentages of intangible resources will decrease the strength of causal arguments of outcome-oriented impact evaluations, and they will decrease confidence in the calculated "return on investment" from cost-oriented impact evaluations.

Intangible resources are explicitly difficult to measure and observe. The effects of organizational culture and reputation on delivery of services can be hard to measure and assign valuations, even

[20] I use "percentage" throughout this section as shorthand for percentage of the total resource portfolio represented by this kind of resource.

if recorded in great detail. Calculations can sometimes be academic or creative in ways that may make them inaccessible to many organizations. Intangible resources may also relate to multiple inputs, activities, and programs, further diluting the strength of causal arguments and their attribution to a particular set of outcomes. For these reasons, as the percentage of intangible resources in the resource portfolio increases, the strength and clarity of causal relationships between the resource portfolio and the observed outcomes as well as confidence in the accuracy and precision of cost calculations will decrease.

Sample Proposition: Justification

> A. Instrumental-justified-only resources will increase the strength of causal arguments of outcome-oriented impact evaluations, and they will increase confidence in the calculated "return on investment" from cost-oriented impact evaluations.

Contributions given solely for instrumental purposes are intended to be immediately applicable and implementable. This may not always be the case due to various reasons. But, because this is more likely relative to expressive-only contributions, it will be easier to associate use of such resources in a program to the observed outcomes. Because of this ease of observation and measurement, there will be less uncertainty in the calculations on the cost effectiveness of programs.

> B. Expressive-justified-only resources will decrease the strength of causal arguments of outcome-oriented impact evaluations, and they will decrease confidence in the calculated "return on investment" from cost-oriented impact evaluations.

Contributions given solely for expressive reasons, while showing solidarity with the purpose of the program or organization, may not be immediately applicable or appropriate to implement. For

this reason, as the percentage of such resources increases, building a causal association between inputs and outcomes becomes more complex, and the expected accuracy and precision will decrease. Cost effectiveness calculations could yield higher returns on investment because more resources are less explicitly allocated to associated with a program. They could also yield lower returns because resources may be over-allocated to a particular program, and so the program seems to require more resources as inputs. Either way, confidence in the calculations will decrease.

> C. Expressive- and instrumental-justified resources will increase the strength of causal arguments of outcome-oriented impact evaluations, and they will increase confidence in the calculated "return on investment" from cost-oriented impact evaluations.

As the percentage of resources that are both expressive and instrumental increase, the increased strength of causal attributions from resources as inputs to observed outcomes will increase because the contributions' orientations will be more functional than those given solely for expressive reasons. In other words, the causal strength gained from the instrumental justification will be greater than the causal strength lost from the expressive justification. This will also contribute to greater confidence in the calculated returns from program inputs. I expect these proposed relationships to hold when the percentage of resources shifts from expressive-justified-only to both expressive and instrumental. If the percentage of resources shifts from instrumental-justified only to both expressive and instrumental, then I expect that the strength of causal arguments and confidence in cost calculations will stay the same or diminish somewhat due to the greater presence of expressive justifications in the resource portfolio.

Sample Proposition: Structural Level

> A. Higher percentages of program resources will increase the strength of causal arguments of outcome-oriented impact evaluations, and they will reduce the expected "return on investment" from cost-oriented impact evaluations.

As more resources are clearly allocated to programs, causal arguments would improve because the entire value or contribution of the resource could be allocated to the proposed relationship between inputs and outcomes. The reduction in complexity could enhance perceived legitimacy and strength of the proposed argument. At the same time, because there are more full values of resources attributed to a program, the measured cost of the program would increase and thus decrease the return on investment.

> B. Higher percentages of organization resources will decrease the strength of causal arguments of outcome-oriented impact evaluations, and they will increase the expected "return on investment" from cost-oriented impact evaluations.

Unless the organization has only one program, organization resources will need to be allocated in some way to a particular program. This results in arguments that connect inputs to outcomes navigating calculations of increasing complexity. In addition, organizations could use multiple different allocation bases for assigning organization resources to individual programs, and justify applying different bases to different resources. If two nonprofits with the same programs and the same costs for each received the same organization resources, then it is possible that the two nonprofits would allocate those organization resources differently such that they identify different required inputs or resource costs for each of their programs. As the percentage of organization resources increases, this distributive effect would create a perception that programs are cheaper to implement by sharing costs across multiple programs. Programs could then be perceived to be

more cost effective, but at the cost of complicating or even reducing the strength of causal arguments between inputs and outcomes.

Sample Proposition: Temporal[21]

> A. As resources from time-bound contributions increase, the strength of causal arguments of outcome-oriented impact evaluations will increase and the expected "return on investment" from cost-oriented impact evaluations will increase.
>
> B. As resources from ongoing contributions increase, the strength of causal arguments of outcome-oriented impact evaluations will decrease and the expected "return on investment" from cost-oriented impact evaluations will decrease.

Time-bound contributions are discrete instances, meaning that they are clearly measurable and observable. Ongoing contributions are non-discrete such that they may be more difficult to measure, observe, and attribute. I therefore propose that resources from these contributions have inverse relationships to both the strength of causal arguments for efficacy and calculated values of cost efficiency. Because of their characteristics, resources from ongoing contributions will increase in value over time. Impact evaluations occur after the start of a program's implementation, and typically closer to or after the program ends. The value of resources from ongoing contributions will have therefore increased relative to when the program started. As the percentage of such resources increase, the calculated effect from a single dollar of input will decrease. On the other hand, the value of resources from time-bound contributions are static throughout the life of the

[21] As noted in the section "Taxonomy Overview," repeating time bound may operate as a special case of either time bound or ongoing. This section only discusses time bound and ongoing.

program. If the percentage of value-static resources in the portfolio increase and value-increasing resources decrease, then I expect that the calculated program cost efficiency will increase.

Sample Proposition: Organizational Need

> Resources' organizational need will not affect the strength of causal arguments of outcome-oriented impact evaluations; resources with greater need will increase the expected "return on investment" from cost-oriented impact evaluations.

Resources that an organization identifies as needed are ones that it has identified as necessary to delivering services and performing activities towards achieving observed outcomes. For this reason, increasing the percentage of resources that an organization identifies are of higher need will not affect the causal arguments because they are implicitly already present.

As the organization identifies a greater need for certain resources to achieve outcomes, then an increasing percentage of necessary resources could have no effect or a positive effect on expected return on investment. Depending on the effect of the needed resource on the outcome, varying degrees of substitution effects could affect cost effectiveness calculations. I anticipate that increasing the percentage of resources that are of high need will either maintain the current cost effectiveness or increase it by allowing organizations to use resources as intended, more effectively, or improve existing systems and processes with the high-need resources. Even if the number of situations with no change is greater than the number of situations with increased cost effectiveness, taking the average of those cases would suggest that increasing the percentage of resources that are of high need will have some degree of increase on expected returns on investment.

Sample Proposition: Example composite

The following proposition is an impact evaluation-focused example of how multiple dimensions can be combined.

> Portfolios with resources that are more tangible, instrumental only or instrumental and expressive, at the program level, time-bound, and with less related expenses are more likely to enhance proposed causal relationships between inputs and outcomes than will portfolios with resources that are more intangible, expressive only, at the organizational level, ongoing, and with more related expenses.

I synthesized a composite proposition by reviewing the proposed relationships between the values of each dimension and the causal arguments at the heart of outcome-oriented impact evaluations. This proposition allows me to compare portfolios in the aggregate, both in terms of total resources and across all relevant dimensions. If this proposition holds, then organizations looking to perform stronger, more rigorous impact evaluations have guidance on the kinds of resource portfolios that they should strive to have when implementing programs. This finding would have multiple consequences, ranging from shaping how organizations craft their fundraising and development strategies, direct resources at the start of and throughout programs, and design theories of change and similar causal relationships. Identifying drivers of programmatic efficacy would also benefit organizations operating in similar spaces, funders and governments looking to partner with nonprofits, and, most importantly, the intended beneficiaries of program services.

Paper 3. Creating A Multi-Dimensional Rationalization Measure

Introduction

Nonprofit outcomes related to health and performance are often captured only through financial measures such as profitability and overhead expense ratios, with significant concerns of generalizability and validity across studies, as well as utility in truly capturing important dimensions of performance (Coupet & Berrett, 2019; Prentice, 2016). Without improved measures, understanding the relationship between a nonprofit's performance and how it uses its resources via processes will continue to be limited. This issue is particularly salient for nonprofits because, compared to private enterprises and government entities, the average nonprofit organization operates under significant constraints idiosyncratic to the sector. For example, restrictions on donations can impose labyrinthine limits on how nonprofits allocate funds towards program and administrative expenses, and nonprofits are often highly reliant on human resources to affect what they can and cannot accomplish (Akingbola, 2013; Kioko & Marlowe, 2016).

The average nonprofit also has a different resource mix than a private enterprise or government, further complicating how nonprofit performance can be reliably understood. This is demonstrated through resource or asset[1] categories such as pro bono work and volunteer contributions[2] of professional services. These services are often heavily underreported or excluded from data sources and, as a result, their effects on nonprofit performance are poorly understood (Cordery &

[1] Assets and resources will be used interchangeably in this manuscript. Distinctions made by authors will be identified as needed. Overall, assets will be considered a type of resource controlled by the firm, in line with Barney's (1991) framing.

[2] As in Paper 2, for consistency, I default to using "contributions" from here on except when a cited or quoted source uses "donations." For additional information, please refer to the relevant footnote in that paper.

Narraway, 2010; Cordery et al., 2013; Gordon et al., 2010). Contributed services are commonly understood as a subset of volunteering, and overall volunteering is an important asset of nonprofits – in 2012 alone there were 8.5 billion hours of volunteer time, valued at $168.3 billion (McKeever & Pettijohn, 2015). At the same time, there are multiple issues with both reporting and valuation of all forms of volunteer labor, such as the lack of reporting on volunteer time data (ex. Cordery & Narraway, 2010) and inconsistent, too complex, and questionable valuations (ex. Callen, 1994; Cordery et al., 2013). In addition, differentiation among volunteers' contributions is little explored. For example, distinctions by level of contributed time and experience has yielded the concept of super-volunteers, for those who contribute above-average time and dedication (Einolf and Yung, 2018). Super-volunteers become a critical asset for nonprofits because they reduce firm expenditures on training and onboarding staff. What is missing from these discussions is a focus on understanding how contributions of professional knowledge and services can affect organizational outcomes such as performance. Nonprofit outcomes related to health and performance are often measured financially but with significant concerns of generalizability and validity across studies, as well as utility in truly capturing performance (Coupet & Berrett, 2019; Prentice, 2016).

The inconsistency in measurement of nonprofit performance and contributed services suggests low rationalization or a lack of consistent, formalized processes among nonprofits, as well as agents for accountability (ex. FASB[3], IRS, funders, researchers). I define rationalization as *the formalization of core (service delivery) and support (management) processes within a nonprofit* (Dart, 2004; Hwang & Powell, 2009; Suárez & Hwang, 2013; Maier et al., 2016). Formalization

[3] Financial Accounting Standards Board (FASB) establishes guidance on how to generate and present information in financial statements.

involves the specification or articulation of exactly *how* a process or a given set of steps should be performed to achieve a desired output, where there are limited or predefined options for what steps consist of and how they are ordered. Inputs include the range of assets controlled by or accessible to an organization – financial, knowledge, human, etc. What differentiates donation of professional services from other forms of volunteer labor is that skilled donors of services[4] have their own predefined notions of what are "acceptable" sets of steps, or permutations, to implement towards achieving a particular output. Their notions of what are and are not appropriate permutations, i.e., different versions of processes that yield the same output, are driven by their professional training and experience. Such contributions may consist of sets of steps that may overlap with or differ different from those already present within the organization, similar to the notions of novel versus possessed knowledge studied in strategic management literature (ex. Sears & Hoetker, 2014). This is just one example of how the fit between nonprofits' organizational characteristics and the assumptions behind donated services must be understood when exploring the effects of contributions on nonprofit performance (Van de Ven & Drazin, 1984).

While such contributions may be useful for some nonprofits, I argue that the degree of beneficial effects on nonprofits depends on both the previous extent of process formalization within the organization and on the fit of a recipient nonprofit's characteristics to its environment. Donors developed their professional services and skills in particular organizational contexts, consisting of internal and external environmental characteristics. Donors' contributions of services thus assume and rationalize processes that are unlikely to be identical between the nonprofit and the original context. Equivalent processes in nonprofits may vary slightly, significantly, or may not exist, and

[4] For additional explanation on why I distinguish donors of skills and knowledge from volunteers, please refer to Appendix A: Identifying Who is a Donor.

the nonprofit's context may be slightly to extremely different to the original context. For this reason, the ways in which contributions rationalize nonprofits and associated necessary and sufficient conditions should be explored further.

For example, a market research specialist may assume that identifying new customers is done best via a particular software and wants to teach their identification process to a nonprofit that wants to learn how to better identify potential financial donors but does not have the software. The specialist could either teach the process without the software, which may reduce the magnitude of the potential gains to nonprofit performance, or convince the organization to purchase the software, which would introduce additional resource burdens such as financial, employee training time, etc. In either case, the organization experiences some degree of rationalization by formalizing how they identify new customers (the outcome), which in turn can affect multiple aspects of the organization. The change in rationalization can yield varying degrees of cost effectiveness and performance improvement.

