Producing Gender Injustice:
Quantified Evaluation as a Logic in International Development

A Dissertation
SUBMITTED TO THE FACULTY OF THE UNIVERSITY OF MINNESOTA
BY

Emily Springer

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Rachel Schurman, Elizabeth Boyle

December 2019

Copyright Page

Acknowledgements

Although PhDs are awarded to individuals, they are most certainly a community affair. I received copious amounts of support in various forms throughout the degree for which I am deeply grateful. My deepest thanks to my co-advisors, Rachel Schurman and Elizabeth Boyle, for their steadfast support and to my committee members, Cawo Abdi and Deborah Levison—I thank you all for your intellectual mentorship and kind support through the challenges of professionalization into the discipline. In addition to this dream team of mentors, there are many others who positively impacted my trajectory through their teaching and/or mentoring, including Ron Aminzade, Karen Brown, Richa Nagar, Fran Vavrus, Joan DeJaeghere, Michael Goldman, Vinay Gidwani, Abdi Samatar, Phyllis Moen, and Carrie Oelberger. I thank you all!

The friendly faces of staff members were a lovely mainstay of my years in the department: Becky Drasin, Elizabeth Cronin, Molly Drew, Hilda Mork, Kerrie Deef, Hollie Schnoover, Nadine Kottom-Dale, Ann Miller, and at ICGC, Laura Bell and Shereen Sabet. Thank you everyone for helping me get things done!

Family and friends provided me with support of all kinds—from reading draft to the labor of everyday life: meals, laundry, rides, and more. My family was incredibly supportive during what often felt like a very selfish period of my life. Sarah, thank you so much for your brilliant feedback on the many drafts that dotted the path toward this degree. My heartfelt appreciation for the kindness of those who understand this journey from the inside: Erin, Anthony, Devika, Sravanthi, Annie, Suzy, Emily, and Alex. And to the real smarties who found a life outside of academia: Venkat, Hanife, Deb, Becky, Feteh, Mike, Eske, Dasha, Danny, Trevor, Sadig, and Myron. The lake walks and brunches breathed new life into me.

And my heartfelt thanks for the time and generosity of the many development professionals who let me into their offices and shared their worlds with me.

I am grateful for the financial support that made this research possible, most notably the Social Science Research Council's Mellon International Dissertation Research Fellowship and Dissertation Proposal Development Fellowship, the Schwartzberg Award for Global Governance through the Interdisciplinary Center for Global Change, and the Yudof Fellowship for Science Policy and Ethics at University of Minnesota.

Any mistakes in this dissertation are fully my own. Please note that the dissertation serves as a timestamp of my thinking about the intersections of gender, measurement, and development. I'm eager for the journey that lies ahead as I deepen these arguments.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

Large donor agencies that fund international development projects create measurement systems to track how their funds are utilized in faraway places, claiming these systems provide data about "what works" for global poverty reduction (DFID 2013; European Commission 2014; GIZ 2017; United Nations Evaluation Group 2016; USAID 2011; World Bank Independent Evaluation Group 2013). These cross-national systems of measurement are termed "monitoring and evaluation systems" or, more colloquially, "M&E." In 2017, I sat at a table amongst 150 international development professionals gathered for a conference called "MERL Tech"—a one-day event where practitioners and technologists come together to showcase new technologies for Monitoring, Evaluation, Research, and Learning in the international development sector. I, like others, listened intently to the speaker on stage—an energetic evaluation director of a project in South Sudan. Partway through her speech she exclaimed: "We know we're lying, and the donors know we're lying. It's called Monitoring and Evaluation!" The crowd laughed, seeming to enjoy the emotional release of this truth telling. She continued, "What data do we *really* have? We're just counting… Look at the spreadsheet! It's full!" More laughter from the crowd. At the end of her speech, she received hearty applause.

This statement, the reaction it garnered, and the setting in which it was said encapsulate the main focus of this dissertation: the contradictions of measurement as experienced by professionals and organizations in the development sector today. In contrast to measurement offering "what works," this evaluation director brazenly called M&E "lies" and cast both evaluators and donors as actors involved in the collusion of such inaccuracies. In doing so, the evaluation director highlighted a lack of trust in M&E data. But this

statement tells us nothing about *how* and *why* these actors would participate in such a thing. Further she drew attention to a spreadsheet—an evaluation document—noting it may be full but calling its worth into question. The laughter, instead of outrage, at this statement demonstrates some level of shared recognition. And lastly, uttered at a conference dedicated to measurement in the development sector, demonstrates how speaking out such a statement can easily translate into applied problem-solving—that the development sector needs a *technical* solution of better tools, technology, and/or data quality—rather than an interrogation of the motivations of development actors who engage in "lies" disguised as objective data.

The development sector has strongly adopted and normalized the use of project-level "M&E" or, put simply, performance metrics (Eyben et al 2015; Hirschmann 2002; Natsios 2010). This has resulted in the proliferation of unintended consequences, such as measurement activities disproportionately garnering staff time and attention, the bureaucratization of community organizations and marginalization of social transformation processes, and the creation of a pressure-cooker-like working environment, among others (Bromley and Powell 2012; Merry and Wood 2015; Sangtin Writers and Nagar 2006; Sharma 2008). This dissertation turns the analytic eye away from perfecting development measurement and toward what measurement does in its presence, addressing a dearth in development scholarship about what metrics *do* in their social worlds (Mosse 2013; Rottenburg et al. 2015; Viterna and Robertson 2015) by analyzing metrics from the "shop floor."

Using the case of gender-related policies, indicators, and programming, I studied how development professionals and organizations react to the pressure to quantify and

deliver results. Gender-related development efforts were selected as a case study because women's empowerment is understood as hard-to-measure and socially embedded (Kabeer 1999; 2003; Sangtin Writers and Nagar 2006; Sharma 2008). The dissertation is empirically based on an East African instance of a 20-country agricultural initiative by a large bilateral donor, said to result in women's empowerment and measured by a standardized evaluation system. For this multi-sited research, I traveled from a recipient country to the donor country, tracing a donor-mandated evaluation system from paper to spreadsheets to an aggregating website to report consumption. I completed 60 interviews with professionals across 4 large implementing organizations, 3 evaluation organizations, and multiple tiers of donor offices. Further, I completed participant observation at international conferences and meetings, and analyzed evaluation-related documents, such as project performance plans, tracking spreadsheets, quarterly reports, and external evaluations.

I theorize that measurement practices are better understood as the primary organizing structure, or logic, of all development labor. In other words, quantified evaluation systems are not merely benign measurements, but structure the relationships within and between organizations, professionals, and "beneficiaries." Evaluation systems actually demarcate what is considered legitimate action in the development sector: constraining professional agency toward localized development while emboldening those who readily participate in "lying" with M&E data, which is disconnected from local project realities.

Regarding the intersection of measurement with gender-related development efforts, my findings demonstrate mandatory measurement as the driving *mechanism* for

organizations to "get women in the door" while simultaneously rationalizing women's marginalization within project activities. The marginalization of gender in the evaluation system is shared across the multiple tiers of development organizations, eliciting bureaucratic micro-battles as professionals (mostly women), push, negotiate, and/or fight for greater inclusion. With professional and organizational status and reputation tied to demonstrable results able to be aggregated across global portfolios of development initiatives, the space for different forms of measuring women's empowerment is foreclosed. This results in a bureaucratic feminism that can only ever reflect that of the donor's interests—recreating second wave "global sisterhood" feminism vis-à-vis metrics despite its failure to account for the diverse experiences of women around the world experiencing different manifestations of structural inequalities.

During the 1980s, as financial accounting principles were applied to new sectors, a societal transformation toward "audit cultures" began (Powers 1997; Strathern 2000). This transformation signaled a qualitative shift in how numbers were utilized, departing from the use of statistics by states to govern populations to an emphasis of performance measurement at the level organizations and individuals (Rottenburg et al. 2015). As calls for evidence of performance rose, they became intertwined with transparency and accountability as moral imperatives. In doing so, performance measurements became mechanisms not of trust, but of *reputation* (Espeland and Sauder 2016; Shore and Wright 2015). Development organizations originally adopted quantitative measurement as a mode of forming trust between development actors to solve the principal-agent dilemma (Best 2017; Ebrahim 2003; Watkins, Swidler, and Hannan 2012): donors are constantly concerned lower-level implementing organizations are not completing the expected work

or are controlling valued "local" information. But as societal expectations of demonstrated performance increased, development organizations came under pressure to provide evidence of their impact on the lives of the global poor. This shift in global expectations towards quantified evaluation enabled a social process known as reactivity to take hold within the development sector.

Reactivity is the idea that, when under measurement, professionals and organizations change their behavior to best accomplish the metrics (Espeland and Stevens 1998; Espeland and Sauder 2007). Once reactivity as a social process is set into motion, metrics stop serving as proxy measures for an organizational vision or strategy and start to shape actions—drawing the attention and energy of those who are measured by them (Harris and Tayler 2019; Kaplan and Norton 1992). In other words, once reactivity occurs, metrics stop measuring the social world and instead induce professionals and organizations to enact a new social world around them. Reactivity also presents new managerial possibilities (Miller 2001): if people and organizations who know they are under measurement react by orienting their behavior toward the accomplishment of the indicator, then managers can harness this process to aid their own goals. Attempts to harness these reactions can be understood as a new form of 21st century governance (Best 2017; Davis et al. 2012; Kelley 2019). The manifestation of managerial governance can be seen in the rise of workplace surveillance through metrics (Ajunwa, Crawford, and Schultz 2017; Levy 2016; Van Oort 2018). The concept of "refractive surveillance" (Levy and Barocas 2018) by metrics is instructive: data collected about one group of people may be used to elicit reactions from an entirely different group. In the development sector, professionals and organizations are not directly measured—the execution of the project is measured in terms

of measures like crop yields, new technologies adopted by farmers, and number of people trained, among others. Professionals and organizations change their behavior to accomplish these simple metrics because these metrics are understood as evidence of project delivery. By delivering contracts to donor specifications, professionals and organizations maintain their reputation and increase their likelihood to be hired or awarded a future project contract.

Research has documented that reactivity to measurement occurs with respect to international indices (Best 2017; Davis et al. 2012; Easterly 2009; Høyland, Moene, and Willumsen 2012; Kelley 2019) and within high level forums of the United Nations (Fukuda-Parr Greenstein Stewart 2013; Merry 2016). Others still have raised concerns about the impact of performance metrics on localized processes that cannot be easily quantified, such as capacity building (Vallejo and Wehn 2016), democratization (Hirshmann 2002), and gender equality (Liebowitz and Zwingel 2014). Although scholarship acknowledges the unintended consequences of the pressure to quantify complex outcomes, what occurs inside *development organizations* around these hard-to-measure social processes, like women's empowerment, has remained a black box. I use sociology of organizations scholarship—namely reactivity (Espeland and Stevens 1998; Espeland and Sauder 2007), gendered organizations (Acker 1990; 2002; 2006), and institutional logics (Friedland and Alford 1991; Thornton et al. 2012)—to analyze what occurs and why inside development organizations.

This research intervenes in sociological scholarship in three core ways. First, joining the growing number of quantification scholars, I shift the analytical eye away from how to perfect evaluation measures to a sociological analysis of what metrics *do* in

development contexts. The negative repercussions of measurement are well documented in the US and Europe (Colyvas 2012; Espeland and Sauder 2007; Shore and Wright 2000), as well as internationally in the form of nation-state ranking (Rottenburg et al 2015; Kelley 2019). I demonstrate that this phenomenon operates *transnationally* within interlinked bureaucracies, cultivating professionals and organizations who work together to collect and aggregate data, write reports, ultimately successfully co-producing evaluation artefacts across vast geographies.

Second, my research identifies a new, important organizational actor in development: in addition to the common "buy it" (the donor) and "do it" (the implementer) organizations, my work documents "check it" organizations (evaluation-specific contracts). "Check it" organizations are tasked with helping "do it" organizations handle the rising demands for knowledge management, learning, compliance, and evaluation as stipulated by "buy it" organizations, demonstrating an organizational reconfiguration. This work invigorates sociology of development with a relational account of measurement within and between organizations, documenting how professional roles interact to successfully co-produce development reports. The durability of performance metrics as an institutional logic becomes legible as analysis moves across these diverse sites.

Third, drawing from anthropologists who have adeptly studied quantification cultures (Merry 2006; 2011; 2016; Strathern 2000; Shore and Wright 2001; 2015), my work offers a sociological account of how performance metrics narrow agency down particular pathways. A few development sociologists have focused on macro-level indices, such as the Gender Inequality Index (Bose 2015), but meso- and micro-measurement in large-scale projects—the predominant form of development work—is key to

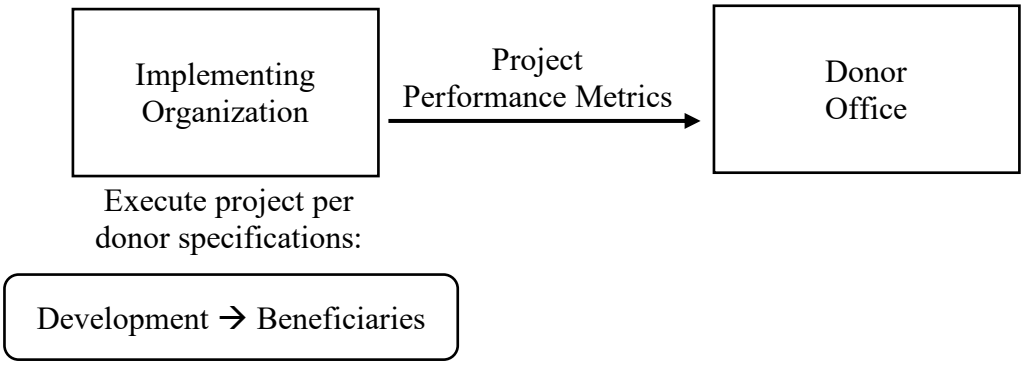
understanding the role of numbers in everyday development. I shift the level of analysis inside development organizations and demonstrate that reactivity may be elicited even by simple counts. This builds the scope condition of extant research by demonstrating the *power of nested measurements*. Indicators are defined as quantified measures that have four traits: (1) name the social world, (2) rank order items, (3) simplify, and (4) are tools for evaluation (Davis et al 2012; Rottenburg et al 2015; Espeland and Sauder 2007). Accordingly, the simple output numbers that make up the majority of development measurement would not, by this definition, be indicators or indicators capable of eliciting reactivity. My research demonstrates that simple counts take on reactive powers because they are nested within greater structures of comparison that allocate significant material resources.

This dissertation begins from the premise that development organizations are not moral, value-driven organizations (DiMaggio and Anheier 1990; Hwang and Bromley 2015; Watkins et al. 2012) but are organizations with imperatives for survival (Brass 2012; Cooley and Ron 2002; Krause 2014; Nagaraj 2015; Roberts 2014). These imperatives mean that development organizations form an *industry*. Dominant understandings of development organizations, or relief organizations in Krause's (2014) case, are that these organizations deliver "development" or "relief" to beneficiaries engaged as participants in development projects and that donors fund these developmental activities. However, Krause argues that this relationship occurs nested within a higher-level relationship: that of a vendor and a contractor. In other words, donors *buy a product*: what Krause terms a "good project." Krause outlined how implementing organizations, or vendors, "produce good projects" that appeal to donors and thus win contracts. Although this establishes

donors as buyers and implementing organizations as sellers in a development marketplace, it leaves what happens during the implementation of development projects underexplored.

I begin where Krause (2014) left off by highlighting transnational performance metrics as the main *mode of communication* between implementing organizations and donors—these metrics "evidence" project contract completion and provide donors with the needed bureaucratic documentation for their own evaluation landscapes (Brandtner 2017). The client-vendor relationship (e.g. donor-implementing organization) generates a principal-agent problem: the principal hires an agent to execute their desired action. In development this results in a simple question: how can donors believe their funds are being spent as they wish? Enter monitoring and evaluation systems. See Figure 1-1 for a representation of the role of project performance metrics as the line of communication between implementing organizations and donors. Notice that donors have no capacity to "see" what occurs in the project—the only insight they have to what occurs at the site of the development project is the information shared by the implementing organization. Donors have to choose to trust that the implementing organizations accurately share information about the project.