Understanding the ways in which professional services contribute to rationalization has been hampered to date by the numerous challenges facing the dominant measure of nonprofit rationalization, such as generalizability to non-U.S. settings, applicability to large datasets, and ways to incorporate quantitative and mixed data. Song & Yin (2019)'s efforts to contextualize the four-dimensional measure first proposed by Hwang & Powell (2009) to a Chinese setting required them to replace one of the four dimensions because it was not relevant, suggesting that future non-U.S.-oriented studies may need to make similar edits. Furthermore, data used to calculate these rationalization measures are not consistently available in ways that take advantage of the large amounts of digital public nonprofit data. Researchers have historically calculated the measure

using solely qualitative data, i.e., interviews and surveys,[5] which they hand-coded for the presence of each of the four dimensions of the rationalization measure. These challenges suggest that the existing measure may be limited to small-scale studies that do not incorporate quantitative or mixed data and require human coders.

Creating a rationalization measure that addresses these challenges can assist in modeling how contributed services affect both a nonprofit's degree of rationalization and its ability to convert resources into desired outputs, e.g., efficiency and efficacy. In addressing these considerations, my proposed generalizable measure would enhance the dominant measure by identifying the degree of rationalization experienced by a nonprofit in a more fine-grained or nuanced way. It also contributes to nonprofit performance research by drawing on common data, i.e., the characteristics of individual organizations, and addressing limits identified in the literature. Publicly available sets of nonprofit data often contain a mixture of this common data in qualitative and quantitative formats and across geographic contexts, meaning that the generalizable measure would immediately enhance understanding across all areas of nonprofit performance and to other areas of nonprofit research.

This paper presents a multidimensional rationalization measure that can take advantage of large datasets and accounts for limitations in existing measures. It outlines the context for why the proposed measure is needed by exploring the case of in-kind contributions' effects on nonprofit performance. I focus on in-kind contributions rather than other rationalizing forces, such as restrictions on donations, because such contributions are more likely to introduce and present possible misalignment across the set of existing processes at an organization. I anticipate that

[5] Interviews were used in Hwang & Powell (2009) and Suárez & Hwang (2013). Surveys were used in Song & Yin (2019).

restricted funds would represent a challenge (solely) for internal resource allocation. While important, their rationalizing effects and consequences are less likely to be observable across the organization as would effects from in-kind contributions of goods and services, in particular donated professional services (ex. Lindenberg, 2001).

After outlining the context, the paper presents a demonstration of how the measure can be created by building from existing literature and leveraging a combination of machine learning and big data. The paper continues with a demonstration of the concept's utility, followed by reflections on potential limits and proposed objectives for future studies.

Context: Outcome Measures Limitations

In-kind contributions of services can target the entire nonprofit or only specific programs, e.g., creating a website for a theater nonprofit vs. teaching Broadway-level choreography for a specific theatrical production. This question of organizational scope complicates efforts to implement a single construct that measures changes in organization-level performance and to attribute those changes to the contribution. In addition, studies can implement many different, and sometimes incompatible or incomparable, constructs to measure just one aspect of performance. For example, one study found 70 unique variables used to measure nonprofit financial health (Prentice, 2016). Not all nonprofits produce tangible or intangible outputs, and trying to account for that subjective variation in financial terms may have limited utility. Furthermore, financial measures have limited success and utility as proxies to capture performance (Coupet & Berrett, 2019). At the same time, efforts to use non-financial performance measures face limited utility if accounting for too few dimensions of performance: in-kind contributions of services can affect multiple parts of the organization, from overhead to individual programs, and most low-dimensional measures are

insufficiently complex to capture this variety. The following section presents an overview of the challenges associated with trying to implement financial and non-financial performance measures when exploring the relationship between in-kind contributions of services and performance.

Financial

Because the effects of in-kind contributions on nonprofit performance measures are difficult to measure, financial performance measures are often used as alternative or proxy measures. Nonprofit financial data is somewhat consistent and formalized across organizations due to requirements for how financial statements and tax returns must be prepared. Nonprofit tax returns are publicly available; as barriers to accessing this data have decreased, financial data from tax returns has been increasingly looked to by researchers and practitioners as a way to understand nonprofit performance. Profitability, for example, is defined as revenues minus expenses, when the difference is positive (Kioko & Marlowe, 2016). This calculation is standard for nonprofits and it is a mandatory field in the annual IRS Form 990 nonprofit tax return (see p.1, Part I, line 19 in USDT, 2019a). Yet, despite this standardization, one review of nonprofit financial literature identifies nine different formulas that various articles have referred to as generating a calculation of "profitability" [6] (Prentice, 2016).

This issue ties into a growing critique of common financial measures used to determine nonprofit performance, e.g., overhead expense ratios (see Coupet & Berrett, 2019). In a comprehensive review of financial performance measures, Prentice (2016) notes that their implementation in nonprofit research introduces two sets of issues, that of construct validity due to reduction of multiple dimensions into one value and that of external validity, i.e., variations in implementation

[6] See Prentice (2016), Appendix, column "Construct referenced in the literature." I counted the number of unique instances in which the author listed "profitability" in a row.

of the same measure across papers resulting in distinct constructs and limited comparability. Despite the seeming objectivity of financial measures based on calculations, the lack of consistent implementation and alignment of calculations and concepts identifies a fundamental weakness to any large studies or cross-study analyses.

Non-financial

Much has been written about the insufficiencies of using purely financial measures to track nonprofit performance. As noted by Gordon et al. (2010), "even with a multidimensional approach to measuring a nonprofit organization's financial performance, financial accounting is unable to show whether the nonprofit organization's mission is being accomplished" (p. 221). Going beyond such measures has been a critical task for various scholars, even outside of the field of nonprofit studies (e.g., Tsarenko & Simpson, 2017). Maier et al. (2016) note a distinction between an organization's performance and its fulfillment of societal functions, further defining performance as "understood within the NPO's own frame of reference, that is, the fulfillment of its mission and the securing of financial and human resources" (p. 75). In Appendix B, I explore a possible solution to measure nonprofit annual mission achievement using a simple percent change calculation. I then consider how this simple and direct measure insufficiently captures the effects of rationalizing in-kind contributions on nonprofit performance. Numerous technical limitations also affect implementations of subjective, mission-oriented performance measures, for example, multiple key outcomes, changes in missions, and limited data on key outcomes. While the limits may vary for practitioners and researchers internal to vs. external to nonprofits, the lack of consistency in implementation greatly reduces the utility of non-financial measures.

Importance of measuring rationalization

Rationalization relates to the core service delivery processes that compose organizational capacity and how organizations work to fulfill their missions through managing activities and resources (Jones et al., 2017; Paynter & Berner, 2014). The extent of rationalization within a nonprofit has implications for its efficacy, or ability to convert resources and inputs into desired outputs, and efficiency, doing so in ways that yield the least wasted inputs and achieves the greatest returns for its intended beneficiaries. Alternatively, too much and too little rationalization could reduce efficacy and efficiency, as discussed in the illustrative examples below, with negative implications for the nonprofit's ability to fulfill its mission. A multidimensional measure of rationalization that can be causally related to various kinds of assets and inputs can contribute to research on how changes to rationalization can affect various measure of efficacy and efficiency, such as the financial measures discussed earlier.

In-Kind Contributions and Organizational Fit

I expect that in-kind contributions of services will be most likely to achieve the intended effects on outcomes when the operating assumptions of the contribution aligns well with the operating reality of the recipient nonprofit. This reality is composed of various internal and external organizational factors that, when taken in aggregate, are unique to each organization. In an engagement between a donor and a recipient nonprofit, the contribution requires various operating assumptions to be met for it to achieve the outcome intended by the donor and desired by the nonprofit. These assumptions could be about whether resources are highly constrained, mission achievement can occur at the expense of profitability, amount of available staff time for training, importance of stakeholder engagement, and numerous other considerations when implementing changes within an organization. To understand how these operating assumptions and contribution

characteristics may facilitate or impede achievement of the intended/desired outcome, I draw on the notion of "fit" in strategic management literature.

Fit refers to the effect of two or more organizational factors on organizational performance; such internal and external organizational factors include environment, strategy, structure, systems, style, culture, etc. (Van de Ven & Drazin, 1984). It has been applied and adapted to numerous questions and contexts, multiple of which are relevant to exploring how the rationalization resulting from contributions can affect organizational outcomes. For nonprofits, one of those key outcomes is making progress towards its mission. As noted earlier, rationalization relates to the core service delivery processes that compose organizational capacity and, ultimately, how organizations work to fulfil their missions through managing activities and resources (Jones et al., 2017; Paynter & Berner, 2014). In this way, rationalizing contributions should ideally relate to the mission of the nonprofit as well as its efficiency and efficacy. Similarly, a category of contributed service that causes the nonprofit to fit less with the demands of its operating environment would reduce its performance, whereas one that increases fit would improve performance (Volberda et al., 2012). A contribution that rationalizes a nonprofit's processes in a way that reduces its ability to respond to its operating environment, while formalizing processes, may have a negative effect on the nonprofit's efficacy and mission achievement.

Another useful perspective on fit to understand how engagements and rationalization interact to affect nonprofit performance is that of novelty of knowledge. While the introduction of novel knowledge such as external practices may lead to a reduction in the possible permutations of steps that the organization can perform to achieve a desired outcome, the extent to which this "benefit"

[7]is achieved depends on various elements of that knowledge. For example, Sears and Hoetker (2014) explore the duplication of knowledge between acquired and acquiring organizations. They find that "novelty outweighs familiarity" (Sears & Hoetker, 2014, p. 51) – as overlap between novel and possessed knowledge increases, there is a decrease in the ability of each of the two knowledges to yield greater-than-expected performance outcomes. In other words, as the percent of possessed knowledge increases and novel knowledge decreases, the potential benefits gained from the novel knowledge are subsumed. Figure visualizes potential overlaps. Assuming that each circle represents a potential benefit of 1, the upper case of no overlap indicates that the novel and possessed knowledge will yield a total benefit of 2. The middle case of partial overlap indicates a total benefit of 1.5, where 1 is from the possessed knowledge and 0.5 is additional benefits gained from the novel knowledge. The lower case of significant overlap indicates a total benefit of 1.2, where 1 is from the possessed knowledge and 0.2 is additional benefits from the novel knowledge. I anticipate that the relationship between contribution-derived novel knowledge and a nonprofit's possessed knowledge is positively related to the nonprofit's performance gains: as a contribution introduces less novel knowledge into a nonprofit, I anticipate that nonprofit performance gains would diminish.

[7] As noted in multiple places in this paper, rationalization may not always yield beneficial outcomes for the nonprofit and its stakeholders. While rationalization may be the desired outcome, it may cause unintended changes to the organization that threaten its capacity to achieve its mission, as discussed below.

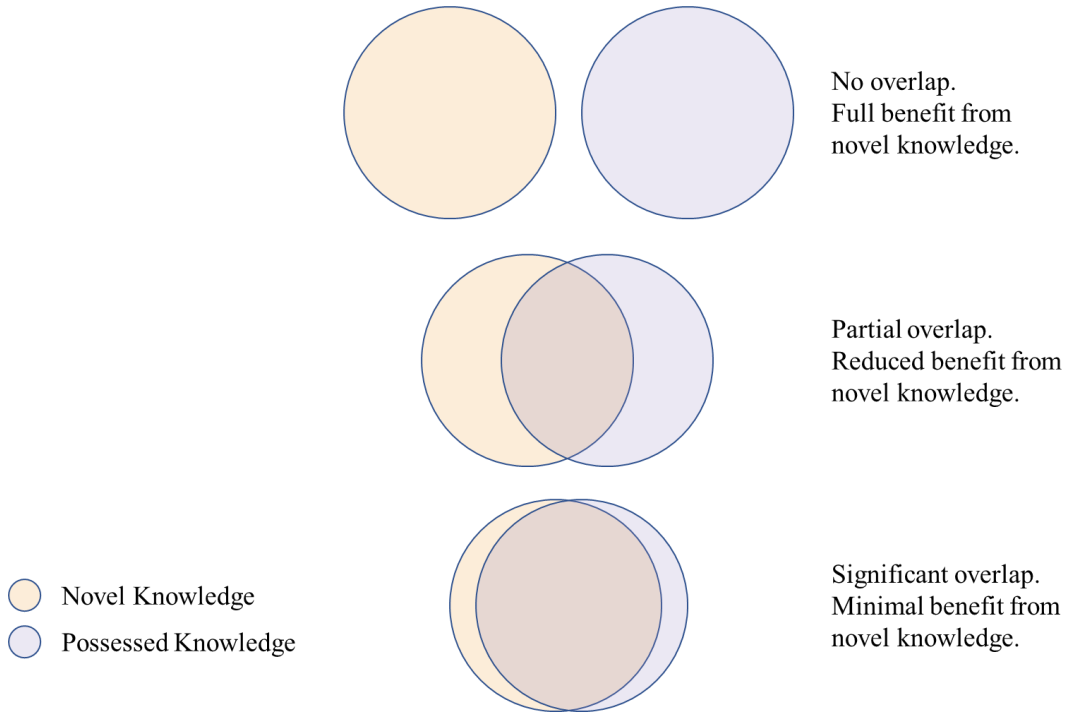

Figure 1. Demonstration of how increasing overlap between novel and possessed knowledge reduces benefits for the recipient organization. For an alternative representation that also distinguishes between the acquiring and target organizations, please refer to Figure 1 in Sears & Hoetker (2014).

This suggests that knowledge composition matters: a contribution that allows an organization to stand up a new program would be of greater benefit than a contribution that provides minor enhancements, and building a new website that highlights the nonprofit's strengths is more beneficial than fixing errors on a website that is poorly designed and does not show those strengths.

Even when a contribution introduces non-overlapping novel knowledge to a nonprofit, that knowledge can be expensive to integrate (Han et al., 2018). In addition, the extent to which it contributes to and complements or fails to complement existing capacities (ex. human resource, finance, infrastructural, strategic) matters (Jones et al., 2017). Returning to the previous example of the market research specialist donor, their contribution of training on identifying new potential

donors requires that the organization buys a particular program, employee time, and other required resources that make it costly to integrate this new process into the nonprofit. Furthermore, if the nonprofit already does this well but really needed training on how to manage and train donor relations in the long term, the disconnect between contribution and actual need reduces potential benefits from the contribution.