Figure 1-1 Performance Metrics as Mode of Communication between Development Client and Vendor



Implementing organizations are better understood, then, as vendors or contractors, serving as proxies who are hired by donors to "deliver development"—this forms a long aid chain that spans continents and organizations (Bebbington 2005; Wallace, Bornstein, and Chapman 2006). Driven by distrust of the vendor and an awareness that their knowledge of the project is inferior to that of the agent, principals' desire to "see" their investment unfold as planned is heightened. Thus, principals create measurements to track and ensure delivery of their ideas in faraway places (Best 2017; Ebrahim 2003, 2005; van Zyl, Claeyé, and Flambard 2018) and impose these as conditions for receiving funding. Quantification is a superior managerial strategy in the development context because it serves as a "technology of distance" by minimizing the need for personal trust while obscuring the people, assumptions, and power dynamics involved in their production (Espeland and Stevens, 2008; Merry, 2011; Mowles, 2010; Porter, 1995; Powers, 1997). For these reasons, quantified evaluation systems have become a normalized component of the development sector.

Although repositioning implementing organizations as vendors or contractors producing a product of value to the donor helps us understand development organizations as *organizations*, it still leaves the mechanisms by which these relationships are managed and maintained undertheorized. In response, my research centers the measurement systems that form the lines of transnational communication between implementing organizations and donors. These systems are meant to produce trust of project execution, but as audit cultures have taken hold, these systems have iteratively become mechanisms of professional and organizational *reputation*—initiating reactivity. Organizations oriented

toward one another in a field of competition (Barman 2016; Krause 2014) advance their reputations by demonstrating their ability to perform *as measured by M&E systems*. Project performance metrics are pieces of greater inter-organizational reputation, interlinking tiers of donor offices and those of implementing organizations. Development organizations, as stated above, play distinct roles in development: "buy it," "do it," and "check it." Furthermore, these offices are located in a variety of locations around the world, including the capital cities of donor countries, capital cities of development recipient countries, and semi-rural areas in recipient countries, and, lastly, in rural areas abroad. While some development projects take place in urban areas, the agricultural initiative studied here executed project activities in rural areas. See Figure 1-2 for a representation of transnational evaluation flows across development organizations, including those who "buy it," "do it," and "check it."

All three types of development organizations are interconnected through a series of policies, procedures, and contracts. Donor policies and procedures outline standard operating procedures and mandate compliance throughout the tiers of the aid chain, pushing down from donor headquarters to donor field offices. Contracts serve as legal documents that outline the desired development project to be implemented on behalf of the donor. Contracts detail and mandate data collection on donor-created performance metrics, including the contracted "amount" of each metric to be delivered by the contractor to the donor. As contracts are nested within the donor policies and procedures, they serve as legal instruments that mandate the compliance of implementing organizations with the policies and procedures of the donor.

Project contracts legally obligate implementing organizations to measure the

project as stipulated by the donor. Donors design measurement systems to track multi-country initiatives so they can report results to legislative bodies that allocate funds, such as Congress or Parliament. These measurement systems include definitions of each indicator, desired levels of disaggregation, and desired directionality (i.e. poverty indicators should decrease, literacy indicators should increase). These measurement systems are outlined in the policies and procedures that push downward through the organizational tiers. At the site of the project, implementing organizations begin counting and measuring projects, monitoring their progress against the donor-mandated indicators. Data collected from individual project sites are then aggregated with data from other projects under the jurisdiction of the same field office. The country office then aggregates data from all of the field offices and shares this information with the donor field office as evidence of progress toward project completion. At the same time, the contractor field office shares this data with their own organization's headquarters for internal managerial purposes, however, this data can now be claimed as evidence of their organizational ability to deliver projects to donor specifications.

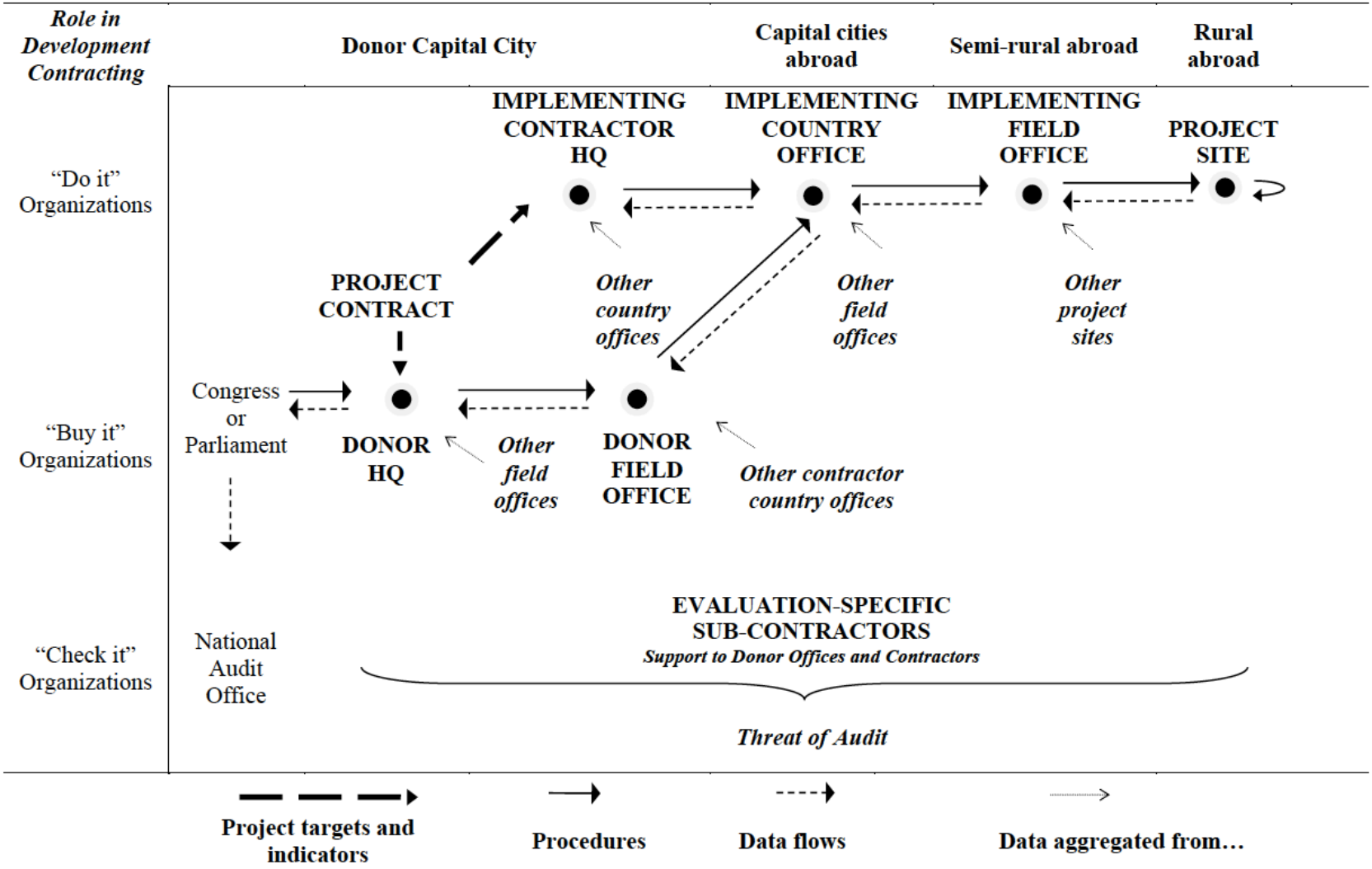
The donor field office manages multiple contractors in the same initiative and aggregates the collected indicator data across the contractors. At this level, the aggregated data no longer is indicative of a contractor's ability to deliver a project to the donor but is now indicative of the donor staff's ability to *manage* projects to successful execution (as measured by the delivery of contracted results). Donor field offices pass the data upward to donor headquarters. At this level, the aggregated data is indicative of a donor field office's ability to achieve their field office priorities. The donor headquarter office then uses the data aggregated from donor field offices as evidence of taxpayer dollar investment

in international development to legislative bodies. The international development office of donor governments is in competition with other departments for taxpayer dollars and thus must maintain a positive reputation with elected officials.

As data flows upward through these organizational tiers, the threat of being audited is present in a variety of forms. In brick and mortar form, the "check it" organizations (under contracts themselves) exist in capital cities in recipient countries and employ professionals to assist the "do it" organizations to manage the knowledge demands of the donor. The donor field office also performs miniature audits of individual indicators at various time through the life of a project, tracing through the data collection forms to ensure the accuracy of the data. At times, the donor government completes audits of entire projects where auditors from the donor government review all project-related documents. And lastly, donor field offices also come under audit from the national auditing office. Though "audit cultures" within the development sector are most often manifested through point-in-time audits such as those described above, the threat of these audits is omnipresent as these organizations and professionals go about their everyday work life.

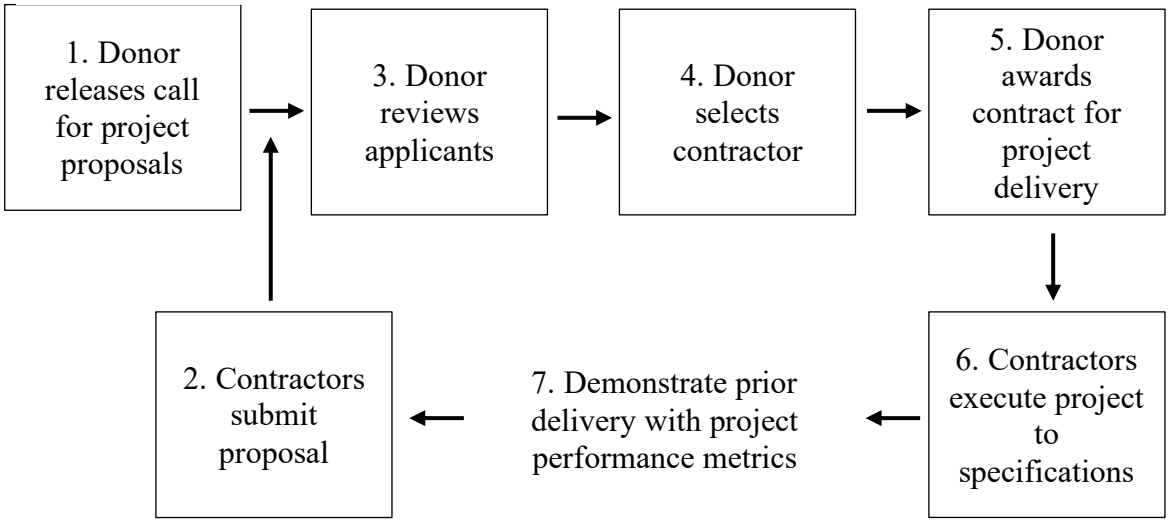
Within this transnational chain of development organizations, data, reputation, and the threat of audit become intertwined. Failure to deliver results as stipulated in a development project contract becomes not only an implementation failure of the implementing organizations, but a managerial failure of the donor offices. Therefore, all professionals have incentives to ensure that data demonstrate successful implementation of development around the globe. Indeed, measurement systems are no longer simple lines of communication to build trust between a donor and an implementer (or a principal and an agent) but have become mechanisms of reputation under the threat of audit.

Figure 1-2 The Transnational Evaluation Chain: Reporting Procedures and Data Flows between Donor and Sub-Contractors



The importance of these metrics for reputation is best seen by taking this systemic view of the development contracting process, the role of metrics, and the audit threat that all development organizations endure. As measurement policies and procedures push downward through organizations, data to be used as evidence of project implementation moves upward. This data is ultimately used as evidence of "do it" organizations' ability to deliver results to their client—the donor. This data plays an important role in the *reproduction* of these organizational relationships as seen in Figure 1-3. Donors outline desired development projects and then release requests for proposals (1). Development contractors, or implementing organizations, develop proposals that outline how and why their organization is best suited to successfully implement the outlined project (2). Donor staff review the proposal and award a multi-million-dollar contract to an implementing organization (3-5). During implementation, contractors track their progress toward the mandated indicators (6), which builds their organization's pool of evidentiary data about their delivery of successful projects (7). Note here that "success" is defined as the completion of a project *bureaucratically*. In other words, even projects that may be interpreted as failures can be recast as bureaucratically successful using simple numerical indicator data as evidence.

Figure 1-3 Role of Metrics in the Contractor Selection Process



Lastly, the goal of research on reactivity is not to outline ways for development bureaucracies to better deliver on hard-to-measure social processes, like gender justice. Instead, this research demonstrates how a focus on quantitative indicators, attached to moral calls for performance, accountability, and transparency, upend *local visions* of women's empowerment. Indicators are created by *people* and the social location of the creator is inseparable from the indicator itself since values are codified into the definition and driven by the motivations of the creator (Rottenburg et al 2015). Therefore, when quantified metrics are enacted as measurement tools, irrespective of location, they define success and create false equivalences between unlike items (Kabeer 1999; 2005; 2015; Bose 2012; 2015). In doing so, metrics embody values and worldviews about the directionality of progress and what counts as progress. Quantified metrics, useful in their ability to aggregate data for policymakers, leave little room for local versions of gender justice. As Abu-Lughod states (2002: 787-88): "We may want justice for women, but can

we accept that there might be different ideas about justice and that different women might want, or choose, different futures from what we envision as best?" The moral problem of quantification is not that we must figure out how to "do it better" but that quantification, performance, and reputation in tandem close the epistemological and ontological space for the *self-definition* of gender justice by communities around the globe.

Case & Methods

This research is empirically based on reactions of professionals and organizations under measurement by a large bilateral donor, here called the International Development Assistance Department (IDAD). The evaluation system, comprised of 52 indicators, is standardized across all 20 countries involved in an agricultural development initiative by the donor. I offer an in-depth case study of the reactions of professionals and organizations in a single East African country and place their experiences in relation to the professionals and organizations located in the donor country and serving organizational headquarters. The agricultural development initiative was selected for its discursive commitment to women's empowerment through agricultural development. The agricultural development portfolio studied in East Africa was valued at $260 million over five years and the 20-country initiative garnered $4.7 billion in total investment.

This multi-sited qualitative research is based upon three data sources. The first and primary data source is that of interviews with 60 development professionals. The second is participant observation fieldnotes taken at two relevant international conferences in the donor country and four multi-stakeholder Gender Meetings in the recipient country. The

third data source is document analysis of initiative and project-related documents, such as evaluation and gender policies and plans, quarterly and annual reports, audit reports, tracking spreadsheets, and more.

Upon hearing that my research is about an "East African case" many seem to envision that I studied development project *beneficiaries*, yet I interviewed donor representatives, consultants, project leaders, evaluation specialists, gender advisors, and researchers. Of the 60 development professionals I interviewed, 40 of them are from the Global North and only 20 are from the recipient country. See Table 1.1 for a description of the organizational sites captured, professional interviews, and the documents collected at different sites.

Table 1-1 Multi-sited Interview Frame: Organizational Sites, Occupations of Interviewees, and Evaluation Artefacts

Organizational Sites	**Sample Interviewee Titles (n=number of individuals)**	**Evaluation Artefacts**
Large Project Field Offices (n=4)	Project Leader, Monitoring and Evaluation Specialist, Gender Advisor (n=13)	Monthly & quarterly reports Annual Reports Standardized Data Tools Performance Monitoring Plan Project Gender Analysis AGMIS (input)
Evaluation Organization Field Offices (n=5)	Project Leader, Deputy Project Leader, Senior Researcher, Research Fellow, External Evaluator, Field Manager, Research Consultant (n=15)	Impact Evaluations External Evaluations (review all project documentation)
Field Landscape	Chief of Party, Country Director, Agriculture & Gender Researcher, Gender Advisor, Policy Consultant, Evaluation Lead (n=10)	--
Donor Field Office (n=1)	Agriculture Officer, Project Officer, Responsible Representative, Evaluation Advisor, Gender Specialist (n=7)	Contracts & Agreements Indicator Justifications Indicators Handbook Country Strategy Gender Strategy & Plan AGMIS (review)
Large Project HQs (n=2)	Director, Program Performance & Quality, Senior Evaluation Director, Director, Gender & Social Inclusion (n=4)	New Business Proposals Performance Monitoring Plan Definition of Indicators
HQ Landscape	Consultant, Senior Advisor & Thought Leadership, Evaluation Director, Learning and Knowledge Management Consultant (n=6)	--
Donor HQ (n=1)	Director, Office of Learning, Evaluation, and Research; Gender Advisor, Africa Department (n=4)	Request for Proposal (RFP) Compliance Rules & Regulations AGMIS (review & publicize)

Notes: HQ=Headquarters; AGMIS=Agricultural Monitoring Information System
"Landscape" includes expert interviews with professionals serving smaller Farming Sustenance projects or closely related agricultural development projects.