This suggests that in-kind contributions, i.e., ones that provide novel knowledge and rationalization of processes, that have some degree of overlap to existing organizational knowledge and contribute to current capacities would be preferable over purely novel knowledge (no overlap) or too much overlap (the nonprofit already possesses the knowledge). In other words, a donation that introduces an entirely novel process or exactly replicates an existing one will not yield as many net benefits as a donation that introduces a modification, such as a novel permutation or reduction of permutations of process steps, that is not costly to implement and yields benefits. While noting that rationalization can both benefit and harm a nonprofit, I propose that contributions that rationalize existing processes rather than add additional organizational capacity are more likely to have direct positive benefits because of lower additional costs and burdens to integrate and maintain that additional capacity, relative to increased costs to the modified, existing processes.

Illustrating Rationalization

I present here two theoretical, illustrative examples of organizations with opposite degrees of rationalization and how I expect them to use resources as inputs, with greater detail for each available in Appendix 3C: Illustrating Minimal and Maximal Rationalization. [8]

Minimal Rationalization

A *minimally rationalized organization* would not have formalized processes or procedures for performing portions of the production process and delivering internal or external outputs, such as measuring and reporting on outcomes, organizing events, or tracking inventory. This organization would create processes ad hoc; while there might be some duplication or replication of processes, it would be on an informal basis and constrained to individuals or groups of individuals that work together. Inefficient use of resources to achieve outputs is highly probable due to reasons such as duplication of steps, performing unnecessary steps or in the wrong order, or including steps that are resource intensive when there are less expensive alternative steps. What this suggests is that a minimally rationalized organization would struggle to pinpoint how it uses its resource to achieve outputs, with consequences for outcome measurement and understanding its efficacy and efficiency.

Another anticipated characteristic is a lack of shared understanding between employees and leadership as to what the process output should look like. When processes are not resource expensive or the organization is not resource constrained, the consequences of mismatched expectations on process outputs may be negligible, with consequences increasing in severity as processes become more resource intensive and the organization more resource constrained. While

[8] The two examples are meant to capture what I expect the consequences of rationalization to be for each extreme degree of rationalization. While these are expected or possible outcomes, they are not necessarily desirable for organizations, researchers, donors, beneficiaries, and other stakeholders that may be affected.

minimally rationalized organizations suffer from a lack of consistent structure, they would see some benefits in responding to internal and external shocks due to their capacity to create and implement as-needed processes.

Maximal Rationalization

A *maximally rationalized organization* would have formalized processes and procedures for delivering all possible internal and external outputs produced by the organization. This implies a high degree of administrative structure including systems for oversight and compliance to ensure that only the "officially" recognized process permutations are being performed by all members of the organization. The act of reducing possible steps and permutations will constrain innovation and flexibility for the purpose of yielding more resource efficiency through reducing duplication, minimizing the total pool of possible resources through only performing steps in orders that the organization recognizes as acceptable, and creating more opportunities for inputs to be standardized such that there are less unique resources and the total resource pool can be used more flexibly. Thanks to the potential for improved input and resource accounting, a maximally rationalized organization should be able to more easily attribute each consumed resource to the outputs produced from its processes for reporting purposes, in response to audits, and to demonstrate greater accountability and transparency. Such organizations should be well positioned to measure their outcomes and understand their efficacy and efficiency, but measuring progress towards these concepts is not the same as achieving them.

Too much rationalization may codify steps and permutations that are not the most efficient or effective. When processes are not resource expensive or the organization is not resource constrained, the consequences of inefficient or outdated steps and permutations on outputs may be negligible, with consequences again increasing as resources become more expensive and the

organization more constrained. The reduction in flexibility and innovation could also insulate the organization such that it is slow or unable to respond to external changes, such as shifting consumer demands and beneficiary needs, new legal requirements, or changing donor expectations.

Weaknesses of Current Measures

These illustrative examples are meant to showcase some limits in our knowledge of

- the nonprofit production process, i.e., transformation of inputs into outputs and outcomes, and
- how rationalization works within organizations to affect organization-wide characteristics and behaviors.

They also serve to provide a shared scale or common understanding of thinking about what rationalization might look like. Yet, the measurement of rationalization itself faces challenges, as existing measures of rationalization are limited to certain data inputs which do not always reflect the many considerations presented in the illustrative examples. For example, the approach outlined in both Hwang & Powell (2009, p. 278-9) and Suárez & Hwang (2013, p. 594) focuses on the use of binary variables (1 = present, 0 = not present) to capture the presence of four dimensions of rationalization within nonprofits: strategic plans, independent financial audits, data to use in quantitative program evaluations, and use of consultants. The studies leveraged the binary nature of the data to generate a matrix of tetrachoric correlation coefficients. Using principal component analysis (PCA), the research teams were able to generate a factor with an eigenvalue greater than 1. Hwang & Powell (2009) extended this further by using the factor loadings generated through PCA to generated a standardized factor score that captures the degree of rationalization.

The goal of this paper is to leverage the same core findings and approach of these two studies while creating a measure that can be generalized and take advantage of more inputs. There is no way to define an *ideally rationalized organization*, but understanding when levels of rationalization are supporting improved performance requires more than just understanding strategic plans, financial audits, quantitative data, and use of consultants. Even if we assume that the combination of these four inputs provides a precise, accurate understanding of rationalization, data on the four are not consistently available across nonprofits. For this reason, creating a more robust, flexible measure of rationalization will shed insights into nonprofit production, performance, and the effects of in- kind contributions.

Creating a flexible rationalization score

The goal of this paper is to demonstrate how a flexible rationalization score can be created by leveraging existing research and large datasets and to suggest how such a score can improve our understanding of the nonprofit production process. As a proof of concept, I use a machine learning algorithm to identify, of the variables available in the electronically filed IRS 990 tax returns, which variables are the most indicative of rationalization. I use the measure proposed by Hwang & Powell (2009) and Suárez & Hwang (2013) as the dependent variable used to train the machine learning algorithm. I do not used the modified version of Song (2017, as cited in Song & Yin, 2019), which takes the previously mentioned measure and replaces "use of consult" for "employing an internal accountant" (p. 8) to contextualize the measure to the Chinese context. Drawing from personal experience and conversations with practitioners in the U.S. context, the employment of an internal accountant at a nonprofit seems to be more a function of size than of the degree to which processes are formalized, and so I retain the original value of consultants.

In addition, I build on those studies by identifying similar binary variables in the IRS 990 Forms that I anticipate will be indicative of rationalization (USDT, 2019a). Building from the definition of rationalization as the formalization of core and support processes of organizations, I identify eight variables reported in nonprofits' IRS 990 forms that indicate the presence of various organizational characteristics. Table reports the variables, descriptions, and the location of the data within the IRS Form 990.

There is some overlap between the variables presented in Table and the studies by Hwang & Powell (2009) and Suárez & Hwang (2013), specifically, the presence of audited financial statements. This is to ground the expected variables within the work of the previous studies, to test the viability of drawing such data from the IRS Form 990s, and to allow for some comparison of the use of the anticipated variables in constructing the rationalization score.

With the passage of the Sarbanes-Oxley (SOX) Act in 2002, it became a mandatory law in the United States that nonprofits establish policies around whistleblowing and document destruction and protection (Nezhina & Brudney, 2012). Around the time, Ostrower and Bobowick (2006) reviewed a recently released survey of nonprofits and their ability to comply with various elements. They found wide variation in adoption and in perceived compliance to the following dimensions of SOX: independent auditors and committees, producing certified financial statements, disclosing financial information, and having policies around insider trading as well as the previously mentioned whistleblowing and document destruction. Broadly speaking, the imposition of this legal requirement set expectations for nonprofits to formalize these processes to at least the legal minimum. Indeed, Nezhina and Brudney (2012) found evidence to support this happening among a percentage of nonprofits, but they note that not all of these dimensions are legally compulsory. Among the variables in Table are some related to SOX; while compliance may not be a legal

requirement for all, they have become somewhat synonymous with best practices of successful nonprofits.

Table 1. Anticipated Variables that Indicate Rationalization within IRS Form 990

Variable	Descriptions	Form Location	Rationale
Website	Does the organization report a website URL?	Header, K.	Having a functioning website indicates that the organization has invested at least minimally in non-programmatic areas.
Audited Financial Statements	"12a. Did the organization obtain separate, independent audited financial statements for the tax year?"	Part IV, 12a	The organization is sufficiently organized and formalized in accounting practices to be able to provide the necessary documentation to an independent party to prepare the statements.
	"12b. Was the organization included in consolidated, independent audited financial statements for the tax year?"	Part IV, 12b	The organization is a part of, at least for financial purposes, a larger organization. This suggests some degree of additional accountability mechanisms, i.e., to other entities, that would compel formalization of necessary processes.
Conflict of Interest Presence	"Did the organization have a written conflict of interest policy?"	Part VI, Section B. Policies, 12a	Part of SOX
Conflict of Interest Compliance	If the organization had a written conflict of interest policy, "did the organization regularly and consistently monitor and enforce compliance with the policy?"	Part VI, Section B. Policies, 12c	Formalization of a process related to monitoring and enforcement of a (presumably) uncommon occurrence would imply some amount of formalization around more common events.
Whistleblower	"Did the organization have a written whistleblower policy?"	Part VI, Section B. Policies, 13	Part of SOX
Document Destruction	"Did the organization have a written document retention and destruction policy?"	Part VI, Section B. Policies, 14	Part of SOX
	The following description relates to the subsequent two variables: "Did the process for determining compensation of the following persons include a review and approval by independent persons, comparability data, and contemporaneous substantiation of the deliberation and decision?" (Part VI, Section B. Policies, 15)		
Compensation Review – Top Management Team (TMT)	"The organization's CEO, Executive Director, or top management official"	Part VI, Section B. Policies, 15a	The various elements identified (review and approval, using data) describe a formalized process.
Compensation Review – Others	"Other officers or key employees of the organization"	Part VI, Section B. Policies, 15b	The various elements identified (review and approval, using data) describe a formalized process.

Data & Dependent Variable

To create the dependent variable, I approximate the approaches outlined in Hwang & Powell (2009) and Suárez & Hwang (2013). As I do not have interview data for nonprofits, I look to acquire the necessary organization-level data via alternative means. I perform a random sample of nonprofits from the healthcare sector, as indicated by the "Health" major group of the National Taxonomy of Exempt Entities (NTEE) charitable codes (Jones, 2019).

I sample from this group of nonprofits because the strong degree of regulation and coercive isomorphism within this subsector would suggest a greater concentration of shared dominant characteristics among nonprofits, e.g., the extent to which organizational processes are rationalized. In other words, as compared to a subsector where there is less regulatory standardization across processes and organizations, nonprofits operating in the Health space are more likely to experience a consistent, baseline degree of rationalization. This standardization among health service-oriented organizations, regardless of tax-exempt status, is driven by legal standards for care as well as financial incentives. For example, according to the Code of Federal Regulations, organizations that provide inpatient and/or outpatient services, such as hospitals and clinics, must comply with an explicit list of conditions to be eligible for reimbursement from Medicare (Conditions Of Participation for Hospitals, 2011; Conditions Of Participation: Specialized Providers, 2011). This threat of not reimbursing if noncompliance is observed is real because of the number of Medicare patients, 58.6 million out of a total of 332.4 million[9] people in the United States in 2022 (Freed et al., 2022; Moore, 2021). In 2020, the National Health Statistics

[9] The number of Medicare patients reported by Freed et al. (2022) is presume to be real-time calculations as of the date of publication, August 25, 2022. The size of the U.S. population was projected by the U.S. Census Bureau for January 1, 2022. Given the sheer volume of these figures, I do not expect the differences to change noticeably and, furthermore, these figures are presented for illustrative reasons of the volume of patients, moreso than to perform precise calculations.

Group estimated $13,490 in expenditures per Medicare enrollee (Centers for Medicare & Medicaid Services [CMS], 2021). Through its combination of regulatory and financial authority, CMS is able to influence the degree of rationalization within organizations that seek reimbursements from it. Healthcare information security and patient privacy is also a growing concern, even as hospitals balance different priorities to comply with multiple federal and state regulations and to perform well in case of data breaches (Kwon & Johnson, 2013). The sources of regulations include but are not limited to CMS, suggesting that even those Health nonprofits not seeking Medicare reimbursements would face isomorphic pressures that would reduce possible permutations of process steps, i.e., cause rationalization.

The data used to generate the pool of nonprofits comes from electronically filed IRS Form 990s. Guidance on accessing these data is provided by the Nonprofit Open Data Collective (n.d.). The IRS Form 990 is filed by all organizations recognized by the IRS for federal tax exemption that meet at least one of the following two criteria by the end of the tax year, 1) total revenues from all sources greater than or equal to $200,000 and 2) "total assets greater than or equal to $500,000" (USDT, 2019b, p. 3). This form is completed by the filing organization or an authorized representative, and covers the organization's finances, some structural components, and activities.

I selected this set of nonprofit tax forms and not other versions of Form 990 because the two filing criteria allow me to set a threshold for a minimal degree of rationalization within each nonprofit. I anticipate that a nonprofit's ability to successfully generate such revenues and control such assets indicates that there is some degree of formalized processes, such as contracting an auditor to generate forms and sufficient levels of financial monitoring and controls to generate the data necessary to complete the form.

I used the most recent tax filings for the 2019 tax year that nonprofits filed in 2020, representing a total of 381,061 filings for 378,897 nonprofits. If an organization filed multiple returns during this period, then I used the most recent entry. This represents the most recent batch of electronically filed forms, which I select to reduce the possibility of a broken URL or website. I then downloaded the set of Exempt Organization Business Master Files published on October 10, 2022. These files contain information about nonprofits that they do not include in their tax filings, such as their NTEE codes. I then matched each organization's tax filings to their NTEE codes on record using their EIN, a tax ID that is unique to each business entity, regardless of for-profit or not-for-profit (USDT, 2022). After matching on the EIN unique IDs, I then filtered onto nonprofits with an NTEE code that is within the Health major group: E, F, G, and H (Jones, 2019).[10] I then created a subset of these filings for health nonprofits that only included URLs so that any entry could contain a (hopefully) working website, for a total of 6,714 electronic tax filings for Health-related nonprofits. To approximate random sampling, I assigned a number generated randomly to each filing and went through each organization in order of smallest to largest number.

I explored the information available on each organization in their tax forms and on their websites, including annual reports and financial statements, to determine the presence or otherwise of the four variables of "strategic planning, conduct of an independent financial audit, collection of quantitative data for program evaluation, and the use of consultants" (Hwang & Powell, 2009, p. 278). I rationalized the alternative identification steps that I performed to assign values for each organization, as presented in Table 2.

[10] With this methodology, I did not account for 620, 974 out of 1,200,324 nonprofits that are missing an NTEE code. This methodology, therefore, assumes inclusion of an NTEE code as another minimum indicator of rationalization. I acknowledge this as a potential limitation that should and can be addressed in future studies, for example using the methodology outlined in Santamarina et al. (2021).

Table 2. Process to identify binarized variables from non-interview data

Variable	Identification method alternate to coding interviews
Strategic planning	Explicit discussion or presentation of some combination of: vision; mission; values; goals, objectives, strategies, and activities (see Dupree & Winder, 2000, p. 44). On some combination of website, statements, and/or reports.
Independent Financial Audit	Availability of audited financial statements, prepared by independent auditor, via website, either standalone or as part of annual reports.
Quantitative Data for Program Evaluation	Presence of quantitative data in reported program evaluations, outputs, or outcomes as presented in annual reports, in addition to or beyond solely financial data.
Use of Consultants	Mention of hiring consultants in annual reports or financial statements, ex. as a line item in Statement of Functional Expenses. "Consultant" is loosely defined to encompass any external or outside employee offering specialized skills and employed at some point by the organization, other than an independent auditor hired to prepare financial statements.

I reviewed 92 organizations' provided websites and identified 77 that I explored for the presence of the four rationalization variables identified in Table 2. Of these 77 Health-related nonprofits, 69 had some indicator of the presence of strategic planning (at minimum a mission statement), or 89.6% of the total. 20 nonprofits (26.0%) had publicly available financial statements prepared by independent auditors. 44 nonprofits (57.1%) presented quantitative data related to their program evaluations, outputs, or outcomes. 10 (13.0%) reported working with consultants (see Table 2 for guidance on how I defined this).

I replicated the approach presented in Hwang & Powell (2009) to calculate a "factor solution for organizational rationalization" (p. 279) for the coded data. Following the guidance of James et al. (2017), I performed a principal components analysis (PCA) using the prcomp() function in the "stats" package of base R (R Core Team, 2022). As suggested by James et al. (2017) and the guidance documentation for prcomp(), I scaled the variables, with the means and standard deviations of each reported in Table 3. All four variables loaded strongly on one factor, with an eigenvalue greater than one. Table

4 presents the factor loadings for the four variables, the eigenvalue of the factor, and the variance explained by that factor, in the style of Table 1 in Hwang & Powell (2009, p. 279). I then recalculated the factor scores for each observed nonprofit as a z-score, with a mean of 0 and standard deviation of 1, to produce a standardized rationalization factor score in alignment with the previously established method.

Table 3. Means and Standard Deviations for Scaled Variables

Variable	**Mean**	**Standard Deviation**
Strategic planning	0.8961	0.3071
Independent Financial Audit	0.2597	0.4414
Quantitative Data for Program Evaluation	0.5714	0.4981
Use of Consultants	0.1300	0.3383

With the previous method's rationalization factor scores in hand, I matched the values back to their nonprofits, as well as the other 6,637 nonprofit tax records. With the full set of identified organizations in hand, I then merged in the complete set of available efiler data from 2019 for each of the organizations. Table 6 in the Appendix presents the sections in the order they worded added and by filename. The resulting dataset captures a total of 1,704 variables for 6,714 Health-related nonprofits.

Table 4. Factor Solution for Organizational Rationalization, N =77

Variable	Factor Loadings, or Linear Coefficients*
Strategic planning	0.3438
Independent Financial Audit	0.5924
Quantitative Data for Program Evaluation	0.5071
Use of Consultants	0.5231
Eigenvalue	1.87405
Percent of variance	46.85%

* To reduce confusion, I present the nonnegative factor loadings here. The values were reported as negative by prcomp() as negative and as positive by princomp(), an alternative method to perform PCA also in base R. James et al. (2017) note that the values are "unique, up to a sign flip" (p. 382) and that each value specifies a direction such that the sum of squared factor loadings equals one, intentionally constraining the variance.

For additional information on the differences between the two functions, please refer to the Details in the help files for each.

With the previous method's rationalization factor scores in hand, I matched the values back to their nonprofits, as well as the other 6,637 nonprofit tax records. With the full set of identified organizations in hand, I then merged in the complete set of available efiler data from 2019 for each of the organizations. Table 6 in Appendix 3D: Data and Results presents the sections in the order they worded added and by filename. The resulting dataset captures a total of 1,704 variables for 6,714 Health-related nonprofits.

Method for Demonstration of Concept

Which of the 1,704 variables that we have is most meaningful in approximating the four-variable factor score approach to measuring rationalization? To answer this question, I implement a dimensional reduction technique known as the lasso. It uses an l_1-norm penalty with an ordinary least squares estimate to minimize the sum of squares and constrain regressions (Tibshirani, 2011).

This equation has become prolific in computer science and machine learning due to its utility in selecting variables through shrinking estimated coefficients "towards zero relative to the least squares estimate. This shrinking (also known as *regularization*[11]) has the effect of reducing variance" (James et al., 2017, p. 204). Because of its ability to estimate a zero value for coefficients, the lasso is able to provide a form of "automatic feature selection" for models (Hastie et al., 2009). In doing so, it can reduce time and computational resource costs associated with training large, complex models on large, complex datasets, as well as reduce variables that introduce more noise and potential for error than contribute meaningful variance which, in the case of machine learning algorithms, can lead to overfitting the model to the data and reducing its generalizability beyond the training dataset (Manning et al., 2009). Automating feature selection in this way can also be helpful in improving the interpretability and understanding of a model and reducing a large number of possible model features or explanatory variables into a more manageable pool (Fonti & Bellitser, 2017)

For the purpose of this demonstration, I aim to follow the straightforward guidance of Hastie et al. (2021) in their vignette on using their R package glmnet, configured to use lasso. As the authors demonstrate, I used a k-fold cross validation approach, which "uses part of the available data to fit the model, and a different part to test it [where,] we split the data into K roughly equal-sized parts" (Hastie et al., 2009, p. 241). Hastie et al. (2021) recommend to use the value of lambda one standard error away (*lambda.1se*) from the value of lambda that minimizes the Mean-Squared Error (*lambda.mse*; see Figure 2). Given that the values of the two terms appear to be almost identical, and that lambda.mse yields 13 variables of interest as compared to two from lambda.1se,

[11] Emphasis in the original.

I will use the 13 variables with non-zero coefficients when running the model with lambda.mse. These 13 variables are identified in Appendix D, Table 7. Variables generated via the lasso approach.

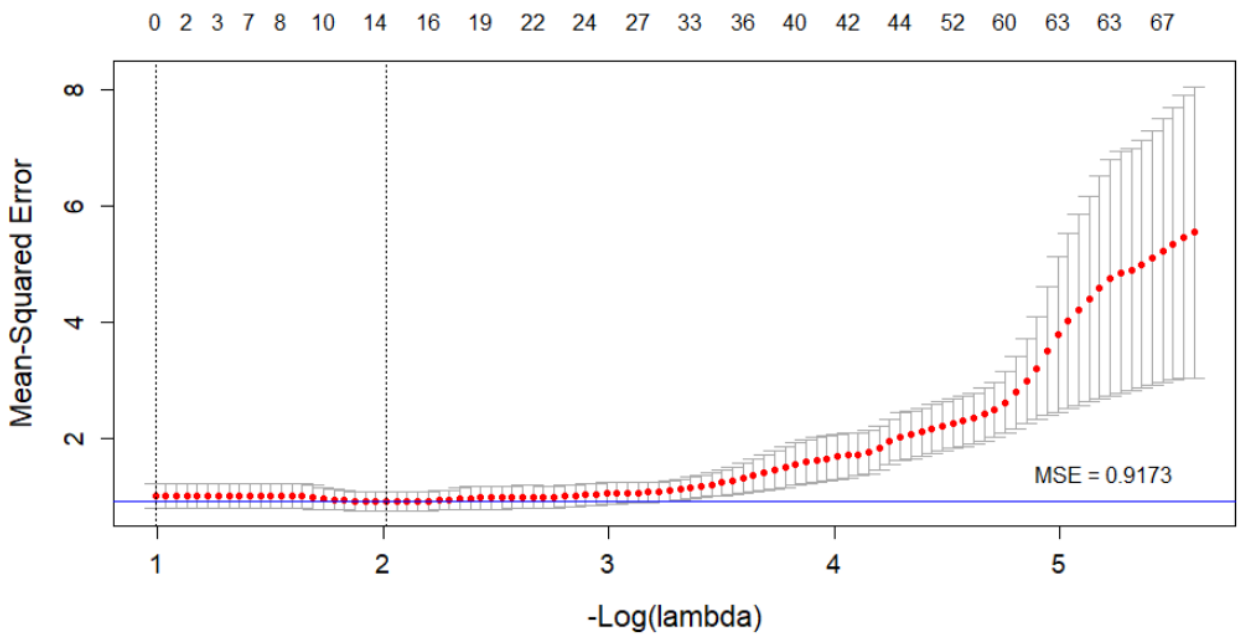

Figure 2. Visualization of the range of variance in potential lambda values. The y-axis indicates the mean of the summed squares of the errors (MSE, y-axis) for each value of lambda on the x-axis. The red dots represent the error at that value of lambda, with the error bars capturing the standard deviation. The rightmost dotted vertical line indicates the value of lambda at which the MSE is minimized, such that the sum of differences between the actual and predicted observations, averaged, is closest to zero. The leftmost dotted vertical line indicates the largest value of lambda within one standard error of the lambda that minimizes the MSE. Refer to Hastie et al. (2021) for further clarification. The blue line indicates the MSE value for the value of lambda at which the MSE is minimized, which is the value reported in the figure.

Testing the rationalization score

I test the algorithm-derived measure against the original methodology, the literature-derived methodology, and a mixture of both algorithm- and literature-derived variables. Returning to the

previous approach for generating rationalization scores, I have identified 21 variables to assist in measuring rationalization, the 13 predicted variables identified via the lasso approach (Appendix D, Table 7. Variables generated via the lasso approach) and the 8 anticipated variables that I identified from the 990 form and its instructions (see Table). I compare a factor score generated with only the variables identified by the lasso algorithm ("predicted variables"), only the anticipated variables derived from theory ("anticipated variables"), and a combined set consisting of both the predicted and anticipated variables ("combined set"). As PCA requires inputs to be numeric, variables that were purely text based (such as addresses or names) were removed. I only used the factor with the highest eigenvalue for each to aid in comparison with the baseline rationalization factor score, which also only used one factor.

In keeping with Figure 2, I report my findings in terms of MSE. The predicted variables score yielded an MSE of 0.8276, the anticipated variables score yielded an MSE of 2.4977, and the combined-set score yielded an MSE of 1.1409. Of the three sets of scores, the predicted variables score performed the best relative to the original method of rationalization factor scores. When looking at the relationships between the variables in the factor used in the combined set PCA, the top five most heavily weighted variables are the anticipated variables related to conflict-of-interest policy, conflict of interest monitoring, whistleblower policy, document retention policy, and whether financial statements were independently audited. The sixth-most weighted variable is the predicted variable Schedule J, which covers compensation information for top performers and practices within the nonprofit (USDT, 2021). This suggests that enhancing theory-derived variables with ones identified by dimension reduction algorithms can improve overall score performance.

This training dataset was trained on a limited pool of nonprofits, that was simultaneously quite diverse, containing hospitals, emergency medical services, healthcare professional groups, addiction treatment and support, elder care, and more. Given that the predicted variables score performed noticeably better than either of the other two, I apply it to the entire set of 6,714 Health-related nonprofits. In subsectors that experience high degrees of isomorphism, I expect to observe that the distribution of the rationalization factor has heavy (or "fat") tails and a high peak, i.e., leptokurtosis. This would indicate that many values fall somewhere in the center of the distribution and do not fall in the tails. In other words, it would indicate that the majority of rationalization factor values for individual organizations are closer to the mean than in a normal distribution.

As described by Breunig (2006), L-kurtosis is a measure of the shape of a distribution ranging from 0 to 1, where higher values indicate more kurtosis. This technique is a linear approach that is conservative ("robust to very extreme observations"), applicable to small sample sizes, and can be used to compare different variables (Jensen, 2009, p. 295). The normal distribution should have a L-kurtosis measure, or fourth L-moment ratio, value of 0.1226 (ex. Hosking, 1990, p. 112). Values above this normal distribution threshold will indicate the presence of isomorphic effects. The degree to which isomorphism is present can then be compared across nonprofit subsectors or other analytic perspectives/subdivisions. Given the overall high concern in the sector that I observed relative to privacy and patient confidentiality, I expect to see some degree of leptokurtosis and not a normal distribution. I assigned rationalization scores to the 6,714 Health-related nonprofits and found evidence suggesting that the distribution is skewed, as seen in Figure 3. To adjust for this behavior, I imposed an upper threshold or cut-off value of z-score < 2.5, not inclusive.