Rather than interviewing every staff member in an office, I sought to interview people by their occupational role—determined by their proximity and role in measurement

tasks and managerial positions that utilize the data generated by project measurement systems. Adopting a multi-sited ethnographic approach (Burawoy et al., 2000; Marcus, 1995), I "followed the thing" of performance metrics as they compel people to collect and aggregate data, produce charts and reports, and alter their work lives and identities. By researching across the sites shown previously in Figure 1-2, moving from raw data to final product, I generated thick description of the behavior of development workers, the social messiness at each site that performance metrics do not record, and the labor of their production. In doing so, I "studied through," viewing the evaluation community as a contested political space in which people who do not share a moral universe work together (Shore and Wright, 1997). Methodologically the research design was informed by Smith's (2005) institutional ethnography, which focuses on the possibilities for the expression of agency based on social location while attending to how texts constrain professional actions.

Interview guides were structured to understand literal happenings in daily work life or special events, corroborate information in reports, and uncover what people think and know about their social world as it pertained to measurement (Rubin and Rubin 2005; Young, 2009). During interviews, I established empirical saturation (Small, 2009) by the occupational role of the individual, identifying people who held experiential knowledge about metrics and organizational imperatives (Weiss, 1994). I did not undertake triangulation for validation between occupational roles but considered dissonance between occupational roles and data sources as an avenue for understanding how some knowledge may be legitimated at the expense of others (Hesse-Biber, 2012). During participant observation, I captured the meanings people attached to concepts and words as well as the specific contexts in which they invoke them (Emerson, Fretz and Shaw, 1995; O'Reilly,

2012). Text analysis of the literal evaluation system—the documents, spreadsheets, and online portals—followed institutional ethnographic principles by considering the social world around the artefact, interrogating the resources and understandings that depend upon or were captured in such artefacts (Campbell and Gregor, 2004). Additional details on analytical strategies are discussed in the empirical chapters.

This research was enabled by my positionality. Trained as a development practitioner and evaluator at Harvard's Graduate School of Education, I was "stamped" as a legitimate member of a traveling cadre of development elite—highly trained and specialized, and taught to embody and own the identity of someone who *manages* diverse peoples in diverse spaces. Familiarity with this milieu not only built my confidence to cold call donor officials and project leaders but also, upon meeting for an interview, to appear as a member that *belonged*. In other words, I immediately presented as someone on the inside, someone to share openly with, and someone who could understand the acronym-laden speech of the development sector. At the same time, not knowing the particular initiative, I was able to ask participants to back up, slow down, explain, and/or repeat themselves when I did not understand. This identity as an insider-outsider (Smith 1999) aided my interviews and my ability to manage research interactions.

Taking an uncritical stance, one might understand my access to the donor and development workplaces as one of chance and serendipity. Instead, I understand my access to these places and people as the consistent unfolding of intersectional privilege. This research was greatly aided by two meetings on the same day several years prior. I was spontaneously and graciously received by two powerful individuals who then quickly chose to share this power, connecting me to additional professionals, some inside the donor

office. Their willingness to connect me to other individuals was undoubtedly due, at least in some part, to my identity as a white, middle class, highly educated American woman. Although I may have garnered access through different means if I had needed to, these "chance events" resulted in approximately 10 interviews with donor staff—an office difficult to gain access. In summary, I studied my own kind—primarily people who look like me and/or were trained like me and/or occupy a social class in their respective countries that enables them to manage complex social processes, implement programs, or develop policies that all are meant to aid in societal development of some kind.

The Chapters

In three empirical chapters, which I develop as standalone article manuscripts, I address the role and consequences of quantified evaluation systems in international development. I find pernicious effects set in motion by high stakes measurement practices and explore the presence of such effects at different levels of the development aid chain. First, I develop a sociological analysis of how evaluation systems structure and constrain the agency of development professionals, as well as the broader managerial relations they call into being. This chapter establishes that reactivity to performance metrics occurs in the development sector—meaning that metrics are less likely to measure the empirical world, but instead, alter the actions of individuals and organizations. Second, I utilize gendered organizations as an analytical frame to understand professional responses to the growing focus on performance metrics in the development workplace. This chapter demonstrates that gender advisors' potential resistance to quantitative metrics is undermined by the

gendered organization in which they work. Instead, gender advisors become promoters of simple counts despite their disinterest in the knowledge this data offers about project participants. Third, understanding quantified evaluation as an institutional logic, I explore the marginalization of gender-related efforts and how these fare within the bureaucracies meant to deliver development programming to the global poor. This chapter documents how both evaluative structures and practices work in tandem to sideline the inclusion of gender in development programming within bureaucracies.

More specifically, in the second chapter, I provide a structural analysis, demonstrating that monitoring and evaluation systems form the status quo of everyday development work due to the managerial relationships that demarcate professional labor and organizational decision-making. Knowing they are *under* measurement, implementing organizations focus their staff resources on activities that match donor-mandated indicators, themselves designed for policymakers rather than the local context. I demonstrate the status quo is reproduced through the *negative sanctioning* of any individual who transgresses—sanctions made possible by the standardized evaluation system which enables comparisons.

In the third chapter, I provide an organizational analysis, documenting how professionals react to measurement within the organizational level. I demonstrate that although gender advisors—the professionals tasked with gender mainstreaming in projects—strongly prefer results about "what works" for development in the form of thick description about structural gender inequalities, they instead advocate for quantitative metrics. I argue this betrayal is strategic: utilizing a quantitative evaluation logic in the form of "bureaucratic tools" helps overcome the gendered nature of the development workplace

and forces reluctant coworkers to comply with gender mainstreaming. This work demonstrates that professional reactions to performance metrics in the workplace are gendered.

In the fourth chapter, I provide a cross-sectional analysis of development bureaucracies interlinked by the evaluation system, documenting what happens to gender efforts in the presence of the logic of quantified evaluation. I demonstrate that the structures and practices associated with quantified evaluation consistently marginalize gender efforts, resulting in a circumstance where gender efforts are individualized rather than institutionalized. Understanding quantified evaluation as a logic demonstrates that professionals and organizations have difficulty *making sense* of the non-quantifiable in the context of development bureaucracy. This shifts the locus of the problem away from empirical evaluation systems and toward the abstract idea of quantified acts of evaluation.

I conclude the dissertation by, first, detailing the implications of this work more broadly for development studies and sociology of organizations and, secondly, with a discussion on the future of evaluation systems in a digital world. In particular, I outline how the original problem of trust between donor and implementing organization (the principle-agent dilemma) is now likely to shift into more digital forms. I outline the importance of blockchain technologies as a "trust machine" and detail how gender is being utilized to garner interest in the adoption of such technologies in the development sector. What new possibilities for gender justice are enabled by this technological innovation, and how might it reconfigure gender relations in development offices and projects?

Chapter 2

Caught between Winning "Rebiz" and Efficacious Learning: Reactivity to Performance Metrics in Development

Abstract: What if our measurement practices undermine our ability to deliver "what works" to the world's poorest? Recent sociological scholarship posits that performance metrics alter professional and organizational behavior, meaning indicators may *recreate* rather than empirically measure social worlds. This perverse reaction to measurement is at odds with the stated goals of international development monitoring and evaluation practices, which are said to promote learning "what works" to reduce global poverty. At the level of development projects, measurement practices abound: During the project, monitoring indicators track progress toward pre-identified targets, while point-in-time evaluations are interpreted as evidence of larger change. This knowledge production aims to hone financial stewardship by routing dollars toward the most evidence-based, high-impact projects. Using the monitoring and evaluation (M&E) system of a large bilateral agricultural development initiative, I examine whether perverse reactions occur in the development sector and how they impact the process of learning "what works." While the M&E system is standardized across the 20 recipient countries, I utilize the experience of a single East African country instance, conducting interviews with approximately sixty development professionals. I demonstrate that the perverse effects of measurement are first initiated by the managerial structure of a single project, forming the status quo of everyday development work. I then demonstrate that the status quo is reinforced by comparisons made possible by standardized M&E systems: during funding decisions, M&E data enables

comparison between different options. These acts of comparison serve to quiet professionals and organizations who attempt to introduce more empirical forms of learning "what works." I argue that high stakes measurement practices create evaluation systems that, instead of providing meaningful learning about "what works" for efficacious programming, produce evidence to garner repeat business, or rather "what works for rebiz." Academics will note the increasing importance of reactivity theory to explain perverse measurement effects in the development sector, while practitioners may heed lessons about the unintended consequences of metrics and the ways they undermine poverty reduction.

INTRODUCTION

> "Monitoring and evaluation (M&E) play key roles in identifying and promoting development work that has the most positive outcomes and biggest impact on the lives of people in developing countries, for the resources available." –Dr. Vinod Thomas, former director general of the World Bank Independent Evaluation Group (Thomas 2010)

> "Most evaluations done in the development sector, you will rarely, if at all, see any that say, 'We failed.'... we are reading many reports about successes, but in reality, it is just failure dressed in a clothing of success." – Abel, field-based evaluation director

The first of these quotes implies that the data produced through development organizations' monitoring and evaluation systems will help donors learn which projects result in efficacious development benefiting the world's poor. The second points out that efficacious development can be reported by evaluations, even when it may not have occurred. This contrast should give us pause.

The world's largest bilateral and multilateral donor organizations,[1] including DFID, USAID, GIZ, EU, UN, and World Bank, cite measurement practices as key to learning "what works" and targeting funding so that results per dollar are maximized (DFID 2013; European Commission 2014; GIZ 2017; United Nations Evaluation Group 2016; USAID

[1] In 2016 dollars, the top three bilateral donors included: Department for International Development in the United Kingdom (DFID), United States Agency for International Development in the USA (USAID), German Corporation for International Cooperation in Germany (GIZ). Whereas the top three multilateral donors included: European Union (EU), United Nations (UN), and the World Bank.

2011; World Bank Independent Evaluation Group 2013). For example, GIZ states: "'Knowing what works': this is the yardstick we use to identify the results our international cooperation measures achieve. Which activities are successful? Which are not?" (GIZ 2017:4), and the European Commission asserts that evaluation output helps "understand not only what works and what does not, but critically why, and under what circumstances" (European Commission 2014: 3). Yet scholars posit that evaluation "creates success" in the development sector (Watkins, Swidler, and Hannan 2012), delivering positive evaluations which may not correspond to empirical evidence in an attempt to garner funding. In this view, monitoring and evaluation (M&E) produces political, socially constructed documents that, when failure has occurred, shift blame elsewhere to ensure organizational reputation is maintained (Venugopal 2018). Yet we know little about how measurement practices influence learning "what works and what does not"—one of the stated goals of monitoring and evaluation. If evaluation systems document "what works," we should expect occasional poor results or reports of failure to be treated as learning opportunities. Conversely, if evaluation systems "create success," then poor results or failure will be quelled and learning from implementation and/or mistakes impermissible.

International development monitoring and evaluation systems are big business, valued at $120 billion USD in 2009 (Ika and Lytvynov 2011). These systems measure the "success" of development projects through a tiered process of downward-moving policies and contractual obligations from donors to grantees and upward-moving data from the project site to donors, aggregated to demonstrate return-on-investment or value-for-money. Although there are different forms of evaluation—from monitoring project performance to external project evaluations to the recent gold standard of randomized controlled trials—

monitoring output indicators are the most widely-practiced form of measurement. and most projects include a final evaluation. In short, development professionals, projects, and organizations are increasingly under measurement. It is more common in economics and social sciences to view reactions to observation and measurement as a methodological issue to be reduced through research design. Instead, I follow Espeland and Sauder's (2007) notion of reactivity as a concept: reactivity draws the analytical eye toward how people and organizations change their actions once they know they are being measured, evaluated, or observed (Espeland and Sauder 2007). Research in the U.S. and Europe has empirically demonstrated that efforts to quantitatively measure and evaluate performance alter people's identities and behavior (Espeland and Stevens 1998) and redefine administrative priorities (Colyvas 2012; Espeland and Sauder 2007; Hirschmann 2002b; Shore and Wright 1997). This process is known as *reactivity*. For example, Davis (2018) demonstrated that universities adopt transgender bathroom facilities in an effort to retain or increase their ranking. Yet detailed analysis of the effects of measurement practices in the development sector is relatively underexplored (See Hirschmann 2002a; Hoey 2015; Rottenburg 2009 for notable exceptions). Development scholars have called for such research, hoping that a shift in focus will move from how to perfect evaluative tools to interrogating what numbers *do* once present in social worlds (Mosse 2013; Viterna and Robertson 2015).

In this study, I conceptualize the mandated and standardized monitoring and evaluation (M&E) system of a 20-country agricultural initiative I call *Farming Sustenance* as a managerial structure that connects staff, projects, organizations, and donors. I interviewed 59 development professionals in East Africa and the donor country, asking about the potential to learn what works and what does not work from project M&E

systems—simple assessments of project performance and point-in-time external evaluations. Probing a cross-section of professionals, who serve implementing, evaluation, or donor organizations, about their ability to discuss or disclose failure reveals the limits of measurement practices to identify what does not work while under organizational imperatives to ensure repeat business, known as "rebiz." Empirically, I demonstrate that reactivity occurs through two dimensions: the managerial hierarchy of a single project and the comparative pressures felt at each tier within project management. A single development project creates a tiered, vertical managerial chain between donor and implementer, and their various offices. At the same time, at each tier—be it an implementing organization, a donor field office, or another agency—there is another entity competing for budget allotment. When combined with performance metrics, these singular and comparative structural relationships set off pernicious organizational effects that undermine the potential to use evaluation systems to learn. I conclude by arguing that reactivity to measurement practices turns evaluation into a knowledge production system that identifies "what works" for winning repeat business rather than "what works" for efficacious development.

Through this in-depth case study of how and why the reactions to measurement are so powerful in international development, I break apart the notion that monolithic donors force implementing organizations to complete measurement. Instead, *all* tiers within the aid chain are under pressure to demonstrate success. A sociological lens identifies where individual agency is constrained or emboldened by context. Here, I establish the theoretical importance of reactivity for making sense of measurement within the development sector, arguing that development professionals' interest in learning "what works" for poverty

alleviation has become constrained by a high-stakes measurement context. In this context, material benefits flow to professionals and organizations who reproduce a focus on securing results for "winning rebiz," while professionals who try to authentically learn "what works" are negatively sanctioned. From the perspective of an engaged academic-practitioner, I contribute a simple ethical question: What if our measurement practices undermine our ability to deliver "what works" to the world's poorest?

REACTIVITY IN INTERNATIONAL DEVELOPMENT

Development scholarship describes funding insecurity as a key driver in development organizations persistent need to demonstrate successful implementation (Best 2017; Cooley and Ron 2002; Krause 2014; Lewis and Mosse 2006). Yet the exploration of development measurement practices role in establishing this success and the consequences of such measurement on learning is relatively sparse (Mosse 2013; Viterna and Robertson 2015). Watkins et al. (2012) outline that evaluations "create success" through three core mechanisms: counting project activities (such as "children immunized"), gathering testimonies of project impact in individuals' lives, and producing consultant-penned reports. Beck (2017) adds that "successful" project outcomes are pushed by professionals seeking their own interests (jobs, promotions, future project funding) and that project coherence is constructed through reports, surveys, and databases. And Venugopal (2018), reviewing World Bank evaluations, demonstrates that failure is carefully constructed to lay blame on sociopolitical context rather than the organization. My contribution is to examine the relationship between evaluation, failure, and learning from the perspective of professionals grappling with the everyday minutiae of measurement practices.