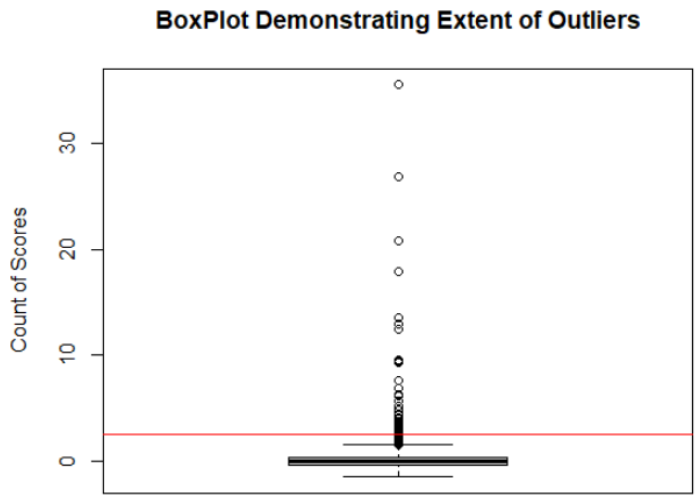

Figure 3. The distribution of the standardized rationalization scores appears skewed. The red horizontal line shows the cut-off upper threshold of z-scores that I set, z-score < 2.5 (not inclusive).

To develop a visual understanding, I plotted the distribution of scores for the remaining 6,670 nonprofits, as seen in Figure 4. When the histogram bins are defined more narrowly (ex. moving from a bin width of 0.5 in graph 1 to 0.1 in the other three), there appear to be more than one mode in the distribution. To confirm whether the distribution is truly multimodal, I used the R package LaplacesDemon to perform logical checks and identify possible local modes before and after setting the maximum score threshold, which yielded four values in both cases (Statisticat, LLC., 2021).

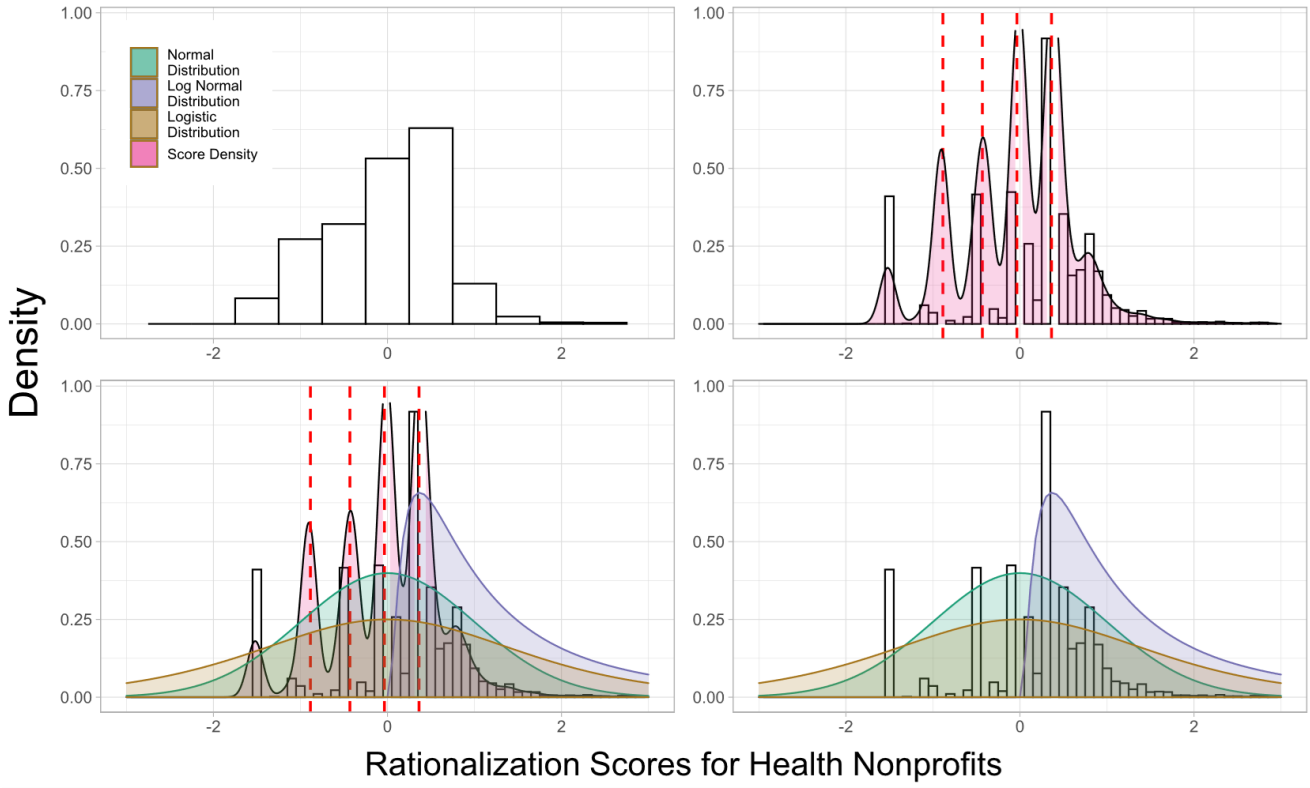

Figure 4. Presentation of the range of calculated rationalization scores for Health-related nonprofits. In clockwise order, Graph 1 shows the distribution of values with bin widths of 0.5. Graph 2 shows how the distribution of data changes when using a smaller bin width of 0.1. The vertical red lines indicate the calculated modal points, and the density plot, in light red, shows overall trends in the distribution of the data. Graph 3 shows the distribution using 0.1 bin widths, with three common distributions overlaid: normal distribution (light green), logistic distribution (light brown), and log normal distribution (light purple). Graph 4 combines graphs 2 and 3, such that the three distributions are juxtaposed against the score density plot and vertical modal indicator lines.

The four modes within the data suggest that there could be clusters of shared experiences of rationalization among Health-related nonprofits. Some organizations seem to be highly rationalized, as indicated by the 44 positive outlier values. One of the modal values, -0.0108, is almost 0, which aligns with my expectation that a large number of nonprofits working in Health

would be similarly rationalized. Two of the modal values are almost equidistant from 0 at 0.3754 and -0.4233. This suggests that there are also clusters of nonprofits that are slightly more and less rationalized than the majority. The fourth modal value, -0.9060, suggests that a small but significant number of Health-related nonprofits are under-rationalized as compared to peer organizations. While these four values were calculated after I removed the 44 outliers, they could still be influenced by the larger variety of organizations being more rationalized than the mean, as evidenced by the tail of the distribution skewing positively. There are two smaller clusters within the distribution that may be potential local modal clusters, around -1.5 and 1. They are each evidence of the non-normal distribution of rationalization within Health, which may also be an artifact of the diversity of nonprofits included in this NTEE Major Area 10 code.

As presented in Appendix E, Health consists of four different Major Area 26 groups. 4,070 nonprofits in the final, cleaned training dataset identified the focus of their activities[12] as "E Health - General and Rehabilitative," 1,714 identified "F Mental Health, Crisis Intervention," 669 as "G Diseases, Disorders, Medical Disciplines," and 261 as "H Medical Research." Table 5 explores which nonprofits are represented within the ranges of the four modes. I defined a mode's range as the mode +/- 0.1930887, half of the smallest distance between the four modes. As a result, only 5,607 of the 6,670 nonprofits with z-scores < 2.5 are included. Most E-group nonprofits appear to be rationalized at average or greater than average levels, given their concentration around the rightmost modes. F-group nonprofits were generally distributed across the modes, with a slight concentration observable around the mode closest to 0, the average. G-group nonprofits were similarly distributed across modes. Both F- and G-group nonprofits were noticeably reduced

[12] Nonprofits self-identify which NTEE code reflects their primary purpose upon application for tax exempt status. For a deeper discussion, please refer to Santamarina et al. (2021).

around the leftmost mode, suggesting that they tend towards more rationalization than not. H-type nonprofits were also generally distributed across modes, but with a noticeable concentration around the rightmost mode.

Table 5. Mix of Nonprofits by Activity per Mode

NTEE Major 26 Group	**~ -0.906**	**~ -0.423**	**~ -0.011**	**~0.375**	**Total**
E Health – General and Rehabilitative	444	508	1046	1344	3342
F Mental Health, Crisis Intervention	269	388	502	392	1551
G Diseases, Disorders, Medical Disciplines	113	131	149	137	530
H Medical Research	45	34	45	60	184
Sum	*871*	*1061*	*1742*	*1933*	*5607*

Discussion

This section explores the results' implications for which rationalizing contributions would be most beneficial, followed by an exploration of model performance and considerations.

What Results may mean for Contributions

The findings from the score test suggest implications for the kinds of contributions that health-related nonprofits would most benefit from, when considering the potential for rationalization to affect nonprofit performance.

Nonprofits that have experienced a greater-than-average degree of rationalization, such as the potential cluster around score = 1 and the rightward tail of nonprofits that are even more rationalized, would benefit from contributions that introduce completely novel knowledge. This implies that donors to these nonprofits should be specialists in their fields. Contributions that introduce completely new permutations for existing processes may reduce rationalization and the

potential negative effects resulting from over-formalization within the organization. As these organizations' rationalization score decreases, I expect potential benefits to increase from contributions that reduce rationalization. At the same time, contributions should be carefully screened to ensure that there is an appropriate degree of fit between what the organization needs, increased flexibility, and the intended outcome from the donor's professional service.

Nonprofits with a less-than-average degree of rationalization, in particular the mode around score = -0.906 and the potential mode near -1.5, may experience greater benefits from contributions that overlap heavily with existing knowledge and processes than from novel knowledge. In particular, contributions that formalize any kinds of processes that exist should increase rationalization and could reduce resource inefficiencies resulting from too much variation. Donors who are subject-matter generalists will likely be as helpful as experts, if not more, depending on the complexity of the targeted processes. For these nonprofits, I expect that any contribution that surpasses a minimum threshold of fit will increase rationalization, but that the burden will be more on the recipient nonprofit to be able to receive the donation and achieve the intended outcome from the contribution. In general, as contribution complexity increases, I anticipate fit to decrease, reducing potential gains in nonprofit performance.

Nonprofits with a (close to) average degree of rationalization, i.e., the three modes of 0.3754, -0.0108, and -0.4233, would benefit from contributions that enhance existing knowledge. This represents the majority of nonprofits in the sample, which suggests that contributions of professional services to health-related nonprofits should be flexible in terms of providing novel knowledge (or new processes, thus reducing rationalization) and enhancing existing knowledge (reducing variation in existing processes, thus increasing rationalization). Donors who are subject matter experts will be able to most effectively alternate between these two realms. That said, since

there are so many nonprofits that fall in this liminal space, careful scoping of engagements is necessary to be able to appropriately draw from the limited pool of specialist donors as compared to the broader pool of generalist donors. I anticipate that engagement scoping will also moderate the importance of fit in determining how the contribution affects nonprofit performance.

These findings may suffer from underlying bias in the data assignment process due to incomplete data. Some of the organizations that had an original-method rationalization score of less than 1 had signals on their websites that they were much more rationalized than other organizations.[13] As a result of these biases in the training dataset, the lasso algorithm may incorrectly identify a large hospital and a small addiction treatment facility as having the same rationalization score, rather than correctly identifying that the large hospital has the same score as another hospital. There may also be legal requirements that did not affect all of the sampled nonprofits, such as having certain overdose medications on hand. Unaffected nonprofits may therefore be less rationalized as compared to the rest of the sample, which could also introduce bias and cause some of the observed variation to be due to categories of health-related work rather than degrees of rationalization.

Score Performance

According to the findings presented earlier, the method to generate the rationalization score seems to be effective. The multimodal distribution is not wholly unsurprising, given the diversity of organizations in the coded dataset. In particular, the use of the Health nonprofit NTEE major code group may have introduced too much variance in the data. When looking at solely the organizations coded to calculated the original rationalization factor score, they included health clinics,

[13] Drawing from personal experiences working with small-, medium-, and large-size hospitals. For example, hospitals above certain income levels and bed counts tend to hire consultants to assist with certain strategic initiatives. These contracts may not be consistently recognized as "consultants" or the like in publicly available reports or financial statements, which would result in the organization receiving a 0 or "no presence observed" for that dimension of the rationalization factor score.

counseling services, healthcare workers' professional societies, volunteer health services providers, and addiction recovery centers, among other kinds of Health-affiliated nonprofits (see Appendix E). That said, the high score performance when applied to the larger dataset that was dominated by nonprofits that identified as "E Health – General and Rehabilitative," 4,070 out of 6,714 nonprofits, suggests that using a diverse "tuning" dataset, in this case to perform the original factor score calculation, to capture variation across the full dataset will yield positive overall performance.

Each of these organizations may have logical, clear reasons to not present or publicize some pieces of information, and the coding process may have introduced unanticipated bias because such reasons were not accounted for, such as quantitative data on services delivered. In the Health space, this lack of information can be multifaceted: elements of the data may be seen as confidential and a part of patient privacy, but also with consequences for employees' capacities to deliver services, ex., for providers of abortion services. On the other hand, not making information available may be done to contribute towards competitive market advantages and not exposing potential weaknesses or critiques, as in the case of larger hospitals, nursing homes, and professional associations that focused on organizing events.

There is also a more expected concern, that of nonprofit resource constraint. For example, volunteer emergency medical service nonprofits may not have the organizational capacity or resources to dedicate to publishing meeting notes, annual reports, financial documents, and/or tax returns on their websites. The organizations most focused on transparency in that regard were community-oriented clinics.