Interrogating these relationships in development is important because donor governments increasingly demand evidence of the achieved results for dollars donated. This "results agenda" includes a growing focus on achieving, ensuring, and delivering (often quantitative) results, and it is underpinned by a linear understanding of project input to output to outcome to impact (Best 2017; Eyben et al. 2015; Vallejo and Wehn 2016). In many ways, the results agenda is part of a broader societal shift that focuses on the moral appeal of accountability and transparency, financial accounting practices extended into other sectors, with funding decisions increasingly determined by demonstration of performance—what Strathern (2000) termed "audit cultures." The development sector has not been immune, and global expectations have trended toward more measurement (Best 2017; Eyben et al. 2015; Hirschmann 2002b; Natsios 2010). This ranges from the Organisation for Economic Co-operation and Development's (OECD) high level fora institutionalizing quantified results as a core pillar for development effectiveness (Best 2017) to sectoral pressures to establish causal linkages between development projects and outcomes through rigorous M&E systems (Béné et al. 2017). Measurement practices are also a core component of the managerialist discourse which aims to bring private-sector efficiency into the public realm; this discourse is furthered by new powerful actors, such as the Gates Foundation, that promote a private-sector organizational culture (Schurman 2018). Managerialism recasts development interventions as objective and neutral through the use of reports, logframes, and results, or "new development management" (Dar and Cooke 2008). Proponents of the results agenda support the idea that indicators drive resources, attention, and aid in accomplishing project goals, while opponents are concerned that non-linear, less measurable processes will lose out as organizations rearrange activities

to fit easily quantifiable outputs (Eyben et al. 2015; Hoey 2015; Owczarzak, Broaddus, and Pinkerton 2016). Do performance metrics emphasize short-term results (Banks, Hulme, and Edwards 2015), work only under particular managerial arrangements (Hoey 2015), or transform project measurement from a temperature-taking thermometer into a pressure-cooking sauna (Hirschmann 2002b)? To provide both a scholarly treatment and a case of interest to practitioners, I explore the unintended consequences of measurement practices with respect to learning "what works."

Funding Uncertainty

The development sector is particularly vulnerable to measurement affecting rather than reflecting knowledge production. Historically, development organizations have been understood as values-based rather than profit-based (DiMaggio and Anheier 1990; Hwang and Bromley 2015; Watkins et al. 2012), and this perception has obscured both the importance of financial survival and the growing role of for-profit organizations in development work (Nagaraj 2015; Roberts 2014). Development organizations operate in a field of shared rules, yet funding, competitive, and principal-agent issues create organizational uncertainty (Cooley and Ron 2002). Thus, NGOs adopt strategies not solely based on beneficiary need but on organizational convenience (Brass 2012; Krause 2014). What unites both for-profit and non-profit development actors is that both serve as contractors, vendors, and implementors for bilateral and multilateral donors and receive contracts through competitive tenders (Cooley and Ron 2002). Krause (2014) demonstrates how development organizations become vendors, *producing* projects that are *purchased* by donors. In other words, donors are the client. With expectations of measurement

increasing under the results agenda, attention to how performance metrics interact with funding concerns is paramount.

The precarity of organizational funding creates a variety of pressures that impact accountability and learning. Because development organizations are proxies—charged with delivering development across long aid chains (Bebbington 2005; Wallace, Bornstein, and Chapman 2006)—donor organizations need accountability mechanisms (van Zyl, Claeyé, and Flambard 2018). This generates a principal-agent problem—the principal, the donor, hires an agent to execute their plan. Principals create measurements to ensure delivery of their ideas in faraway places (Best 2017; Ebrahim 2003, 2005) and impose these as conditions for receiving funding. This arrangement means that accountability shifts upward, to donors, rather than downward, to beneficiaries; securing donor funding is a precursor to delivering development to beneficiaries (Nogueira 1987; van Zyl et al. 2018). Learning, then, is shifted away from localized, contextualized knowledge that resists scaling toward quantifiable results that are better able to "travel" up to the donor (Ebrahim 2002). Thus, the demand for scalable knowledge for policymakers has a negative impact on localized processes that cannot be easily quantified, such as capacity building (Vallejo and Wehn 2016), democratization (Hirshmann 2002), or gender equality (Liebowitz and Zwingel 2014). Although scholarship acknowledges the outcome of these pressures, what occurs inside the development office to deal with these pressures has remained a black box.

Reactivity and Metrics

Reactivity as a concept, rather than a methodological concern to be reduced through triangulation or statistically representative sample sizes, focuses attention on what

measurement practices elicit. Reactivity is the idea that people change their actions once they know they are being measured, evaluated, or observed (Espeland and Sauder 2007). Humans are heavily reflexive, exhibiting reactivity under evaluation of almost any sort, not least when indicators create, name, define, and simplify what they seek to measure, and then are institutionalized to evaluate performance (Davis et al. 2012; Merry 2011). Studies of reactivity in other sectors, namely education, have documented the pernicious effects set in motion by institutionalized performance metrics, including gaming strategies and "teaching to the test" (Bowman and Bastedo 2009; Colyvas 2012; Davis 2018; Espeland and Sauder 2016; Sauder, Lynn, and Podolny 2012). Evaluative pressures are all around, creating evaluation landscapes of multiple measures (Brandtner 2017), although the effects of such pressures are likely to be greatest when a single measurement regime is present (Lamont 2012). In order to assess if and how reactivity occurs in the development sector, I study a single, mandated evaluation system, tracing the effects of project measurement within development offices.

Espeland and Sauder (2007) identify two mechanisms through which reactivity occurs. The first is the self-fulfilling prophecy, through which measurement practices stop measuring the social world and instead start to recreate and reorganize the social world. Using *U.S. News and World Report's* annual law school rankings as their case, Espeland and Sauder (2007) argue self-fulfilling prophecy is likely to occur when four pathways are present: strong attention to external audience reactions to the rankings, prior rankings' influence on current rankings, funding decisions made using rank performance, and when activities conform to what the metric measures. For example, when students (an external audience) began using the *U.S. News and World Report* to decide where to attend law

school, students indirectly forced schools to pay attention to their ranking in a bid to ensure a strong applicant pool. Further, if a law school invests staff energy and resources into activities that not included in the ranking indicators, they risk falling in ranking. Over time, schools are enticed into performing activities to maintain or advance their rank and forgo activities that are not captured in the ranking.

Together, these four factors harness professional and organizational energy toward the metrics—a managerial success. Yet, over time, as professionals orient behavior toward achieving the metric, the validity of the metric is increased in a self-referential loop. In doing so, these metrics iteratively shore up their own power. Social scientists have begun to document various effects of numbers in development (Adams 2016; Rottenburg et al. 2015), from emboldening experts to buoying the hope of people living with HIV/AIDS with irregular medicine supply by watching their CD4 counts. Thus far, practitioners rather than academics have produced structural critiques (Eyben et al. 2015; Natsios 2010). Utilizing the pathways identified by Espeland and Sauder (2007), I document the existence of a self-fulfilling prophecy in the managerial relationships created by project contracting in the development sector. In doing so, I account for large scale development initiatives and present a structural analysis which gives insight beyond individual projects.

The second mechanism through which reactivity occurs is commensuration. This is the social process of bringing different units under a common quantified metric, enabling ranking or comparison across categories of possibly unlike things (Espeland and Stevens 2008). The logic of commensuration is comparison: *U.S. News and World Report* compares colleges, the U.S. Census allows city comparisons, and actuarial assessments compare potential insurance policyholders (Espeland and Stevens 1998). Commensuration metrics

typically include rankings, ratings, prices, and cost-benefit ratios—any metrics that enable comparison (Espeland and Stevens 1998; Rottenburg et al. 2015). Rankings are particularly potent measurement forces because they widen even minute differences to a difference in ordinal ranking, raising the potential cost or gain in a rank order change (Espeland and Sauder 2007). Commensuration scholarship in development has largely focused on large-scale indices, such as the World Bank's *Doing Business* indicators and the United Nations Human Development Index (Davis et al. 2012), or on how countries react to such measures. For example, countries then make policy changes to enhance their rankings—termed "rank seeking" by Høyland, Moene, and Willumsen (2012)—qualifying them for resources, such as Millennium Challenge Corporation funding (Best 2017). Ravallion (2012) cautions that the questionable robustness of "mashup indices"—indices that combine multiple measures into a composite score—is obscured by the objective appearance of such scores. Additionally, by creating rankings that measure all countries, even empirical successes may be perceived as failures. For example, relative to other countries, African countries succeeded in addressing the Millennium Development Goal targets (Easterly 2009), yet this progress was perceived as failure because the targets were set at a consistent level rather than in relation to country-level baseline indicators.

Surprisingly, a structural account of the effects of comparison at the level of international development *projects* is absent. At the level of projects, rankings are uncommon and M&E systems are the preeminent form of measurement. Measurement at this level has often been considered a simple act of "counting" (Watkins et al 2012) and reactivity scholars note that "counting" does not elicit reactions by professionals and organizations. Yet when such counting is undertaken as part of a *standardized* evaluation

system, comparison between projects, implementing organizations, donor field offices, and entire development funding streams becomes possible. In this paper, I map the existence of comparative pressures in the development sector to the horizontal relationships at all levels of the "aid chain" from the perspective of development projects. Using the case of success and failure in evaluation systems, the following empirical sections document that reactivity does indeed exist at the level of development projects. I argue that reactivity to measurement hinders the ability to truly learn "what works" for efficacious development.

CASE & METHODS

To illustrate the complex role indicators, results, and organizational pressures play in the development sector, I conducted interviews with professionals who contribute to a single donor-mandated monitoring and evaluation system (M&E) in field and donor offices. A large bilateral donor, here called the International Development Assistance Department (IDAD), funded a 20-country agricultural initiative, *Farming Sustenance* (operating mainly in Africa). Adopting a multi-sited ethnographic approach (Burawoy 2000; Marcus 1995), I followed standardized performance metrics across offices as they compel people to collect and aggregate data and produce reports to meet bureaucratic demands in a single East African country (a recipient of the initiative) and in the donor country. This design was informed by my professional experiences in international development. I use pseudonyms for individuals and organizations and scramble some demographic and project details for anonymity. Anonymity is essential for ensuring the

analytical focus is on the effects of measurement rather than assigning blame (see Eyben et al. (2015), Mosse (2006), and Rottenburg (2009) for additional discussion).

I began interviews with staff serving the four main agricultural development projects of *Farming Sustenance,*[2] the evaluation-specific organizations (which focus on strengthening project M&E, completing external evaluations, and promoting knowledge learning and sharing, and donor field staff. I then interviewed experts or former staff to document the landscape *Farming Sustenance* operates within. Fieldwork then moved to the donor country where I completed expert interviews with donor staff and global program directors. This systemic approach is documented in Table 2-1. This systemic sampling frame documents the evaluation community as a contested political space in which people who do not share a moral universe work together (Shore and Wright 1997). By moving from field offices to headquarters, my interview data capture the ways development is translated across organizations and professionals (Lewis and Mosse 2006; Rottenburg 2009), which enables the underlying organizational logics shared across these sites to emerge.

The 59 interview participants were contacted due to their employer's role in *Farming Sustenance* or expert status. This includes professionals serving the initiative in the East African recipient country (21 recipient-country nationals and 25 foreigners working outside of their home country), and 13 professionals living and working in the donor-country (mostly donor-country nationals). In the recipient country, these

[2] With the exception of one project, which had concluded, and former staff were unreachable. The Farming Sustenance initiative in this country accounts for approximately $260 million in bilateral agricultural development assistance over 5 years.

professionals represent almost all of the management and evaluation professionals involved in the initiative. Interviews were 1-2 hours, semi-structured, outlined by a base

Table 2-1 Multi-sited Interview Frame: Organizational Sites and Occupations of Interviewees

Organizational Sites	Sample Interviewee Titles (n=number of individuals)
Large Project Field Offices (n=4)	Project Leader, Monitoring and Evaluation Specialist, Gender Advisor (n=13)
Evaluation Field Offices (n=5)	Project Leader, Deputy Project Leader, Senior Researcher, External Evaluator, Field Manager, Research Consultant (n=15)
Field Landscape	Chief of Party, Country Director, Agriculture & Gender Researcher, Gender Advisor, Policy Consultant, Evaluation Lead (n=10)
Donor Field Office (n=1)	Agriculture Officer, Project Officer, Responsible Representative, Evaluation Advisor, Gender Specialist (n=7)
Large Project HQs (n=2)	Director, Program Performance & Quality, Senior Evaluation Director, Director, Gender & Social Inclusion (n=4)
HQ Landscape	Consultant, Senior Advisor & Thought Leadership, Evaluation Director, Learning and Knowledge Management Consultant (n=6)
Donor HQ (n=1)	Director, Office of Learning, Evaluation, and Research; Gender Advisor, Africa Department (n=4)

Notes: "Landscape" includes expert interviews with professionals serving smaller Farming Sustenance or closely related agricultural development projects. To check that the evaluation system was mandated and standardized as reported in donor documents and observed in the primary case, I conducted 2 interviews with Evaluation Directors serving large projects under Farming Sustenance in another East African country. HQ=Headquarters. A total of 62 interviews were conducted with 59 professionals.

interview guide and took place in English—the working language of these projects. Verbal consent was obtained for all recordings. Analysis was completed in ATLAS.ti. Transcripts were first grouped according to their location and occupational role and then coded for mentions of contracting relationships and learning. Co-occurring quotations were taken as the basis of this article and then analyzed for differences and similarities between the tiers involved in the M&E system.

The purpose of interrogating metrics is not to indict the practice of monitoring and evaluation, but to attend to the ways it creates new templates of human interaction and resource allocation in an interconnected yet deeply unequal world. While my interview data comes from the donor country and one East African country, the indicators under study are standardized and implemented in all 20 countries. These managerial techniques, bureaucratic policies, and measurement processes are widely practiced by multilateral and bilateral development funders around the globe.

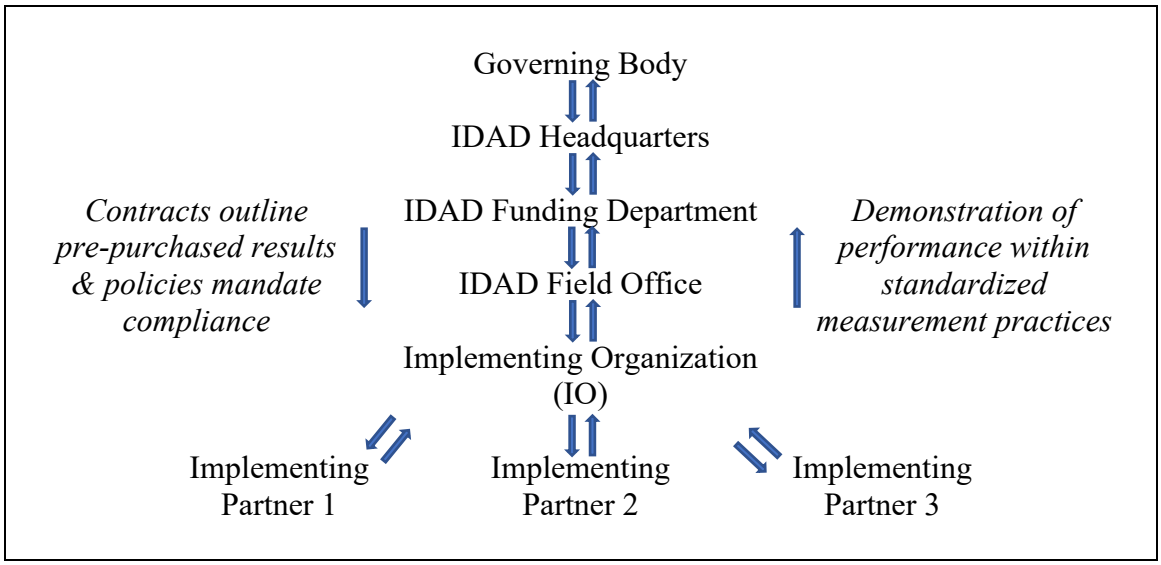
THE SELF-FULFILLING PROPHECY OF INDICATORS WITHIN A SINGLE PROJECT

Following Espeland and Sauder's (2007) notion of a self-fulfilling prophecy as a mechanism of reactivity, I demonstrate that reactivity is at play in development projects in three keys ways: (1) strong attendance to external audiences, (2) the mandate to track progress against standardized indicators, and (3) funding disbursement tied to performance. I argue these dynamics form the status quo of everyday development, resulting in an environment which strongly prohibits learning outside of the mandated project indicators.

The structural relationships involved in single large-scale development projects are hierarchical and tiered. When a bilateral donor hires (by awarding a contract or a grant) an implementing organization (IO) to deliver development to a group of beneficiaries, it is sourced through funds allotted by a legislative body. Bilateral donors often maintain oversight offices in implementation countries and IOs set up field offices as well. The IOs will then contract local organizations, often located in rural areas, as partners to implement

the projects (See Figure 2-1). I argue these organizational relationships, mediated by legal contracts that codify the expected results by the end of the project, set organizations "running" to deliver and manage results. Ultimately, these relationships determine the status quo of everyday work in development organizations.

Figure 2-1 Vertical Data Flow Schema on a Single Project



"It's Not for Us, It's for them": Data for External Parties

Are development professionals, far from the purview of donors, concerned about external parties? In other cases of reactivity, such as university rankings and tourism, the external audience for metrics are clear: say, students choosing which school to attend or families using "best vacation" lists to decide where to travel. In long aid chains, external audiences are harder to identify. However, development professionals spoke of an omnipresent awareness that numerous "others" are invested in project performance, perhaps none so prominently as the legislative body which granted the funding and the taxpayer.

Interestingly, the citizens of donor countries are the intended final stakeholder of bilateral aid money, yet geographic distance and layered bureaucracy limits their insight

into development activities. The penultimate stakeholders are elected officials in legislative bodies, such as Congress or Parliament, that disburse funds to government agencies and ministries. Several development professionals noted the importance of these two audiences in relation to admitting project failures. For example, a headquarters-based staffer who had also worked "in the field" said, "I think the 'talking about failures' part, it happens at the donor field offices. But they're going to keep it internal to the extent possible because of the awareness that people don't necessarily think international development is the right way to spend taxpayer dollars." Despite the geographic distance, development professionals are aware that performance is judged by tax-paying citizens.

Given the stated purpose of M&E in donor Evaluation Policies, I expected that indicator data must help project staff learn about efficacious development (what works and what does not). So I asked Project Leaders, the top manager of projects worth approximately $50 million each, whether the M&E system was useful for them to learn. "Well," one began, "you understand what it's for, right? It's for aggregating data to present to officials. I think there's a common misconception that it's the donor helping us select good indicators and helping us—it's for them. So we're happy to do it, but it doesn't help us, per se." For him, there was no utility to the M&E system; instead it is an obligation to an external party who uses it for their own purposes. This perspective was carried all the way up to an Evaluation Specialist at IDAD headquarters, who noted that their primary stakeholder was government officials "who need a clear, simplified version of the data." Project Leaders worried about external audiences beyond elected officials, too. They took part in measurement practices to help ensure that their organization would not, as one Project Leader said, land on the front page of newspapers for financial mismanagement,

or, as another worried, experience a "Gotcha!" from a government auditor should indicator data be incorrect.

Other IO staff concurred and spoke about the data eventually reaching legislative bodies and being used to demonstrate transparent taxpayer-dollars spending. Donor staff also articulated the importance of the *form* of data for officials. One IDAD representative in the field office noted how his own interests in capturing the project's systemic impacts on development beneficiaries are frequently overridden by the data needs of legislative bodies. A donor Agricultural Officer explained, "There is pressure to show results. Always. As an organization and as a culture, I think we have a need for immediate gratification." Thus, he continued, "IDAD presents to legislative bodies what they are getting for the money. [They ask] 'Is it having an impact?' It is harder to tell systemic impact. The story we give is: 'These many kids are immunized.'" So while the field donor staff want to report more meaningful results, he said that if his team reported, "We worked with [agricultural] extension workers to increase their capacity so that the seeds we're promoting are supported once in the farmers hands," it wouldn't be "a sexy story" or a "one- or two-sentence soundbite" amenable to policymakers and elected officials. Therefore, field staff alter their priorities to fit the needs of an external audience.

A Responsible Representative (donor staff who are legally responsible for supervising contracts) continued, noting that "legislative bodies are pushing. They want evidence and they want data, and for them that means numbers." The demand to deliver valuable data, as defined by elected officials, creates strong incentives for donors like IDAD to develop and mandate standardized M&E systems. The *Farming Sustenance* M&E

includes approximately 30 standardized indicators IOs must track. See Table 2-2 for examples.

Table 2-2 Examples of Required Indicators

Value of rural loans (monetary)
Number of farmers who have used new agricultural technologies
Number of firms that have increased profit
Value of target agricultural commodities exported (monetary)
Percentage of female direct beneficiaries consuming a diet of minimum diversity
Number of people who have received short term training on agricultural productivity

IDAD defines the indicator and requires varying levels of disaggregation (such as by age, gender, job type, religion, etc.). Once indicators are mandated, implementing organizations must collect data, not for their own use, but for the use of a party external to the development project.

Success is Defined by Indicators

If development project "success" is already defined by the indicators, can M&E systems produce actionable knowledge for project efficacy? International development contracting elicits reactivity by inducing activities to conform to the indicators' definitions of success. Project performance is measured against the contract, which outlines the implementing organizations responsibility to deliver "x amount of indicator A" over the course of the project. The overall amount for the life of the project is pre-identified and codified in the contract. This is then broken down into annual achievement targets that represent the sum total of contracted results. These bureaucratic procedures matter greatly for project management and the interpretation of timely progress. Time and again, professionals cited the difficulty of working in development. Some were glib, "I mean, c'mon, we're not

selling Hondas here," while others were solemn, "It's really hard to do it right, and really easy to do it wrong." Rather than flexible and responsive measurement practices, M&E systems are hardened sets of indicators that outline the expected delivery of results contracted by donors. An external evaluator describes how management is focused on achieving targets rather than thoughtfully utilizing data:

> As it stands, they [the IOs] collect the data, it goes into a table, it goes into a report, and it goes up the line. But it doesn't feed back into management the way it should. Management has already been set running on the contractual basis of the project, which is—you are contracted to do this, to achieve these targets, and off you go!

This external evaluator describes managers running to achieve the indicator targets outlined in the contract, orienting the behavior of professionals and organizations toward the *accomplishment* of the indicators, rather than the *substance* of what they were meant to achieve—reduction of poverty through enhanced agricultural practices. Although the logic of an evaluation system is that if the indicator targets are achieved it will lead to the reduction of poverty, professionals focus on demonstrating performance against the indicator targets and do not have time or space to interrogate if the accomplishment of indicators creates the sort of socioeconomic change envisioned.

Pre-purchased project performance creates a bureaucratic scaffolding that donor representatives can use to force implementing organization actions. A professional who frequently visits donor field offices commented on the managerial role of the IDAD

Responsible Representative in the field office as they manage the project contract with the implementing partner staff: "I think if the IDAD rep has the awareness that we cannot expect everything to go perfectly; we're not implementing development assistance in an experimental situation where we can control all variables, that's just not how it works." Imitating the stern tone of the IDAD rep chastising a contractor, she went on: "'They didn't do what they said in the contract!'" And then in an explanatory tone, she continued, "because it's about expectations. You [the donor rep] can't expect things to be held constant, and there is a lot of uncertainty and a lot of complexity when we're implementing projects." For her, donor inflexibility was a huge frustration: "you're going to have to figure out what that is [the complexity] and you're going to need to change." She contrasts the needs of the project to learn about the complexity at the project site and design activities accordingly against the managerial power expressed by the Responsible Rep, enabled by contracts that outline the results to be achieved. For example, one Project Leader mentioned that he was unable to "slow down" to address the needs of illiterate women who needed extra support to engage in business proposal development because doing so would detract time and resources away from accomplishing other "more important" indicator targets.

This also creates a quandary for IOs who realize they want to alter project implementation. The donor Agricultural Officer linked design and indicators, saying, "if the implementing organization wants to change, they have no incentive to because the terms by which they will be judged are already set and signed upon." This dynamic was confirmed by an Evaluation Director for a field implementing organization, who revealed that working to alter the project indicators to better fit the project delivery:

> Becomes a null and void debate, because [the indicators] will be the basis upon which you are judged, so you are better off just paying attention to achieving those results rather than spending your energies debating… you realize that if you change the design, then the indicators you are going to be measured against are different, so then you are stuck.

Even if a development professional observes a mismatch between the metrics and what project beneficiaries actually need, the measuring stick is mandated. Knowing that judgment will be based on these approved metrics not only curtails motivations to work toward enhanced project management and learning, but also serves as a lock or restraint. Since bilateral donors require aggregable data for legislative bodies, a single project may not alter the greater standardized M&E system. The only option is to be measured by it or not—and not being measured would result in the withdrawal of donor funds from the project.

The donor definition of success as "bringing the figures" and codified in the contract means that implementing organizations have no space to learn or ask questions during implementation. This is compounded by their fear of loss of funding each year. An external evaluator explained, "If projects are constantly feeling like they need to demonstrate that they've met all of their goals and targets, then there is no room to say, 'Maybe those goals and targets weren't the best designed' and learn from what hasn't worked, why those aren't appropriate and what maybe would be better." Several IO staff thought the indicators did not make sense for their projects but were resigned to complete them because "they're required." Evaluation Specialists in the donor country felt that a

"strong evaluation" was a report that addresses the project indicators clearly. This suggests "good" evaluations are ones that meet bureaucratic needs for audit and stay within the boundaries of the M&E system rather than foster local learning.

In essence, M&E practices, though discursively acknowledged as a well-intentioned and valuable activity, take on different meanings in everyday work life. Performance targets increasingly cultivate the attention and resources of IO staff, pushing meaningful project-based learning to the side. And within the vertical relationships between donor and implementer, contractual obligations and standardized indicators mean that the easiest path for implementing organizations is to comply with bureaucratic requirements. Organizations react by working to fulfill metric targets, with little opportunity to alter the project design to enhance utility to the beneficiary because such innovation could disturb the aggregable measurements of success and jeopardize funding.

Rebiz Rules the Day

Can implementing organizations share failures with their current and future client—the donor—without negative repercussions? Plausibly, this might come in the form of post-project learning, which could elicit corrections for future projects. And yet, since implementing organizations are interested in maintaining their reputations and relationships with the donor (a potential future funder), practitioners noted that successes are over-emphasized in multiple formats even after project completion. A Gender Adviser with an IO mentioned a wrap-up presentation at a fancy hotel in the capital city. When I asked about lessons learned from the project, she interjected jovially, "Only positive

lessons!" She laughed as she continued, "You know, this is like promoting us for the coming [project], maybe fundraising from IDAD… I don't know, but nobody was talking about negative things, so I kept silent because I didn't want to disturb the environment." Moments later, she back pedaled, noting that IDAD is willing to hear both positives and negatives about projects, but her instant, in-the-moment reaction was to describe how she, a seasoned professional, sat in silence at a wrap-up presentation, not wanting to rock the boat. In this circumstance, the agency of the development professional is constrained by the overall tone of the event, which is focused on ensuring future organizational funding.

Today, Silicon Valley touts failure as the ultimate learning opportunity. But within development circles, the *absence* of failure, errors, and mistakes in events, such as the one above, is triangulated by another data source: project evaluations. These are point-in-time evaluations that assess a project at the beginning, middle, or end of a project. A field consultant previously employed at donor headquarters, explained, "It would be good if they [IOs] could say 'we're failing.' That's a recognition that is important for learning. But a lot of projects will be afraid to say they failed at anything. For funding purposes, it doesn't look good." She mused, "Maybe at the end of the project?" before concluding, "No, no one, you don't say it and you're not asked about it either… If people do go back and look, then the goal is to have a successful evaluation. From the point of view of the IOs, they don't want others to know the negative, they want evaluations that show success. The next time they apply [for funding], that will come up."

The need to secure repeat business—"rebiz"—was a theme that carried through all my interviews. In this relationship constellation, all actors are locked into a vertical pattern of success: If an implementing organization failed in a project, it would reflect backward

onto the donor, reverberating up a chain of invested professionals and organizations. As the quotes above demonstrate, professionals frequently demonstrate the interlinked nature of "data success" with repeat business. Speaking in this manner demonstrates the naturalized and omnipresent way that development professionals are aware their performance is under measurement and that funding consequences may ensure. Regardless of winning future bids, funding is uncertain even for current projects; each year, multi-year contract funding can be pulled at any time by the donor. In this context, organizations are on constant alert for what several professionals called "being dinged." A donor Agricultural Officer agreed: "Sometimes we need to rewrite the project intentionally, but IOs can get dropped or canceled every year, it's written into the contract, dependent on funding or a documented reason." Here a donor staff member admits that sometimes the project design outlined in the contract may be ineffective yet raising this discussion with the donor introduces uncertainty into the next funding disbursement. That is, even when professionals recognize a purposeful change may be needed to enhance a project, this becomes a risky organizational discussion—one that may result in the cancelation of funding.

It is not surprising, then, that IOs carefully manage metrics to highlight success alone. An Evaluation Director said, "I don't want to make this sound as bad as it probably is going to sound, but [the Project Lead's] job is to keep IDAD happy. His job, when he got here, was to mend fences because we were not in IDAD's good graces. And so he did that, and we're now considered doing very well. And it's not because we are doing well, it's because he keeps them happy. He's taking them out when they want to go out. He emphasizes the good numbers we've got, and it's all the nice things." Project Leaders must

be skillful relationship managers and "good numbers" form an important currency within the development landscape's focus on results.

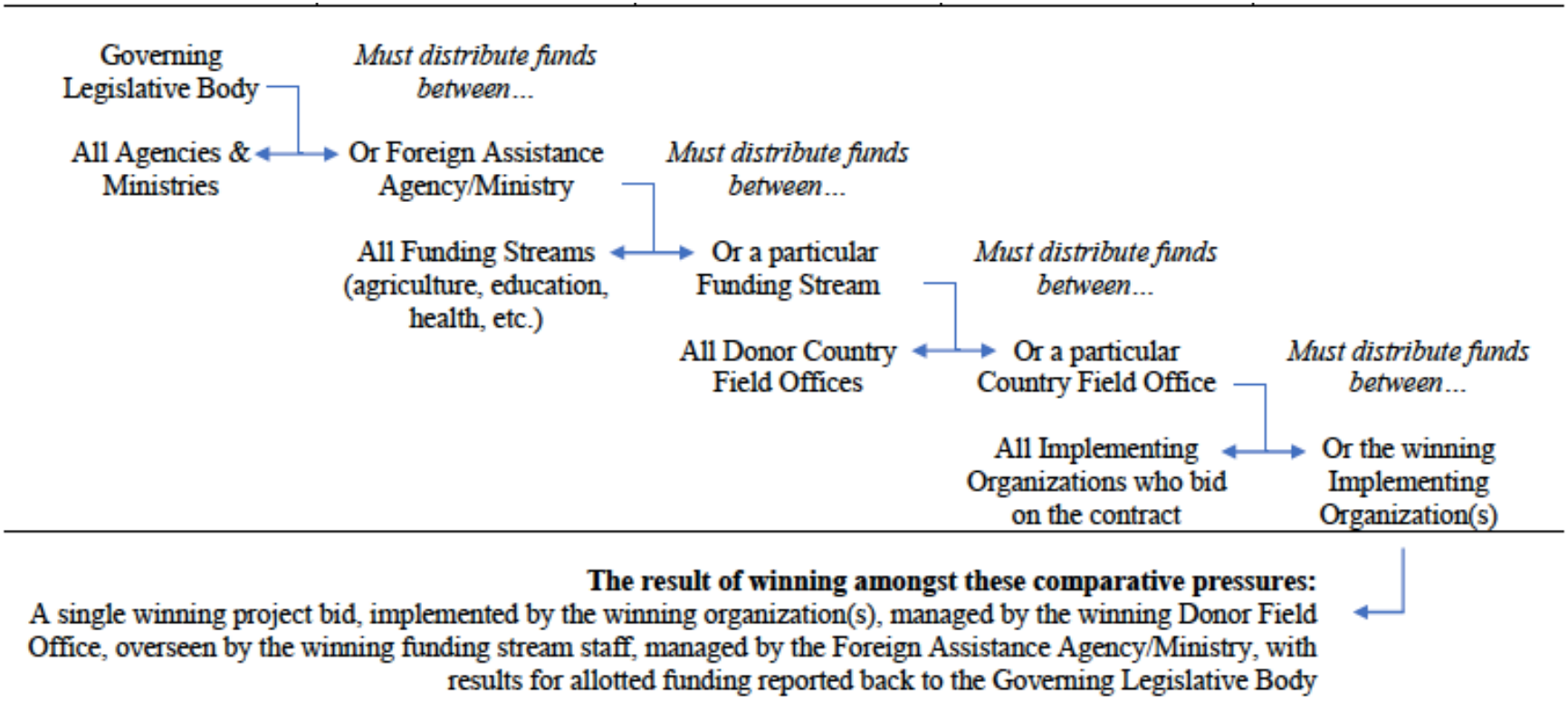
In sum, implementing organizations, under persistent funding uncertainty, view the demonstration of results as a basic task needed to secure future work. Current performance is carried forward through the inclusion of "prior performance" in future Request for Proposals. Magnified by the rising expectations of accountability, transparency, and results in an actuarial era, the ability to demonstrate, document, or present results is highly valuable in securing development funding.