Similarly, an organization's website and data that is publicly available there may not be a good indicator of overall organizational rationalization. A large hospital chain that does not have audited

financial statements publicly available is, nonetheless, more likely to be formalized in its processes and acceptable permutation of steps than some of the other nonprofits in the Health space because of the sheer volume of regulations that they face. There were also numerous instances where a smaller nonprofit had done a good job of making public reports, meeting notes, and the like, but with a sudden decline or even cessation of this practices starting around 2018 to 2019. In other words, the effects of the pandemic may have affected some processes whose outputs, in turn, were used to measure overall rationalization; such missing data could exaggerate the extent or lack of rationalization, which could be another source of bias.

These various examples underscore both the wide array of organizations in the dataset as well as create encouragement that there could realistically be a multimodal distribution in the space, rather than the distribution emerging solely from issues with the measurement constructs.

I was surprised at how the lasso-predicted variables and literature-derived or anticipated variables interacted when loaded together in the principal components analysis. The finding that a mix of theory-derived and algorithm-predicted variables can yield high performance suggests that some variables (i.e., theory derived) may be expected to have a relatively steady utility in predicting or measuring rationalization across various nonprofits sectors, and a combination of both drawing on theory and discerning patterns from data can improve analytic capacities and resolve long-standing issues. Given the sheer volume of qualitative text data in nonprofit tax returns, the inability of PCA to leverage text data indicates a need to look for alternative analytic methods and modeling to assign rationalization scores once variables have been analyzed, consider how to implement methods to transform that text into meaningful quantitative data, or a combination of both.

Conclusions

This paper presents a method on how to develop a flexible, customizable rationalization measure or score. The measure was used to support theory-derived expectations that Health nonprofits are rationalized in some similar ways, while also suggesting trends within the space that merit additional investigation. This measure seems to be a viable way of measuring changes in organizational rationalization resulting from engagements with donors who contribute skills, and could serve as a useful performance outcome and dependent variable in models. I identify three key areas of improvement for the method that should be considered for later studies.

Future implementations of this model should consider alternative ways of calculating the score. For example, Liu et al. (2020) have published an R package called HDCI that allows one to create meaningful OLS models using variables selected via lasso. While there is some debate as to how to interpret a model created using variables that were calculated with a focus on sparsity, the integrated approach could allow for a direct calculation of a nonprofit's rationalization score using lasso-selected variables. This would present a smoother workflow than evidenced here and with greater ease of interpretation for many scholars, as the final output would be in commonly understood OLS terms.

Future extensions of this project should include a rationalization measure trained on a dataset where the existing score has been assigned to a much larger sample of nonprofits. Taking two different approaches would yield different but important findings. For example, creating a larger, sector-specific sample that includes a wider array of nonprofits could help develop differences in creating a measure that is topic-specific and accounts for such nonprofits' unique considerations. On the other hand, creating a sample of organizations with similar characteristics but operating in

different spaces could shed light on how sectors' idiosyncratic isomorphic pressures result in varying degrees of rationalization.

Thirdly, creating a rationalization measure that can be generalized across all nonprofits in a particular geography, such as the United States, should be performed via stratified random sampling, where the strata reflect as many categories of nonprofits in the geography as possible. This approach is likely to yield some quintessential categorization issues and challenges, which can be resolved through methods such as using machine learning to assign the unidimensional, single-value NTEE codes to all nonprofits' returns in a year of efiled data. This approach should leverage the findings of Santamarina et al. (2021) to identify, for each nonprofit, the probability match for that nonprofit to each NTEE (major group) code. Once these probabilities are calculated, researchers can then identify the nonprofits that are most "purely" of one code and nonprofits that are evenly split between two or more codes. This would create a sampling framework with strata or groups of nonprofits that are more accurately representative of each code or code combination. Sampling from these multi-dimensional strata would allow researchers to have a richer and more accurate representation of the range of nonprofits, such that random sampling from each stratum would yield organizations that are more internally similar and externally distinct to the other strata.

Regardless of the stratification and sampling approaches used, future studies should consider ways to take advantage of more complex and robust methodologies to create scores that address nonprofits' contexts, characteristics, and the work that they perform. Doing so can shed additional light on the ways that rationalization can enhance or diminish nonprofit capacities and provision of services and social capital.

References

Akingbola, K. (2013). A model of strategic nonprofit human resource management. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *24*(1), 214-240.

Barney, J. B. (1991). Firm Resources and Sustained Competitive Advantage. *Journal of Management*, *17*(1), 99-120.

Breunig, C. (2006). The more things change, the more things stay the same: A comparative analysis of budget punctuations. *Journal of European Public Policy*, *13*(7), 1069-1085.

Callen, J. L. (1994). Money donations, volunteering and organizational efficiency. *Journal of Productivity Analysis*, *5*(3), 215-228.

Centers for Medicare & Medicaid Services (2021). National Health Expenditure Data, Table 21 Expenditures, Enrollment and Per Enrollee Estimates of Health Insurance, United States, Calendar Years 1987-2020 [Data file]. Retrieved from: https://www.cms.gov/files/zip/nhe-tables.zip

Conditions Of Participation for Hospitals, 42 C.F.R. § 482 (2011)

Conditions Of Participation: Specialized Providers, 42 C.F.R. § 485 (2011)

Cordery, C. J., Proctor-Thomson, S. B., & Smith, K. A. (2013). Towards communicating the value of volunteers: lessons from the field. *Public Money & Management*, *33*(1), 47-54.

Cordery, C., & Narraway, G. (2010). Valuing volunteers: expanding the relevance and reliability debate. *Australian Accounting Review*, *20*(4), 334-342.

Coupet, J., & Berrett, J. L. (2019). Toward a valid approach to nonprofit efficiency measurement. *Nonprofit Management and Leadership*, *29*(3), 299-320.

Dart, R. (2004). Being "business-like" in a nonprofit organization: A grounded and inductive typology. *Nonprofit and voluntary sector quarterly*, *33*(2), 290-310.

Dupree, A. S., & Winder, D. (2000). Choosing Structure and Mission. In *Foundation Building Sourcebook: A practitioners guide based upon experience from Africa, Asia, and Latin America* (pp. 38-50). New York: *The Synergos Institute*.

Einolf, C. J., & Yung, C. (2018). Super-Volunteers: Who Are They and How Do We Get One?. *Nonprofit and Voluntary Sector Quarterly*, *47*(4), 789-812.

Financial Accounting Standards Board (FASB). (2008). *Statement of Financial Accounting Standards No.116: Accounting for contributions received and contributions made.* Original Pronouncements as Amended. Retrieved from https://www.fasb.org/jsp/FASB/Document_C/DocumentPage?cid=1218220128831&acceptedDisclaimer=true

Fonti, V., & Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam research paper in business analytics*, *30*, 1-25.

Freed, M., Fuglesten Biniek, J., Damico, A., & Neuman, T. (2022, August 25). *Medicare Advantage in 2022: Enrollment Update and Key Trends*. KFF (Kaiser Family Foundation). https://www.kff.org/medicare/issue-brief/medicare-advantage-in-2022-enrollment-update-and-key-trends/

Frumkin, P. (2012). The Idea of a Nonprofit and Voluntary Sector. In J. S. Ott and L. A. Dicke (Eds.), *The nature of the nonprofit sector* (pp. 17-30). Boulder, CO: Westview Press.

Gordon, T. P., Khumawala, S. B., Kraut, M., & Neely, D. G. (2010). Five dimensions of effectiveness for nonprofit annual reports. *Nonprofit Management and Leadership*, *21*(2), 209-228.

Han, J., Jo, G. S., & Kang, J. (2018). Is high-quality knowledge always beneficial? Knowledge overlap and innovation performance in technological mergers and acquisitions. *Journal of Management & Organization*, *24*(2), 258-278.

Hastie, T., Qian, J., & Tay, K. (2021). An Introduction to glmnet. *CRAN R Repositary*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.) Springer Science & Business Media.

Hosking, J. R. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, *52*(1), 105-124.

Hosking, J. R. M. (2022). L-Moments. R package, version 2.9. URL: https://CRAN.R-project.org/package=lmom.

Hwang, H., & Powell, W. W. (2009). The rationalization of charity: The influences of professionalism in the nonprofit sector. *Administrative science quarterly*, *54*(2), 268-298.

Independent Sector. (2022, April 18). *Value of Volunteer Time*. https://independentsector.org/resource/value-of-volunteer-time/

Internal Revenue Service. (2010). Instructions For Requesting Information On Exempt Organizations (Effective January 2010). Published January, 2010. https://www.irs.gov/pub/irs-tege/p4838.pdf

Itami, H. (1991). Invisible assets. In *Mobilizing invisible assets* (T.W. Roehl, Trans.). Harvard University Press. (Original work published 1984)

Iyer, E. (2003). Theory of alliances: Partnership and partner characteristics. *Journal of Nonprofit & Public Sector Marketing*, *11*(1), 41-57.

Jensen, C. (2009). Policy punctuations in mature welfare states. *Journal of Public Policy*, 287-303.

Jones, D. (2019). IRS activity codes. Published January 22, 2019. https://nccs.urban.org/publication/irs-activity-codes.

Jones, G. J., Edwards, M., Bocarro, J. N., Bunds, K. S., & Smith, J. W. (2017). Collaborative advantages: The role of interorganizational partnerships for youth sport nonprofit organizations. *Journal of Sport Management*, *31*(2), 148-160.

Kioko, S., & Marlowe, J. (2016). Financial Strategy for Public Managers. Rebus.

Kwon, J., & Johnson, M. E. (2013). Security practices and regulatory compliance in the healthcare industry. *Journal of the American Medical Informatics Association*, *20*(1), 44-51.

Lindenberg, M. (2001). Are we at the cutting edge or the blunt edge?: Improving NGO organizational performance with private and public sector strategic management frameworks. *Nonprofit Management and Leadership*, *11*(3), 247-270.

Liu, H., Xu, X., & Li, J. J. (2020). A bootstrap lasso+ partial ridge method to construct confidence intervals for parameters in high-dimensional sparse linear models. *Statistica Sinica*, *30*(3), 1333-1355.

Maier, F., Meyer, M., & Steinbereithner, M. (2016). Nonprofit organizations becoming business-like: A systematic review. *Nonprofit and Voluntary Sector Quarterly*, *45*(1), 64-86.

Manning, C. D., Schütze, H., & Raghavan, P. (2009). *Introduction to information retrieval*. Cambridge university press. Online edition. Retrieved from https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

McKeever, B. S., & Pettijohn, S. L. (2015). The nonprofit sector in brief 2015. *Public Charities, Giving and*.

Moe, T. M. (1991). Politics and the Theory of Organization. *JL Econ. & Org.*, *7*, 106.

Moore, D. (2021, December 30). *U.S. Population Estimated at 332,403,650 on Jan. 1, 2022*. U.S. Census Bureau. https://www.census.gov/library/stories/2021/12/happy-new-year-2022.html

Nezhina, T. G., & Brudney, J. L. (2012). Unintended? The effects of adoption of the Sarbanes-Oxley Act on nonprofit organizations. *Nonprofit management and leadership*, *22*(3), 321-346.

Nonprofit Open Data Collective. (n.d.) "Open Data for Nonprofit Research." Retrieved from https://nonprofit-open-data-collective.github.io/overview/

Ostrower, F., & Bobowick, M. J. (2006). Nonprofit governance and the Sarbanes-Oxley act. *Urban Institute National Survey of Nonprofit Governance Preliminary Findings*.

Paynter, S., & Berner, M. (2014). Organizational capacity of nonprofit social service agencies. *Journal of health and human services administration*, 111-145.

Prentice, C. R. (2016). Why so many measures of nonprofit financial performance? Analyzing and improving the use of financial measures in nonprofit research. *Nonprofit and Voluntary Sector Quarterly*, *45*(4), 715-740.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Risse, T. (2010). 11 Rethinking advocacy organizations? A critical comment. In A. Prakash and M. K. Gugerty (Eds.) *Advocacy organizations and collective action* (pp. 283-294). Cambridge University Press.

Santamarina, F. J., Lecy, J. D., & van Holm, E. J. (2021). How to Code a Million Missions: Developing Bespoke Nonprofit Activity Codes Using Machine Learning Algorithms. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 1-10.

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., ... & Jinks, C. (2018). Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & quantity*, *52*(4), 1893-1907.

Sears, J., & Hoetker, G. (2014). Technological overlap, technological capabilities, and resource recombination in technological acquisitions. *Strategic Management Journal*, *35*(1), 48-67.

Sirmon, D. G., & Hitt, M. A. (2009). Contingencies within dynamic managerial capabilities: Interdependent effects of resource investment and deployment on firm performance. *Strategic management journal*, *30*(13), 1375-1394.

Song, C., & Yin, J. (2019). "The advancing of management": Cross-sector agents and rationalization of nonprofits in Eastern China. *Nonprofit Management and Leadership*, *29*(4), 529-548.

Statisticat, LLC. (2021). LaplacesDemon: Complete Environment for Bayesian Inference. Bayesian-Inference.com. R package version 16.1.6. https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software

Suárez, D. F., & Hwang, H. (2013). Resource constraints or cultural conformity? Nonprofit relationships with businesses. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *24*(3), 581-605.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 73(3), 273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.x

Tsarenko, Y., & Simpson, D. (2017). Relationship governance for very different partners: The corporation-nonprofit case. *Industrial Marketing Management*, *63*, 31-41.