Through the intertwining of external audience, definitions of success codified in mandatory indicators, and funding decisions based on performance, the cost of admitting failure is heightened. And so reactivity is perpetuated in the international development landscape, with metrics overshadowing dynamic learning about "what works" for recipients. A consultant bluntly describes a focus on successful bureaucratic documentation of goals, rather than on the changes in people's lives: "You [the IO] focus on what was successful. The funder doesn't want to raise [failure]. They [donors] look at goals achieved, not the impact on people." M&E systems deliver knowledge about "what works" for organizations to garner future business rather than "what works" for efficacious development. This vertical relationship-management across the tiers of project implementation creates the status quo of everyday development work.

COMPARISONS OF PERFORMANCE: HOW COMMENSURATION OPERATES

Within development contracting, there are numerous decisions to be made about how funding will be allotted, to whom it will be disbursed, and if disbursement will continue. Standardized M&E systems enable comparisons between potential funding recipients. Comparisons are made at a variety of moments throughout the project cycle and at different tiers in aid chains. The managerial relationships described above, which form the status quo of everyday development work, occur not in isolation, but in multitude. Here, I demonstrate that commensuration, or the ability to compare, between peers at each managerial level of development negatively sanctions any professional or organization that transgresses the status quo by acknowledging failure or questioning indicators. That is, professionals and organizations, irrespective of their intentions, have little ability to escape the comparative gaze of the tier above them, tasked with making decisions about funding disbursement. Learning has already been placed in a precarious position because of structural concerns between the donor and implementing organization around the need to deliver results and ensure "rebiz." As I demonstrate below, the marginalization of learning is *maintained and reproduced* through a series of comparisons that punish professionals or organizations who attempt to introduce more authentic forms of learning—forms that acknowledge mistakes, activities that did not produce results, or slow progress. See Figure 2-2 for a representation of how the decision to distribute funds to the tier below creates a waterfall of comparative assessments at every level of development contracting. These assessments place pressure on professionals and organizations to present strong performance consistently.

Figure 2-2 Moments of Comparison for Funding Distribution Down a Bilateral Aid Chain



Comparative pressures are popularly understood at the Request for Proposal (RFP) stage—when a call for project bidders is released by a donor, and implementing organizations build teams to research and write applications in the bid for funding. And while professionals confirmed this stage as important and spoke of the need to win funding, there are many other moments of comparison—between donor representatives in fields offices and projects that share a technical focus to name a few.

Wanting to understand where possible sources for changing the "measuring stick" problems described above, I asked professionals where M&E systems begin. Most replied "in the RFP." However, an Evaluation Director in the donor country noted that although the RFPs usually include a mishmash of items, yet for the IOs, "their incentive is to include everything. All the keywords, whatever, regardless of whether or not it entirely makes sense…if there are mismatches or things that don't make sense in the RFP, they're not going to push that too hard because they wanna win it." Another example was provided by a Project Leader, who noted that IOs are careful to include results that are appealing to the donor in their bid submission:

> Now IDAD projects are competitive bid, so there is very little space for learning because you don't get many points for having a very elaborate learning agenda saying, 'Okay we're not going to do $100 million in exports, we're gonna do $5 million but we're going to give you a lot of lessons learned for the future.' That doesn't sell. Under the [competitive bid] model that's why there is less time and attention for learning.

Learning and indicators are fully divorced in this rendering, and competition between implementing organizations partially to blame. Later, the same Project Leader noted that implementing organizations can be like "two dogs fighting." To play the game—to "win" funding—means being strategic in presenting an organization under a logic of comparative evaluation with their peers. Since donors create RFPs with their own concerns in mind—ensuring elected officials and taxpayers are satisfied—they effectively initiate a process in which IOs present their ability to accomplish the desired donor indicators, divorcing M&E systems from the project context.

Understanding the "measuring stick" structure created by M&E systems in a single project, comparisons at each level serve as additional locking mechanisms to ensure an implementing organization maintains the original design, even if poor project design is acknowledged. In the event that an implementing organization pushes for a design change and takes that request to the donor, the donor is disincentivized to address the request. The donor *Farming Sustenance* manager in the field noted, "But if you change too much [during implementation], then other bidders get upset. If tweak here or there, it's okay, but if you alter from the original scope, they will say 'we could have won this. You put a bid out for project x but now implementing project y, and we could have won project y.'" In this circumstance, implementing organization headquarters put winning contracts ahead of meeting the needs of the people projects are meant to assist. An External Evaluator agreed, noting that "if you see, even a third of the way in, this is nonsense and it's not going to really get us where we want to go, you've got to have the nerve to change the design of your program" but that other bidders to the original project will become angry, "People get

really worked up about this because its millions of dollars." This is a downstream effect of competitive contracting that results in a horizontal squeeze, rather than the more obvious vertical lock-in.

Horizontal pressures of comparison play out at donor Field Offices, too. A Project Leader of an evaluation contract discusses the working dynamics within the donor Field Office, drawing a distinction between professionals interested in advancing their own careers and professionals who arrive in the Country Office interested in learning from the projects. He identifies a process of negative sanctioning that trains professionals interested in learning "what works" locally to instead focus on standardized indicators, which allow for the ability to compare projects:

> Other guys are just as ambitious but more ambitious about what the project is teaching them. Unless the Head of Office, the top 2-3 people set that agenda, you end up with different people doing different things. The super ambitious [career-focused] guys do not ask any tough questions to their team. The other guy [project-focused] forces tough questions and, if so, you lose 20% of your whatever because you've done a much more rigorous assessment. When they all come together it's, 'Oh, how's your project?' 'Oh, how's yours?' [snide tone and mimics looking over and down one's nose at a colleague's paperwork]

Those who want to learn "what works" for development programming ask tough questions, trying to ascertain what works for project beneficiaries. Those who want career

advancement stick to the project indicators. When these two different approaches come under a performance review, the donor staff that asked harder questions and did a "more rigorous assessment" is negatively sanctioned precisely because his counterpart claims success. The "career" professionals work within the logic of the standardized indicators, primarily based on output numbers. In other words, those who hope to effect change in people's lives may disregard the ontological status of the project indicators and seek a more empiricist interpretation of data about "what works" but they do so to the detriment of their careers.

The daily tasks of managing projects, results, and career paths are interlinked in ways that make it hard for professionals and organizations to put learning before concerns about their reputation and status in a context of constant comparisons. This idea was reinforced by a Country Director who had worked as a Policy Consultant to the donor field office. He described the same dynamic but between implementing organizations. This Country Director built a coalition of individuals across organizations to openly discuss and share empirical findings on "what works and what does not", only to find that comparative pressures arose within the "technical area" of the project compared against performance in other countries:

> What's interesting is even with demonstrating these accomplishments, I'm sure we could have done lots more. It's not a game-changer because it hits up [against] rigidities everywhere. For example, I think a lot of the stuff being done on agroforestry next door is just really dumb stuff and hasn't been thought through, and yet all the reports are coming up being

> successful, right? Whereas, here, we've learned some stuff and we're doing really good stuff and we're able to prove things are happening. But because we said, "This hasn't worked, this hasn't worked," the perception was "Oh, well, the neighboring country is doing a much better job because all of *their* projects are working, whereas yours are not producing." And I just thought that's just so completely opposite of the truth.

Note that the *documentation of success* matters more than the evidence built or shared. The Country Director explicitly noted he had formed a "critical mass" of in-country development professionals and thought that would be sufficient to alter what is permissible to say around results and learning—that some things work and other do not. But reported results in the same technical area in a *different country* resulted in a donor complaint: comparatively, his project was not performing as well. The importance of how comparative pressures can enable one tier in the aid chain, while constraining another is clarified by a Project Officer with the donor Field Office, who noted that aggregable indicators are: "important because we're able to aggregate data across countries globally to get a good picture of the impact of the programming which we aren't always able to do in our technical sectors." In other words, this donor staff member appreciated the comparability of the very indicators that hindered the Country Director from showcasing his learning of what is most efficacious for development.

The intertwining of documented success, organizational reputation, and securing future contracts undermines learning from multiple angles. An evaluation Project Leader

offered an example of his "worst" work in development and how he wanted to share it with others for the purposes of learning,

> I want to say to others "Don't do this again, because I did this whole pile of a mess here." We don't do that. I think that we are still very conscious about failure. We need to be able to get big contracts. So we have to change the culture, we have to make it so that instead of data and monitoring, there has to be rewards for learning and humility. We need reward systems. Ultimately, we go where the money is, and if there are big bucks for big success stories that are pretty spurious—and that's how organizations get rewarded again and again and again. The guys at the tops of these organizations, they're for keeping it, because these CEOs are a massively competitive group of people, they have egos the size of flags. But I would like some guidance about their greed, [it should be] about evidence-based stuff rather than just about success stories.

Rewards, data, organizational reputation, and professional drive are an interconnected, locked system that blocks evidence-based learning and instead promotes spurious success stories. This Project Leader mentions that these dynamics are generated by IO CEOs, people who are not present in the field, because they experience a comparative pressure when competing against their peer organizations to win additional contracts. He raises the idea that these CEOs could compete on more rigorous terms, evidence-based terms, rather than simple success stories that may not be well-evidenced.

Most of the professionals I spoke with suggested that learning inherently requires failure or poor performance, while also noting that there is little space in the development sector to acknowledge such occurrences. My closing question in interviews was "What keeps you up at night?" Nearly all staff, serving implementing organizations and the donor, responded with an iteration of: "Are we getting it right? Are we having an impact? Does this matter?" Professionals are heavily interested in "what works" for global poverty reduction yet are constrained by the development structures that demand "what works" for repeat business.

Comparative pressures operate throughout the aid chain system. IOs imagine their competitiveness to win a bid and develop content accordingly. Responsible Representatives compare the projects under their management. Country Offices compare implementing organizations across countries by technical area. CEOs at the largest development contractors compare the projects in their portfolio. This creates a system in which learning is squeezed out; entertaining failure, error, or mistakes openly invites negative sanction. Thus, these horizontal comparisons keep development professionals and organizations "in line" and raise the opportunity cost of transgression. In the process, agentic attempts at changing the culture around learning (and failure) are structured out. To understand professional and organizational decisions under the results agenda, we must attend to the pressures of comparison made possible by standardized M&E systems. While these comparisons make sense from a managerial perspective, taken together they have detrimental effects on the development sector's ability to learn "what works."

CONCLUSION

This article documents the ways performance metrics have become integral to the structural dynamics of the development sector; recruiting the labor, attention, and anxieties of professionals and organizations they are being measured and evaluated in a competitive landscape. I analyzed how and why professionals located throughout the transnational evaluation chain are drawn into a focus on achieving timely progress against performance metrics rather than a concerted assessment of project efficacy on its own terms. Two empirical findings emerge from this work of import to development scholarship. First, reactivity, or changes elicited from professionals and organizations because of measurement, is indeed occurring in development projects. This establishes the analytical and explanatory value of reactivity to development scholarship and encourages the use of reactivity as an analytic framework in future research. Second, I expand the scope conditions of extant research—which focuses solely on rankings—and demonstrate that simple project output indicators, once part of a standardized M&E system, set off pernicious effects that undermine knowledge production about efficacious development.

The first contribution is to establish that reactivity to measurement practices in development is, indeed, occurring through the self-fulfilling prophecy of measurement in a single project and acts of comparison at every managerial level. The presence of the self-fulfilling prophecy calls into question the data produced through M&E systems as empirical measures that document "what works and what does not." Instead professionals and organizations make decisions and determine actions around the metrics. The circular logic of the self-fulfilling prophecy occurs through three main activities: attending to the

needs of external audiences, matching project activities to the indicators, and concern for future funding. These activities set the status quo of everyday labor with a development project. The forms of communication between an implementing organization and the donor create the structural backdrop for managing large scale development projects, within which results serve as a core mode of bureaucratic communication. By defining project "success" *as* meeting bureaucratic demands for results, attention and resources are routed away from the substance of the project and toward the achievement of the performance metrics. This serves to strengthen the validity of the metrics—a successful managerial strategy for donors—but upends learning "what works" for improving people's lives. Instead, organizations engage in data collection and reporting "what works" to maintain their relationship with the tier above them in a long aid chain, with the aim of repeat business. The status quo of development then becomes to deliver successful results to the client—the donor—rather than a focus on the *substantive activities* meant to deliver such results.

This status quo is maintained and reproduced through comparative pressures at every level, negatively sanctioning any professional or organization that attempts to disrupt this status quo by introducing more authentic forms of learning. Acts of comparison exists across the multi-tiered playing field in which everyone's reputation is interconnected and passed around via documented performance. Implementing organizations are commonly understood as being perverted by donor demands, yet my research demonstrates that donor staff and offices experience these same pressures. These pressures influence how professionals navigate relationship-management with those in the levels above and below them. Watkins et al (2012) describe the *forms* through which success is evidenced, noting simple output counts, evaluation reports, and performances for donors. I contribute

reactivity as the *mechanism* by which success is created. More broadly, the evaluative structures created by development practitioners to use as a tool to better implement development have now become structures that use development professionals as a tool to implement development within a very narrow frame of action. In this context, professional agency and intention to learn from the local experience of the project is constrained.

Secondly, this study demonstrates that simple counting measures, at the level of development projects, are capable of eliciting reactivity—expanding the application of reactivity to development activities. Thus far, development scholarship has focused on country-level reactions to large-scale indices, such as the Human Development Index and World Bank's *Doing Business* Index. Most commensuration scholarship argues that rank ordering—which widens even minute differences and punishes those who attempt to "opt out"—invites a particularly insidious form of reactivity (Espeland and Sauder 2007; Davis et al 2012; Rottenburg at al 2015). In order to gain more explanatory power from reactivity, I bring attention to the importance of metric *type* and *scale* in the development sector. Global *ranking* has been shown to elicit *government* reactions to garner international foreign investment or development assistance (Best 2017; Davis et al 2012; Høyland, Moene, and Willumsen 2012). I widen the analytical territory for reactivity by demonstrating that *standardized* M&E systems elicit *organizational* reactions to ensure reputation and resources for the next project. This finding is important because the structure of development contracting and projects obscures the greater comparative landscape simple indicators exist within. What appears to be a simple count for a single project is actually a single standardized indicator, against which data is collected, aggregated and tabulated in projects around the world. Furthermore, it is not quantification itself that

necessarily results in the deprioritization of immeasurable activities, it is *the act of comparison* made possible through *standardized* M&E systems. By taking this broader view, intentions to learn "what works" for development efficacy would do well to not engage in a politics of comparison and instead produce localized knowledge not meant to scale. Quantification in and of itself is not the concern, but quantification for purposes of comparison likely reduces the ability and space for concerted learning.

Regarding scale, I bring reactivity to the level of development projects—arguably the fundamental unit of international development programming. It is through the unit of the project that development is meant to be delivered, said to ensure qualitative changes in the lives of the global poor. As the development project is the means by which the development sector interfaces with the global poor, the effects of measurement practices within projects will most impact how "beneficiaries" experience development. Understanding the role of metrics at the level of projects also opens up more analytical possibilities—the careers of development professionals are made around the unit of the project and organizations ensure survival by winning project-based contracts. Therefore, understanding that reactivity operates through simple project output metrics and at the level of development projects demonstrates the importance of reactivity theory to development scholarship and invites rich future research.