U.S. Bureau of Labor Statistics. (2022, November 4). *Table B-3. Average hourly and weekly earnings of all employees on private nonfarm payrolls by industry sector, seasonally adjusted*. Retrieved November 6, 2022, from https://www.bls.gov/news.release/empsit.t19.htm

U.S. Department of the Treasury, Internal Revenue Service. (2019). *Form 990 Return of Organization Exempt From Income Tax* (Cat. No. 11282Y). Retrieved from https://www.irs.gov/pub/irs-pdf/f990.pdf

U.S. Department of the Treasury, Internal Revenue Service. (2021). *2021 Instructions for Schedule J (Form 990): Compensation Information*. Retrieved from https://www.irs.gov/pub/irs-pdf/i990sj.pdf

U.S. Department of the Treasury, Internal Revenue Service. (2022. October 25). *Apply for an Employer Identification Number (EIN) Online*. https://www.irs.gov/businesses/small-businesses-self-employed/apply-for-an-employer-identification-number-ein-online

Van de Ven, A. H., & Drazin, R. (1984). *The concept of fit in contingency theory*. Strategic Management Research Center, University of Minnesota.

Volberda, H. W., van der Weerdt, N., Verwaal, E., Stienstra, M., & Verdu, A. J. (2012). Contingency fit, institutional fit, and firm performance: A metafit approach to organization–environment relationships. *Organization Science*, *23*(4), 1040-1054.

Appendices

Appendix 3A: Identifying Who is a Donor

FASB has specific guidance around reporting contributions of services in financial statements (FASB, 2008). Rather than distinguishing volunteers by amount of dedicated time (ex. volunteers vs. super-volunteers), I draw on the language from FASB's guidance to focus on volunteers who contribute their professional knowledge and services. Professional services are recognized if they "(a) create or enhance nonfinancial assets or (b) require specialized skills, are provided by individuals possessing those skills, and would typically need to be purchased if not provided by donation" (FASB, 2008, section 9). In other words, they cannot be provided by the average member of a population, and are rare, not easily imitated, and not easily replaceable (Barney, 1991, 1995). This difference is also visible in generalized financial estimates: the average estimate of a volunteer hour for 2021 falls between $23.77[14] to $29.95 an hour, whereas the average estimate of a private, nonfarm employee's hourly pay in October 2011 is $31.11 (Independent Sector, 2022; U.S. Bureau of Labor Statistics, 2022). The difference increases even further when looking at pay for people who may volunteer to build a website or prepare financial statements.[15] Because this class of volunteer should have their contributions and time valued at a different scale than general volunteer activities, I refer to such individuals as donors and their actions as donations or contributions of professional services.

[14] Projected using a straight-line calculation from values provided for 2007-2013 in Table 6 of McKeever & Pettijohn (2015, p. 14).
[15] Respectively, "Information" was listed as at $44.60 and "Financial Activities" as $40.55 in October 2021 (U.S. Bureau of Labor Statistics, 2022).

Appendix 3B: Measuring Non-Financial Outcomes

In this section, I illustrate how a possible solution is an inappropriate measure for capturing the effects of rationalizing in-kind contributions on nonprofit performance.

Maier et al. (2016) note a distinction between an organization's performance and its fulfillment of societal functions, further defining performance as "understood within the NPO's own frame of reference, that is, the fulfillment of its mission and the securing of financial and human resources" (p. 75). To that end, I propose a subjective performance measure using year-to-year changes in the key outcome(s) reported by nonprofits in their annual reports, standardized across all nonprofits in the sample. A key outcome is one that explicitly aligns with the nonprofit's stated mission and is not financial. Measures that exist in at least two years' worth of reports will be extracted, and the percent change in each measure will be calculated using the formula,

$$\frac{New\ Value - Old\ Value}{Old\ Value} \times 100. \qquad (1)$$

The percent changes will be averaged across measures and time periods by nonprofit. The organization-level averages will then be recalculated as Z-scores to standardize values across the organizations in the database.

Consider a child nutrition nonprofit that provides elementary school students with food-filled backpacks, and has reported the number of provided backpacks in its annual reports for fiscal years 2016 through 2019. A supply chain management expert engaged with the nonprofit in January 2019, donating their experience to identify potential bottlenecks in acquiring the materials and supplies for backpacks and ways to resolve them. To determine the effect of the donated service

on the key outcome of provided backpacks, we could calculate the percent change in the number of backpacks using Equation 1:

$$\frac{Backpacks_{2019} - Backpacks_{2018}}{Backpacks_{2018}} \times 100\,, \qquad (2)$$

which captures the year-to-year change in one key outcome measure reported by the nonprofit before and after receiving the donated service.

Numerous challenges exist with an approach like this one, which is intended to grant flexibility and subjectivity to understanding performance. A nonprofit must have at least two years' worth of reports available for the calculation to function. When the nonprofit receives contributed services, at what point relative to when it produces the annual reports, could also affect calculations: if the service is not completely implemented in year *t* but is by the middle of year *t+1*, then values for year *t* will not be reliable for calculating before-and-after effects. More years of data would be helpful in establishing a more precise calculation of true, overall outcomes, but could also introduce bias from unobserved or exogenous influences, such as market fluctuations that reduced outcomes in one year more so than in other years.

There are also broader, technical limitations. Some nonprofits may have multiple key outcomes; in such cases, the researcher will have to determine how to weight them. Equal weights may be mathematically easier but not truly representative of the organization's mission, vision, and values, introducing measurement bias. Mission statements are almost always present in annual reports, but they are not static. If a mission statement changes year to year such that the key outcome varies, then the calculation will not be possible for that nonprofit.

Even more critically, nonprofits do not always include key outcomes in annual reports in clearly quantified terms. Gordon et al. (2010) reviewed annual reports for 75 nonprofits and found that 97.3% included narratives of accomplishments but quantitative data was much more limited: only 36% included physical output measures and 33.3% included number of service recipients (Table 3, pg. 222). In addition, nonprofit activities may not be easily measured numerically, introducing additional bias into calculating changes in performance. The proposed non-financial measure therefore rests solely on values that are inconsistently present in the data.

Appendix 3C: Illustrating Minimal and Maximal Rationalization

Minimal Rationalization

A *minimally rationalized organization* would not have formalized processes or procedures for performing portions of the production process and delivering internal or external outputs, such as measuring and reporting on outcomes, organizing events, or tracking inventory. This organization would create processes ad hoc; while there might be some duplication or replication of processes, it would be on an informal basis and constrained to individuals or groups of individuals that work together. For example, someone whose role within the organization include tracking inventory may do so in roughly the same way on two separate occasions, but the focus is more on the overall output of finishing the inventory rather than performing the individual steps. They may also teach another employee how to track inventory in this way but not transmit formal, explicit guidance on which steps must be completed and "acceptable" permutations of steps for an inventory to be considered appropriately completed. Inefficient use of resources to achieve outputs is highly probable due to reasons such as duplication of steps, performing unnecessary steps or in the wrong order, or including steps that are resource intensive when there are less expensive alternative steps. What this suggests is that a minimally rationalized organization would struggle to pinpoint how it uses its resource to achieve outputs, with consequences for outcome measurement and understanding its efficacy and efficiency.

Another anticipated characteristic is a lack of shared understanding between employees and leadership as to what the process output should look like. When processes are not resource expensive or the organization is not resource constrained, the consequences of mismatched expectations on process outputs may be negligible, for example, stacking shovels in the wrong location and then having to move them. But, as processes become more resource expensive and

the organization more resource constrained, the consequences increase in severity, for example, lost funding opportunities because of missed or incomplete grant applications. While minimally rationalized organizations suffer from a lack of consistent structure, they would see some benefits in responding to internal and external shocks due to their capacity to create and implement as-needed processes. This possible benefit would also be highly dependent on other internal characteristics, such as skills and competence of employees, shared understanding of vision, and various other factors that may not actually be achievable in an organization with minimal rationalization.

Maximal Rationalization

A *maximally rationalized organization* would have formalized processes and procedures for delivering all possible internal and external outputs produced by the organization. Standard operating procedures, process guides, manuals, and other forms of knowledge management would be essential for employees to know which steps to perform and in what orders to achieve outputs. This implies a high degree of administrative structure including systems for oversight and compliance to ensure that only the "officially" recognized process permutations are being performed by all members of the organization. The act of reducing possible steps and permutations will constrain innovation and flexibility for the purpose of yielding more resource efficiency through reducing duplication, minimizing the total pool of possible resources through only performing steps in orders that the organization recognizes as acceptable, and creating more opportunities for inputs to be standardized such that there are less unique resources and the total resource pool can be used more flexibly. Instead of each input only used for one possible step, multiple steps can use the same input, allowing the organization to reduce costs by sourcing that input at scale and, presumably, at discounted costs. Thanks to the potential for improved input and

resource accounting, a maximally rationalized organization should be able to more easily attribute each consumed resource to the outputs produced from its processes for reporting purposes, in response to audits, and to demonstrate greater accountability and transparency. Such organizations should be well positioned to measure their outcomes and understand their efficacy and efficiency, but measuring progress towards these concepts is not the same as achieving them.

Too much rationalization may codify steps and permutations that are not the most efficient or effective. When processes are not resource expensive or the organization is not resource constrained, the consequences of inefficient or outdated steps and permutations on outputs may be negligible, for example, generating a report using multiple complex computer programs when it can be produced using just one simple computer program. As processes become more resource expensive and the organization more resource constrained, these consequences increase in severity, for example, multiple stages of internal sign-off and review causing delays that lead to missed or incomplete grant applications. The reduction in flexibility and innovation could also insulate the organization such that it is slow or unable to respond to external changes, such as shifting consumer demands and beneficiary needs, new legal requirements, or changing donor expectations.

Appendix 3D: Data and Results

Table 6. Sections of Form 990 in order of compilation and by file name

Order	File Name	Order	File Name	Order	File Name
1	F9-P00-T00-HEADER-2019.rds	21	SA-P99-T00-PUBLIC-CHARITY-STATUS-2019.rds	41	SF-P99-T00-FRGN-ORG-GRANTS-2019.rds
2	F9-P01-T00-SUMMARY-2019.rds	22	SCHEDULE-TABLE-2019.rds	42	SG-P01-T00-FUNDRAISING-ACTS-2019.rds
3	F9-P02-T00-SIGNATURE-2019.rds	23	SC-P02-T00-LOBBY-2019.rds	43	SG-P02-T00-FUNDRAISING-EVENTS-2019.rds
4	F9-P03-T00-MISSION-2019.rds	24	SC-P03-T00-LOBBY-2019.rds	44	SG-P03-T00-GAMING-2019.rds
5	F9-P03-T00-PROGRAMS-2019.rds	25	SD-P01-T00-ORGS-DONOR-ADVISED-FUNDS-OTH-2019.rds	45	SH-P01-T00-FAP-COMMUNITY-BENEFIT-POLICY-2019.rds
6	F9-P04-T00-REQUIRED-SCHEDULES-2019.rds	26	SD-P02-T00-CONSERV-EASEMENTS-2019.rds	46	SH-P02-T00-FAP-COMMUNITY-BENEFIT-POLICY-2019.rds
7	F9-P05-T00-OTHER-IRS-FILING-2019.rds	27	SD-P03-T00-ORGS-COLLECT-ART-HIST-TREASURE-OTH-2019.rds	47	SH-P03-T00-FAP-COMMUNITY-BENEFIT-POLICY-2019.rds
8	F9-P06-T00-GOVERNANCE-2019.rds	28	SD-P04-T00-ESCROW-CUSTODIAL-ARRANGEMENTS-2019.rds	48	SH-P05-T00-FAP-COMMUNITY-BENEFIT-POLICY-2019.rds
9	F9-P07-T00-DIR-TRUST-KEY-2019.rds	29	SD-P05-T00-ENDOWMENT-2019.rds	49	SH-P99-T00-FAP-COMMUNITY-BENEFIT-POLICY-2019.rds
10	F9-P08-T00-REVENUE-2019.rds	30	SD-P06-T00-LAND-BLDG-EQUIP-2019.rds	50	SI-P01-T00-GRANTS-INFO-2019.rds
11	F9-P09-T00-EXPENSES-2019.rds	31	SD-P07-T00-INVESTMENTS-OTH-SECURITIES-2019.rds	51	SI-P02-T00-GRANTS-US-ORGS-GOVTS-2019.rds
12	F9-P10-T00-BALANCE-SHEET-2019.rds	32	SD-P09-T00-OTH-ASSETS-2019.rds	52	SI-P99-T00-GRANTS-US-ORGS-GOVTS-2019.rds
13	F9-P11-T00-ASSETS-2019.rds	33	SD-P10-T00-OTH-LIABILITIES-2019.rds	53	SJ-P01-T00-COMPENSATION-2019.rds
14	F9-P12-T00-FINANCIAL-REPORTING-2019.rds	34	SD-P11-T00-RECONCILIATION-REVENUE-2019.rds	54	SL-P01-T00-EXCESS-BENEFIT-TRANSAC-2019.rds

Order	File Name	Order	File Name	Order	File Name
15	SA-P00-T00-HEADER-2019.rds	35	SD-P12-T00-RECONCILIATION-EXPENSES-2019.rds	55	SL-P02-T00-LOANS-INTERESTED-PERS-2019.rds
16	SA-P01-T00-PUBLIC-CHARITY-STATUS-2019.rds	36	SD-P99-T00-RECONCILIATION-NETASSETS-2019.rds	56	SM-P01-T00-NONCASH-CONTRIBUTIONS-2019.rds
17	SA-P02-T00-SUPPORT-SCHEDULE-170-2019.rds	37	SE-P01-T00-SCHOOLS-2019.rds	57	SN-P01-T00-LIQUIDATION-TERMINATION-DISSOLUTION-2019.rds
18	SA-P03-T00-SUPPORT-SCHEDULE-509-2019.rds	38	SF-P01-T00-FRGN-ACTS-2019.rds	58	SN-P02-T00-DISPOSITION-OF-ASSETS-2019.rds
19	SA-P04-T00-SUPPORT-ORGS-2019.rds	39	SF-P02-T00-FRGN-ORG-GRANTS-2019.rds	59	SN-P99-T00-LIQUIDATION-TERMINATION-DISSOLUTION-2019.rds
20	SA-P05-T00-SUPPORT-ORGS-2019.rds	40	SF-P04-T00-FRGN-INTERESTS-2019.rds	60	SR-P05-T00-TRANSACTIONS-RLTD-ORGS-2019.rds