There are two additional findings of import to understanding this moment of measurement, possibilities for learning, and efficacious development programming. The first is that data has become a particularly strong form of currency to be traded within and between organizations. The lived experiences of M&E practices as told by development professionals is in stark contrast to rhetorical calls for learning "what works." Cooley &

Ron (2002) outlined agency problems, competitive tendering, and multiple principals as the underlying mechanics of development organizations' self-interested behavior. I find that these dynamics are persistently present yet now deal in the currency of auditable data. The ability to claim success in the form of results is now a requirement for development projects, and thus organizational survival, in a competitive field. To hold this currency is to have power: Professionals use it to advance their careers, organizations use it to advance their reputation and manage donor relationships, and donors, too, use it to manage departmental reputations with legislative bodies. In this context, professionals and organizations have strong incentives to work together to create success vis-à-vis project monitoring and evaluation. This comes at the cost of using the M&E systems to learn "what works" because, in a high stakes context in which performance metrics are institutionalized, admitting slow progress or failure in an auditable form is a risky professional and organizational move. Professionals describe an interest in promoting learning from mistakes, only to be silenced by their peers or organizations. This has serious implications for practitioners: The impetus toward standardized, comparable metrics, mandated by development donors and reported by implementing organizations then undermines rhetorical claims to openness and a willingness to learn from mistakes. Absent learning, development measurement may become a shell industry which uses output numbers for funding legitimation without a serious intention to measure, understand, and progressively deliver efficacious development.

Secondly, auditable data should be considered part of the product that implementing organizations sell and donors buy (Krause 2014). Scholarship frequently critiques quantification for focusing on short-term output over long-term transformative

change (Banks et al. 2015; Fukuda-Parr, Greenstein, and Stewart 2013; Merry and Wood 2015). I argue, instead, that commensuration reminds us to take a wider view to understand how comparative pressures are shifted and transformed as they move through aid chains across geographies. Although "counting" may begin at the terminus of the aid chain, as Watkins et al. (2012) state, commensuration processes begin when donors determine a mandated evaluation system. Contrary to much scholarship, then, measurement choices are not made by implementing organizations to please the donor but are made within donor headquarters as a managerial strategy to please the external audience of legislative bodies and taxpayers. In other words, data amenable to policymaking must be delivered and implementing organizations are legally contracted and paid to share this burden. Measurement practices, then, are better understood as a key product feature that implementing organizations sell and donors buy.

In making this argument, I hope to draw attention to the importance of data, results, and measurement in international development and how metrics are a strengthened currency in an era with stated interests in "evidence-based decision making." Learning provides an interesting case study because it holds the moral appeal of "doing better" and enhancing development efficacy for the benefit of people living in poverty around the world. It also identifies the outer limits of the knowledge generated through standardized M&E systems. Although the development sector is premised on alleviating poverty, I argue that high stakes measurement practices supersede this mission by recreating rather than measuring the social world. Community based or local organizations may stand to lose the most under these circumstances. Winning a contract with a large bilateral donor may enhance an implementing organization's status yet meeting donor monitoring and

evaluation demands may leave organizations "feeling like a stranger to ourselves" (Eyben et al. 2015) as they reconfigure their work toward reporting requirements. Rather than advancing the effectiveness and efficiency of international development programming, performance metrics as managerial tools instead reorganize development activities around the metrics. This ultimately constrains professionals' ability to learn "what works," share and discuss the challenges of development work, and iteratively improve the impact of development projects in the lives of people around the world. Development scholarship must critically attend to how the organizational pressures we have created, taken part in, and reified have then crystallized into structural *constraints* that undermine the very intention of development: ending global poverty.

Chapter 3

Bureaucratic Tools in (Gendered) Organizations:

Performance Metrics and Gender Advisors in International Development [3]

ABSTRACT: This article contributes to a growing conversation about the role of numbers in promoting gendered agendas in potentially contradictory ways. Drawing from interviews with gender advisors—the professionals tasked with mainstreaming gender in development projects—in an East African country, I begin from the paradox that gender advisors articulate a strong preference for qualitative data to best capture the lives of the women they aim to assist while voicing a need for quantitative metrics. I demonstrate that (women) gender advisors come to imagine metrics as expeditious bureaucratic tools able to inspire cooperation from otherwise reluctant (men) coworkers. I argue that development organizations are gendered in ways—acutely seen in how advisors struggle, are sidelined, and attempt to advance their goals with numbers—that lead to the utility of valuing quantitative metrics over qualitative ones. I establish two theoretical contributions: (1) Gendered organizations theory is essential to understanding the adoption and globalization of performance metrics, and (2) in an age of evidence-based decision making, the utility of quantified data to garner resources is heightened, rewarding those who adopt quantified knowledge production. I coin the term "the paradox of quantified utility" to describe how these material advantages encourage even skeptics to value quantitative metrics.

[3] This chapter was published in *Gender & Society* in 2019. The version included here is the author's manuscript and does not represent that final manuscript, which is available online: https://doi.org/10.1177/0891243219874058

> Quantitative data—Ha! I'm saying we need more of it!—is also a bias against the global women's movement and Global South intelligence, expertise, and knowledge.
>
> Irene, donor country-based gender advisor

Social scientists are interested in the increasing role numbers play in modern society through valuation, evaluation, knowledge production, and allocation of resources (Espeland and Sauder 2007, 2016; Lamont 2012). Numeric data and indicators have recently captured the attention of gender scholars across disciplines (Bose 2015; Liebowitz and Zwingel 2014; Merry 2016). Still, little attention has been paid to how professionals respond to indicators and why. Indicators, or quantifiable measurements, are typically used as performance metrics to measure progress, such as prenatal care rates at a hospital, women staff and promotion statistics in *Fortune 500* companies, and global progress toward Sustainable Development Goals (UN General Assembly 2015). Some development professionals, however, state a preference for qualitative data, which they see as more able to capture the experience of the people with whom they work than numbers. So why might professionals, like Irene, want *more* quantitative measures? I examine this puzzle using the occupational case of gender advisors—the professionals tasked with mainstreaming gender and engaging women as participants in international development projects. I demonstrate that this question can be answered by understanding the highly gendered nature of their occupation and the ability of performance metrics to alter behavior and decisions within organizations that prioritize metrics. Indeed, quantitative indicators help make women's

work visible and build solidarity around issues that disproportionately affect women, among other uses (DeVault 1996). Amid the rise of organizational decision making based on quantitative performance metrics, scholars must interrogate how gendered organizations and quantification intersect and with what implications for knowledge production.

The growing scholarship on *governance by indicators* focuses on how performance metrics combine with organizational structures to alter organizational and professional behavior (Berman and Hirschman 2018; Muller 2018). For example, Alexander Davis's (2018) case study demonstrates how the quest for high rankings drove universities to unexpectedly adopt transgendered bathroom and housing facilities. My research advances scholarship on the role of numbers in society and the paradoxical reactions that quantitative performance metrics—a rising tool of neoliberal management—may elicit in organizations and from professionals, particularly in gendered organizations.

Getting gender equality on the development agenda was a hard-won victory of the 1980s and 1990s. Today, many development organizations in the Global North mandate gender mainstreaming—bringing a gendered lens to the design, implementation, and evaluation of development projects (Cornwall and Rivas 2015; for an insightful overview of gender in development, see Ransom and Bain 2011)—and employ a class of professionals, "gender advisors," to accomplish mainstreaming (Ferguson 2015) in the Global South. Although advisors and mainstreaming exist to alter gender dynamics, raising the gender saliency of workplace interactions (Ridgeway 1997), Acker (2006) cautions that what appears as a progressive step may be a reconfiguration: The experience of a job may vary, yet occupational roles and gender typing may reinforce hierarchies. In response, I

analyze the occupational experiences of gender advisors across development organizations—a sector vigilant about achieving results.

This article is based on interviews with international development professionals employed in donor and implementing organization offices in an East African country. I use the case of donor-mandated performance metrics in an agricultural initiative to show how and why gender advisors, skeptical of quantitative metrics as *measurement* tools, come to embrace them as *bureaucratic* tools able to alter coworkers' behavior. By using gendered organizations as an analytic framework (Acker 1990, 2002, 2006; Ridgeway 1997, 2009), I demonstrate how gender advisors are structurally subordinated and emotionally fatigued from workplace interactions, which generates their interest in strategies capable of engaging their coworkers in gender mainstreaming. Gender advisors see the power of indicators in the development office workplace; and, although they may view them as problematic measurements of women's lives, I argue they come to value such numbers for their bureaucratic, micropolitical power. This case contributes to our understanding of how and why resistance to quantitative metrics is silenced in gendered organizations and their impact on knowledge production.

In what follows, I synthesize literature from development organizations, gendered organizations theory, and governance by indicators. Empirically, I detail gender advisors' quotidian work experiences and their dismissal of quantified indicators as valuable knowledge about women's lives, and demonstrate that their paradoxical desire for such metrics is driven by their imagined future workplace once gender-related metrics are more strongly present. I conclude by highlighting two theoretical contributions: First, professionals' reactions to performance metrics in the workplace are indeed gendered,

suggesting that future research must incorporate gendered organizations to understand the nuanced pathways by which metrics are globalized, and particularized, as managerial tools. Second, quantitative indicators are not necessarily a superior form of knowledge, but are perceived as such, which draws even those who may contest quantitative metrics into valuing them. Within a culture focused on demonstrating results, this produces what I term "the paradox of quantified utility"; consequently, knowledge production increasingly tightens around quantitative ways of knowing.

GENDERING DEVELOPMENT ORGANIZATIONS

Nongovernmental organizations were once considered altruistic, values-driven organizations operating outside of market logic (DiMaggio and Anheier 1990). These beliefs derived from the groups' socially oriented goals and nonmarket funding sources (mostly grants, donations, and government contracts). Indeed, development organizations endure persistent funding instability, which alters their everyday work activities. Large bilateral and multilateral funding typically occurs through contracting—a chain of donors, brokers, and implementing organizations (Lewis and Mosse 2006) in which donors outline their developmental aspirations and contract development organizations to *implement* them. That is, donors are *consumers* (Krause 2014). Because funding is often sourced from taxpayers, politicians (and associated donor nations) are constantly concerned that lower-level implementing organizations are not completing the expected work and seek evidence capable of casting their financial stewardship in a positive light (Best 2017; Watkins, Swidler, and Hannan 2012). This peculiar funding source creates strong incentives for donors to mandate evaluation systems to monitor progress. These performance metrics

become a key mode of communication between donors and the organizations they hire to implement development.

More recent work, however, has demonstrated that development organizations—both for-profit and nonprofit entities—operate within a shared social space of rules and, along with their donors, share broad similarities with other organizations (Barman 2016; Krause 2014; Watkins, Swidler, and Hannan 2012). In particular, they are gendered. Gender is omnipresent, operating through a "gender substructure" (Acker 1990), yet often unacknowledged until it is directly relevant to the workplace circumstance or the people interacting differ in sex category (Ridgeway 1997; Ridgeway and Correll 2004). Gendered hierarchies are reinforced and entrenched through mundane, often textual, organizational processes, such as the division of labor, wages, performance evaluations, and job descriptions (Acker 2002, 2006; Martin 2003; Ridgeway 1997, 2009). Moreover, intersecting systemic oppressions of race, class, and sexualities, among others, create inequality regimes (Acker 2006). One example of how workplace structure may coproduce interactional experiences (Ridgeway 2009) is the differential experiences of "tokens." Employees may be relative minorities in their workplace (Kanter 1977); women may be numerically scarce in gender-inappropriate occupations (Yoder 1991); or tokenization may be mitigated by accessing the cultural resources of dominant groups (Turco 2010). The powerful emotive residuals these and other experiences produce for individuals are difficult to capture (Martin 2003). This scholarship offers valuable tools to analyze the often-invisible gendered dimensions of workplaces.

A small body of work confirms that gender influences what occurs *inside* development organizations, yet this typically focuses on international discourse at

conferences and donor agendas (Ferguson 2015; Merry 2016). Others document how race, sexuality, and citizenship affect the development workplace, which may be understood as a contemporary manifestation of colonial relationships (Baines 2010). As a policy, gender mainstreaming within the workplace has been stymied by gendered power differentials and a failure to challenge dominant structures within organizations (Benschop and Verloo 2006; van Eerdewijk 2014). There is much room for gendered organizations theory to enhance analyses of development workplace dynamics, and specifically the field offices where the labor of gender advisors is understood as the implementation of gender mainstreaming.

REACTIVITY TO INSTITUTIONALIZED MEASUREMENT

Development organizations, like other organizations, are invested in measurement practices. Metrics are crucial to a variety of organizational processes: rationalization (Weber 1978), standardization (Timmermans and Epstein 2010), and political governance (Scott 1998), to name but a few. Understanding neoliberalism as a broad term, I draw attention here to the role of metrics in the devolution of responsibility onto individuals rather than organizations (Connell, Fawcett, and Meagher 2009). Metrics play a role in this process because of *reactivity*—the idea that people alter their actions once they know they are being measured, evaluated, or observed (Espeland and Sauder 2007). Reactivity, then, is valued by managers and other stakeholders looking to mold worker behavior (Miller 2001; Power 1997), yet may also result in worker gaming strategies and "teaching to the test" (Lipsky 2010). Scholarship that explores the relationship between gendered organizations and metrics is nascent: Levy (2016), for example, demonstrates in a study of

truck drivers that hypermasculine gender identity prompts resistance to metrics, and Van Oort (2018) coins "the emotional labor of surveillance" to capture how fast-fashion store cashiers, marginalized along multiple dimensions, absorb the anxieties of customer-time-to-payment metrics. The stress of long lines shifts from managers who must hire more employees to employees who must work faster. Acker (2006, 181) notes that "struggles for power and control are often struggles over bureaucratic tools." This repositioning focuses attention on how performance metrics, as bureaucratic tools, may be wielded in organizational contexts and with what gendered effects.

The increased use of performance metrics to alter worker and organizational trajectories has fostered scholarship about governance by indicators. This literature focuses on how the production and communication of numbers are leveraged for the expression of power (Espeland and Stevens 2008). Beginning in the 1980s, financial accounting practices extended into other sectors, with funding decisions increasingly determined by demonstrations of performance (Strathern 2000). In this new global "audit culture," metrics may become technologies of governance when they name and define the social world, simplify what they measure, and are institutionalized to evaluate performance and inform decision making (K. Davis et al. 2012; Merry 2011). The organizational effects are most visible when a single evaluation regime is present (Lamont 2012), such as that of the *U.S. News & World Report* higher education rankings or the No Child Left Behind rating system (A. Davis 2018; Colyvas 2012). Considering the global shift toward institutionalizing performance metrics across sectors, scholarship must attend to whether, how, and why measurement practices influence organizational and professional decision making.

In the development sector, "audit cultures" have manifested in administrative reconfigurations to meet "results" demands (Dar and Cooke 2008; Eyben et al. 2015; Natsios 2010). Donor mandates have created what Merry (2016) calls "evidence-based governance" and Best (2017) terms "measurement-driven governance." These ideas are enshrined in the highest levels of development discourse and planning, from the Organisation for Economic Co-operation and Development's Paris Agreement on Aid Effectiveness—where quantified results are classified as one of five pillars for development (Best 2017)—to the adoption of goals, indicators, and targets in the Millennium and Sustainable Development Goals (Fukuda-Parr, Greenstein, and Stewart 2013). At the meso level, development organizations risk budget cuts unless they quantify results for donors and politicians (Rottenburg et al. 2015; Watkins et al. 2012). A culture of vigilance around accountability, transparency, and performance metrics now permeates the development sector (Eyben et al. 2015), laying the institutional and normative context for indicators to proliferate and offering a meaningful case of performance metrics in gendered organizations.