Table 7. Variables generated via the lasso approach

Variable	Coefficient
OBJECTID	1.94E-18
ORG_NAME_L1	1.01E-03
F9_00_ORG_WEBSITE	3.94E-03
F9_01_REV_CONTR_TOT_PY	2.57E-09
F9_01_EXP_FUNDR_TOT_CY	2.26E-07
F9_01_EXP_OTH_PY	1.56E-09
F9_04_REP_FOOTNOTE_FIN48_X	1.42E-02
F9_06_DISCLOSURE_BOOK_ADDR_CITY	3.13E-03
F9_09_EXP_TOT_FUNDR	2.29E-10
F9_12_FINSTAT_AUDIT_X	1.14E-02
SCHEDD	2.22E-02
SCHEDJ	1.49E-01
SCHEDM	2.37E-01

Appendix 3E: Comparison Table of NTEE Codes

Table 8. NTEE & IRS Activity Codes, by 10 Major Groups

NTEE Major Group (10) NTEE Major Group (26)	**Count of IRS Activity Codes**	**NTEE Major Group (10)** NTEE Major Group (26)	**Count of IRS Activity Codes**
Arts, Culture, and Humanities	39	N Recreation, Sports, Leisure, Athletics	30
A Arts, Culture, and Humanities	39	O Youth Development	24
Education	28	P Human Services - Multipurpose and Other	45
B Education	28	International, Foreign Affairs	22
Environment and Animals	38	Q International, Foreign Affairs, and National Security	22
C Environmental Quality, Protection, and Beautification	20	Public, Societal Benefit	123
D Animal-Related	18	R Civil Rights, Social Action, Advocacy (R)	21
Health	120	S Community Improvement, Capacity Building	23
E Health - General and Rehabilitative	28	T Philanthropy, Voluntarism, and Grantmaking Foundations	18
F Mental Health, Crisis Intervention	24	U Science and Technology Research Institutes, Services	19
G Diseases, Disorders, Medical Disciplines	34	V Social Science Research Institutes, Services	23
H Medical Research	34	W Public, Society Benefit - Multipurpose and Other	19
Human Services	191	Religion Related	21
I Crime, Legal Related	25	X Religion Related, Spiritual Development	21
J Employment, Job Related	15	Mutual/Membership Benefit	22
K Food, Agriculture, and Nutrition	19	Y Mutual/Membership Benefit Organizations, Other	22
L Housing, Shelter	19	Unknown, Unclassified	1
M Public Safety, Disaster Preparedness, and Relief	14	Z Unknown	1

Data sources include Jones (2019) and Internal Revenue Service (2010).

Conclusions

Rationalization among nonprofits and NGOs can affect their ability to achieve their mission and the size of their impact on society. Theoretically, a nonprofit should pursue rationalization up to the point where they are able to best express that ability. There are numerous internal and external factors that can impede a nonprofit's ability to achieve, stay at, and/or not surpass its idiosyncratic, ideal level of rationalization relative to ability for mission achievement. Each of the three papers in this dissertation builds towards a foundation for future studies into this relationship between rationalization and mission achievement, the ultimate performance measure for nonprofits and NGOs. While it made progress towards shedding light on the guiding questions related to rationalization as process formalization among nonprofits and NGOs, each of those questions in turn feed into larger questions around the relationship between actors (nonprofits, NGOs, and others), nonprofits' processes, and the resources that are used and consumed to provide services that fulfil critical social and community roles.

Table 1. Paper Questions

	Paper Title	Guiding Question	Overarching question
1	Meta-Rationalization of Impact Evaluation	How might standard-setting documents cause NGOs to rationalize processes through formalizing impact evaluation?	What influences from outside actors can affect nonprofits and NGOs, not just broadly but their very processes and the ways in which they provide services?
2	Rationalizing the Valuation of Goods and Services	What are ways to value in-kind contributions of goods and services that can improve how nonprofits understand their outputs and outcomes?	How do these organizations use resources to serve their targeted communities?
3	Creating a Multi-Dimensional Rationalization Measure	How can we better measure rationalization within nonprofits?	How might the degree of rationalization at each organization help or hinder their services?

Paper 1. Meta-Rationalization of Impact Evaluation

Paper 1 finds that exploring the presence and expressions of meta-rationalization in impact evaluation can broaden the understanding of influences on an NGO's performance and overall societal impact. There are also negative implications for an NGO to sign on to more than one standard, as it could lead to resources being diverted away from the NGO's core missions and allocated in ways that can restrict it or even clash with the organization's values. Referring to Table , outside actors' influences can have negative effects on NGOs' processes by diminishing the effect of or reducing the number of provided services.

Ultimately, these expressions of meta-rationalization indicate at trends across the set of standards. The ways that different organizations, segments of the international aid chain, geographic perspectives, and kind of standard communicate different vs. overlapping expectations can shape what standards an NGO should pursue signing on to. A standard produced by an influential global reputation may be an obvious choice, but reviewing its expressions of meta-rationalization may help an NGO realize that the required resources would be too large of a diversion from current activities, and it would be better to go with a locally produced standard or not commit fully to the global one. This perspective allows NGOs to become better "consumers" of standards and to not center the expectations and requirements of standard-setting organizations at the expense of their activities, values, and served communities.

Paper 2. Rationalizing the Valuation of Goods and Services

The taxonomy provides an enhanced resource valuation method that practitioners and researchers alike can use to improve assessment of nonprofit performance. It is designed to give users freedom in selecting which of the eight different dimensions are most appropriate for their specific context

and purpose: by performing the current, required fair market valuation as well as one of the more subjective valuations, the user can better respond to regulatory and reporting requirements as well as gain insights into what nonprofits see as valuable or useful. Users will be most successful in selecting which dimensions to implement in addition to fair market value when they have a clear purpose, e.g., summative evaluation of an event and identifying opportunities to reduce costs for future events, formative evaluation of a successful pilot program to identify necessary resources so that the program can start elsewhere right away. Nonprofits could use the taxonomy to identify and measure the use of in-kind contributions in service delivery and programs. The tool encourages collection of more financial data than what most nonprofits collect per the literature, and at the same time lets nonprofits identify how essential in-kind contributions are to achieving their missions.

Referring to Table , this taxonomy allows nonprofit managers and researchers to understand how nonprofits use all resources, not just financial ones such as grants, fees, and monetary donations. For example, nonprofit managers could use this to inventory all resources consumed to deliver a program, service, or activity that may be delivered in the future. With financial and subjective valuations from the taxonomy, managers can then track changes in underreported assets over time and identify which are most important (then pursue relevant donations or even purchase) and which can be replaced or even dismissed. The simple act of creating a resource inventory or portfolio and filling out at least two dimensions would engage nonprofit managers with data that can provide critical insights to replicate or improve services in the future.

Paper 3. Creating a Multi-Dimensional Rationalization Measure

The presented method for creating a flexible rationalization measure allows large-scale observation of process formalization across an organization, which can be linked in future studies to effects on organizational performance. My findings, while preliminary, suggest that coupling theory-derived variables with variables predicted by algorithms can yield a high performance; this would then allow for researchers to have a deeper understanding of what, exactly, the variables used in these measures are communicating about the population of nonprofits that they are studying. How rationalized an organization is has implications for its performance, efficacy, efficiency, and the effect that it has on its intended communities. As I proposed in the introduction of the dissertation, the ultimate performance measure for nonprofits and NGOs might just be their ability to achieve their stated mission. Measuring the degree of rationalization allows us to better situate where a nonprofit is relative to that ability, and determine if it is approaching at, or past its peak rationalization relative to that performance (see Figure).

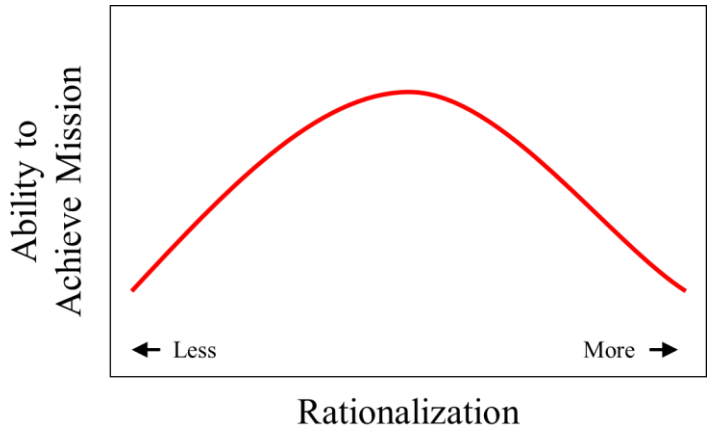

Figure 1. Proposed relationship between rationalization and mission achievement for nonprofits and NGOs. Reproduction of Figure 1 in the introduction of this dissertation.

We can gain additional insights into where an organization lies on this curve by understanding how well the organization is able to convert resources into its desired outputs and outcomes: this efficacy should theoretically be maximized at peak rationalization relative to mission achievement ability. The degree of wasted inputs and size of returns, or cost effectiveness, can be used for a similar purpose. In other words, understanding efficacy and effectiveness provide further insights into an organization's performance, i.e., degree of rationalization and ability to achieve its mission. Mission achievement can be best measured and understood in terms of the degree to which the organization achieved its desired effect on its intended communities. Given the possible relationship between rationalization and this kind of performance, the flexible rationalization score provides a key element in building the relationship between performance and effects experienced by communities, which ties back to the question for this paper that is indicated in Table .

The methodology for a flexible rationalization score will allow researchers to understand these four aspects of a nonprofit, compare to peer organizations, and better analyze the relationship between costs or resources consumed and services delivered. Decision-makers such as donors and foundations can use these findings to better direct their funds as well as identify nonprofits that can be strengthened. Finally, researchers and practitioners can challenge dominant narratives around financial "deservingness" and the reliance on certain financial measures (e.g., overhead expense ratio) by better identifying what is necessary for nonprofits in different sectors and fields to achieve impact.

Opportunities for Future Research

There are multiple opportunities for future research and testing the papers' findings and outputs in multiple contexts to determine external validity and reliability. Expressions of meta-rationalization

can be explored in contexts, industries, and sectors beyond impact evaluation and international development. Doing so will provide data on potentially unobserved secondary influences that result from primary pressures, and in turn allow organizations to make better informed decisions in response to those primary pressures to anticipate, mitigate, or enhance those secondary influences as they see fit. Presence of the identified expression of meta-rationalization in Paper 1 can be explored among the processes of signatories to the analyzed standard-setting documents. By performing interviews and collecting relevant data, researchers will shed light on the process-level effects of meta-rationalization, e.g., potential resource allocation issues and value conflicts from signing on to one or more standards. These two approaches can help establish a broader understanding of how meta-rationalization appears across contexts and within processes.

The taxonomy could be used to propose new and explore existing relationships between theories and how they view, understand, and account for resources. This could reinforce or challenge existing assumptions about resources in ways that will deepen our understanding of organizations through those theoretic perspectives. Where data are available, the taxonomy could also be implemented in other sectors. While some dimensions may not be valid outside of a nonprofit context, the principal concern around underreported assets' effect on organizational performance is also found among governments and businesses, which suggests that at least some dimensions of the taxonomy may be useful for organizations other than nonprofits. This would yield additional data and standardized valuation methods in many policy and operating spaces that lack both, as well as contribute a solution that allows more organizations to assign their own meaning and value to resources while also being able to compare such values externally.

The flexible rationalization measure can be generated in other contexts to test and identify possible limits or issues. Such contexts include kinds of nonprofits (e.g., more homogenous dataset of

nonprofits that engage in the same activities; activity codes other than health-related ones) and geographies (i.e., testing explanatory power when applied nationally, as compared to levels of region, state, county, city). By analyzing other mixes of nonprofits, researchers can increase understanding of trends in rationalization and challenges in measuring rationalization. This will help to narrow down potential peaks of rationalization relative to ability to achieve mission for different kinds of nonprofits, as well as identify data cleaning methods or issues for specific populations. The methodology for the measure can also be tested for ways to improve performance around identifying variables, i.e., implementation of algorithms more complex than lasso. This could improve the accuracy of rationalization scores as well as contribute to broader discussions around big data, policy, and nonprofits.

Finally, I and future scholars can develop best practices to facilitate adoption of the findings and methods outlined in this dissertation across nonprofits and research communities. Only by encouraging adoption and testing can these outputs achieve maximum utility and social impact.

In this dissertation, I seek to begin laying a foundation for how process formalization among nonprofits, rationalization, is related to nonprofits and how they seek to achieve their missions. My work is able to show that outside actors can have secondary influences on how NGOs rationalize, outside of whatever pressures those actors intend to have, how rationalizing valuation can enhance the data a nonprofit already has and improve its understanding of how to approach its theoretical peak of mission achievement, and how to measure rationalization at a nonprofit using whatever data is accessible. My hope is that future research can build on these findings to identify how rationalization can best help nonprofits and NGOs achieve their missions and maximize their potential impact on intended communities, and that such information can be used by nonprofits to

make the best-informed choices to maximize their impact. In doing so, I am optimistic that, to borrow from Saint-Exupéry (1943), we can take what is essential and make it visible to the eye.

References

Saint-Exupéry, A. de, & Woods, K. (1943). The little prince (K. Woods, Trans.). Reynal & Hitchcock.