METHODS

To illustrate how gendered organizations, social indicators, and professionals intertwine, I use the occupational vantage point of gender advisors serving organizations involved in a single-country instance of a 20-country bilateral agricultural development initiative, here called Farming Sustenance. Informed by my lived experiences as a development practitioner and seeking to understand how everyday development happenings are shaped by the rising "results agenda," I completed 60 interviews with

professionals serving offices across the evaluation system of Farming Sustenance. This study is based on the perspectives of 19 gender advisors, contextualized by those of 41 of their coworkers (project leaders, evaluators, and donor representatives), participant observation at gender-related meetings, and document analysis of project evaluation artifacts. Of the 19 gender advisors, 8 work in the donor country and 11 in the recipient East African country. Of these 11, 5 serve in supervisory or consulting positions and 6 are frontline gender advisors who are responsible for gender mainstreaming across a 5-year portfolio worth approximately $260 million.

These interviews represent nearly full coverage of the initiative's core projects,[4] including gender advisors (all local staff), their managers (all expatriate staff), and gender experts and donor staff (local and expatriate staff). Interviews were 1 to 2 hours, semistructured, and took place in English—the working language of these projects. Verbal consent was obtained, and recordings were made when individuals consented. Using ATLAS.ti, I began analysis by grouping data by office location and job title, and was struck by how gender advisors consistently voiced frustration, foregrounding gender as the most important category in their work life. Understanding the development office as gendered, I then coded transcripts for how professionals value performance metrics. To access the structural nature of how gender advisors are, in turn, valued by the institutionalized, donor-mandated evaluation system, I analyzed their narratives in relation to evaluation artifacts. Viewing discrepancies in narratives as demonstrable of power contestation and therefore analytically valuable (Hesse-Biber 2012), I privileged and contextualized the standpoint of

[4] The exception was one project that had concluded, and former staff were unreachable.

gender advisors. I use pseudonyms for individuals to protect their anonymity[5] and to ensure analytical focus on the interactions between gendered organizations and evaluation practices. Although the data are limited to a single context—with its particularized and variegated gender dynamics—the managerial techniques, bureaucratic policies, and employment roles at the heart of this research are widely practiced around the globe.

GENDER ADVISORS' EVERYDAY WORKLIFE

Ridgeway (2009) encourages a multilevel analysis of gendered organizations that places individuals and organizational structures into relation. I first synthesize interview data and evaluation artifacts to build a structural account of the gender advisor occupation, and then use this organizational context to make sense of their interactional experiences. By employing gendered organizations as an analytical framework, the gendered nature of the development workplace becomes visible.

The Structural Subordination of the Gender Advisor

Although gender advisors are hardly a monolithic group (ranging in age and marital and parenthood status, they are educated through degrees and/or work experience and wear everything from brown flats to pink stilettos), they share a structural location within large-scale development projects. Official development discourse encourages strong investment in women at the site of the project, yet when these intentions are refracted through development organizations, gender-related work becomes tenuously positioned. A woman

[5] For discussion regarding the importance of anonymity when researching evaluation practices, see Eyben et al. (2015), Mosse (2006), and Rottenburg (2009).

manager pondered "How can we deliver women's empowerment when it's not in our offices?" In partial response, I show that the structural subordination of gender advisors results from low numbers of women professional staff, the tokenization of the gender advisor role, the placement of gender metrics at the lowest level of the monitoring system, and team orientation to performance metrics.

In each of the field offices of Farming Sustenance, gender advisors are uniformly women and local nationals. Among the local project staff, men are technical staff with knowledge in agricultural development—for example, livestock fattening or market systems—whereas women serve as the singular gender advisor. In contrast to tokens in the United States (Kanter 1977; Yoder 1991), women here entered the development sector in a role not previously held by men *while* also being a numerical minority. Further, inequality regimes at the intersections of citizenship and gender are present: Although each office has at least one other woman employed in a non-gender–related professional position (technical agricultural staff, upper management, or titled gender expert),[6] these women are expatriates. The only man engaged in gender-related tasks holds the title consultant and the primary focus of his work is not gender-related. Taken together, the gender advisor role is both visibly gender-typed, with the women serving these roles as numerical tokens, and with a tokenized responsibility to address gendered power relations at the project site.

The tokenization of the gender advisor role is created through the reactions of the technical agricultural staff to the task of gender mainstreaming. Project leaders confirmed that most staff do not consider gender dynamics, with one project lead saying this results in "some funny things like groups of men being trained by male trainers about improved

[6] This article is focused solely on technical project staff. Women play crucial yet gender-typed roles in development offices as secretarial and cleaning staff.

breastfeeding practices." Conceptually, gender mainstreaming is a collegial team effort, with the gender advisor working as a facilitator among coworkers similarly committed to mainstreaming. Yet in practice, "gender work" is offloaded onto gender advisors by coworkers who do not incorporate a gender perspective in their own technical work. Hamdiya, a frontline gender advisor who recently joined the project, described attending a workshop where the nutrition technical team had written scripts for a theater drama about improved child-feeding practices. She lamented their reproduction of gender norms, saying "The developed scripts were promoting a strong husband who shouts at the household and has a traditional way of household management. Even he has two or three wives. Why would the project promote polygamy?" As her job requires, Hamdiya requested that the all-men technical team rewrite the scripts to dismantle, rather than reproduce, gender norms. Through scenarios like this, gender advisors are tasked with disrupting the men-dominated space of the development field office. Several told me that simply hiring a gender advisor like themselves generates resistance to incorporating gender into the project activities, to the point that one project leader refused to hire a gender advisor. In this context, women gender advisors are perceived as vanguards for gender, as they singularly carry out their occupational tasks by correcting and advocating for the inclusion of gender to reluctant, predominantly men coworkers.

Gender advisors face resistance from their coworkers, partly because of how gender mainstreaming is incorporated into the measurement practices of development projects. These practices matter because the donors may stop disbursing funds if they become unsatisfied with an implementing organization, thus incentivizing strong, consistent performance on metrics. In other words, performance metrics have been institutionalized

and budget allocation is made based upon "results." This institutionalization occurs in two forms: The evaluation of gender mainstreaming exists at the lowest level of the monitoring system, and staff teams are configured toward the accomplishment of nongendered project objectives.

First, gender mainstreaming is a complex process meant to bring a gendered lens to all aspects of a project, yet it is often only tracked through sex-disaggregated data and the achievement of a target for women's involvement. For example, an indicator on an education project may be "the enrollment of students" with a target of 30 percent girls. Although indicator targets are used to track progress toward the overall project objective, they may also be established for any level of disaggregation—typically breaking down counts of project participants by age and sex. This latter disaggregation is typically used to ensure women's involvement in project activities, reducing *gender* to a simplistic understanding of woman or man. In a context firmly committed to measurement practices and a discursive focus on reaching smallholder farmer women, one would expect to see gender present in the monitoring system; yet Table 3-1 illustrates how gender is actually obscured.

Table 3-1 Example of Gender's Disappearance within Performance Metrics

Metric	*Project example*	
Core Project Objective	Reduce poverty by improving the productivity of value chains	
Programmatic intermediate result	Enhance value chains of selected crops	
Intermediate result	Enhanced [crop] value chain	
Indicators	Value of sales from farms	Number of individuals who received productivity training
Indicator disaggregation	Not applicable (monetary)	Sex-disaggregated targets

Development staff further report that achieving the agricultural-related indicator targets takes precedence over achieving sex-disaggregated targets. One project lead, voicing frustration with the donor's emphasis on results and given women's current levels of literacy and numeracy, regretted that he could not "slow down" to better reach women and hit sex-disaggregated targets, noting that it is "more important" to achieve the main project results. Despite the conceptual underpinning of gender mainstreaming as gendering all aspects of a project, it is tracked only at the lowest level of these measurement practices, allowing "more important" components to take precedence.

The second way gender mainstreaming, and thus gender advisors, is precariously positioned occurs through staffing configurations by *performance metric*. Although it is common in bureaucracies for staff to be divided by work task, spurred by contractual obligations to donors to deliver results, development organizations often arrange their staff by "intermediate result" (commonly referred to as "IR"). Teams are designated as "IR1 team," "IR2 team," and so on, and progress reports to the donor are compartmentalized by IR. Using this lexicon normalizes the accomplishment of results as the primary goal: The work task *is* to achieve results, not the substance it is meant to measure. In contrast, gender advisors are not aligned to a single IR but must "mainstream across," measured solely by hitting sex-disaggregated targets. Elaine, who is local staff and advises multiple projects, explained how their role is positioned outside of the core project:

> We have to show our coworkers how gender is important to reach *their* objective. Usually projects are designed to meet some other goal, like increase agricultural productivity, not to increase women's standing.

> [Increase] nutrition outcome—sure, but to improve nutrition you must improve the status of women. It's not enough to address gender through other [objectives], but people aren't committed enough. [Emphasis added.]

Thus, gender advisors face a structural conflict: They need to work alongside the IR teams to *programmatically* mainstream gender, but the IR team does not inherently need to incorporate gender—they must accomplish the IR's associated indicators, not the sex-disaggregated targets.

In addition to gender advisors being numerical minorities in a newly-created and gender-typed role, the work of gender mainstreaming is buried at the lowest level and positioned relative to the core project objectives in ways easily dismissed as peripheral. These practices intertwine to compound the marginalization of "gender work" in the development workplace. This structural location affects how gender advisors interact with and are treated by their coworkers, eliciting emotional fatigue in gender advisors.

The Emotional Fatigue from Workplace Interactions

Gender advisors are required to campaign for the inclusion of gender in the project through overtly gendered interactions in workplaces in which they are *structurally subordinated*. In other words, structural subordination and interactional resistance coproduce emotional fatigue in gender advisors, which creates a differential work experience for them.

Despite official policies, gender advisors and project leaders describe efforts to incorporate gender as piecemeal and driven by individuals working "against the grain." In

contrast, most technical IR staff report positive engagement in gender efforts—albeit bureaucratically (e.g., "We have sex-disaggregated targets and target accordingly")—while acknowledging gender's relative absence from projects. None recognized themselves as a potential barrier. Elaine explained:

> I have to plan to meet people and assist them to better integrate gender. How can I help them? How can I convince them? You have to be a professional. I hate my job. Really. I have to analyze and interpret, I have to use diplomacy, advocacy, knowledge, and expertise to convince people.

If Elaine's coworkers were truly participating in gender mainstreaming, "convincing" them would be unnecessary. This interactional resistance breeds frustration, resulting in Elaine hating her job as *a de facto* vanguard for gender. Thus, the gender advisor must constantly muster her skill set to garner coworker support.

Such feelings were persistently present as the gender advisors described their job tasks. Although gender mainstreaming is a goal for the *project*, gender advisors describe being "exhausted" by engaging their coworkers in the *office*. An expatriate gender expert, Kathee, spoke of such gendered fatigue, saying:

> I have to come to terms with how my coworkers think and take them on a journey to a new way of learning. And that is an exhausting process [she laughs] and there's not a lot of stamina or support for gender advisors on how we do that, on how we communicate, on how we have to stand up and

> be ridiculed as step number one. And how we deal with that and how we move past that. And how we constantly *feel* like a nag, and then we're *told* we're a nag. And then we're buying into gender stereotypes again because generally we're female. [Emphasis added.]

Gender advisors "feel like a nag," partly because they are positioned outside the core project structure and required to correct their coworkers—typically after they have completed planning, design, or even partial implementation. Not only are interactions substantively about gender, they are also about gendered power dynamics: It is hard to imagine women calling men "nags" in the workplace.

At times, these dynamics manifest as intimate, overtly-gendered interactions between coworkers. Two gender advisors, working in separate offices, independently mentioned that men coworkers commented they were annoyed by the noise from women's heeled shoes. In the first instance, both people were expatriates and, in the second, both were locals, foregrounding gender. The expatriate gender advisor said she was walking down the hallway, when another expatriate man ran out of his office, yelling "It's you! You're the one! Your heels—clack-clacking down the hallway! I cannot even concentrate!" Another example of highly personal interactions includes a local gender advisor who recounted a conversation with a local donor representative. He segued from the women's empowerment agenda of his employer by saying to her "If you ladies keep on talking like this, you'll be divorced." The donor representative generalized from her to all gender advisors through the use of "you ladies" and then portrayed the donor goal of women's empowerment as undermining the valued social institution of marriage. She

rebuffed his statement, but contextualized the story as demonstrative of men's resistance in the workplace.

In sum, the common practice in development projects of having one gender advisor who must "get gender in" to nongendered programmatic activities means that gender advisors must interact with their reluctant IR team coworkers. These interactional realities *individualize* the gender advisors' structural subordination, producing emotional fatigue and a work environment markedly different from that of their men coworkers.

GENDERED VALUATIONS OF MEASUREMENT PRACTICES

Gender advisors face marginalization partly because of measurement practices that place gender mainstreaming at the lowest level, as sex-disaggregated indicator targets, whereas other staff focus on accomplishing the higher-level intermediate result indicators. Given this circumstance, how do gender advisors interpret the value of sex-disaggregated data? Most gender advisors noted that numbers do not provide insight into what "really" happens in women's lives as a result of development programming: "We can say '20 percent women' but it doesn't show how we can transform their lives unless we put a qualitative indicator." All the gender advisors recognized that quantitative data provide insight into the prevalence of a given issue, but, like Uwimana, a former frontline gender advisor turned consultant, believed that qualitative data provide a better picture of project efficacy. Uwimana said:

> Qualitative [data] really tells you women's position: access to resources, how women really *own* it… If you just have numbers, you'll see the 40

> percent [target is reached], but you have to look beyond the numbers to see the *real* situation.

Uwimana understands qualitative data as more "real" and worries that quantitative indicators may be misleading. Chungwa, a former frontline gender advisor and current expert overseeing national agricultural policies, was frustrated that her office could not hire a consultant to qualitatively investigate the reasons for women's lower project participation rates. She said:

> We couldn't find a good gender researcher apart from "sex-disaggregated data." Gender has different faces. If you go to the agricultural technology people, they say, "Women are not there, women don't use extension." *Why is that*? Because of culture. Okay, what kind of culture? It's religion. When you go to religion, what kind of religion is that? Why is the religion-biased? So you want to talk to people. [Emphasis added].

In Chungwa's interpretation, qualitative methods can trace backward from socially embedded *effect* to *cause*—an insight that sex-disaggregated data do not provide. Meti, a young frontline gender advisor recently trained at the master's level in research methods, was skeptical of quantitative methods. She said:

> I believe that quantitative methodologies are predetermined. Maybe it's the data collector, maybe the questions are leading, and the women just answer

> how they think they should. Qualitative is best, with in-depth interviews we can capture the real change of the women. It'd be better, the reliability, the validity. By using quantitative we may not get the real experience, it's not enough on its own, it needs to be triangulated with qualitative. [With quantitative data] we can't get reliable info.

This is a surprising reversal from dominant understandings of qualitative as more interpretive: By invoking "reliability" and "validity," Meti describes qualitative data in words typically reserved for quantitative data. Meti, like her counterparts, views qualitative data as better able to accurately capture women's lives in all their intersectional complexities.

Despite this skepticism, gender advisors voice a simultaneous desire *for* quantitative metrics. One advisor, discussing large-scale surveys, remarked that "It's statistical, it's scientific. I appreciate that we get some number." Under the logic established by these gender advisors regarding qualitative data, we would not expect them to also request quantitative measures. So why does this happen? I demonstrate that their desire is tied to their understanding of quantified metrics not as meaningful measurements of the women they seek to assist in the project, but as tools for gaining power in negotiations with coworkers in the office.

Bureaucratic tools: The appeal of quantified metrics

The appeal of quantified metrics to gender advisors becomes clearer once viewed through the lens of gendered organizations. These workplaces are gendered in ways that

affect the value of performance metrics, existing and imagined, to gender advisors. Although quantitative indicators are often interpreted as useful *externally*, gender advisors understand their utility *internally* for potential workplace changes. Gender advisors are interested in quantitative indicators to, first, convince coworkers who are dismissive of gender mainstreaming as a project imperative, and, second, to force their coworkers' compliance with gender mainstreaming.

Gender advisors believe that numbers appeal to their coworkers as "evidence" and, as such, value numbers as a tool for garnering support in an otherwise hostile workplace. Elaine explained:

> I have to convince people why it's important to include gender. They don't want to [include gender], but it must be there. We must have data or evidence, otherwise how can we get people to work on this? I must be equipped enough to convince them.

Her use of "enough" implies a search for the minimum standard of evidence needed to engage coworkers in gender mainstreaming. Chungwa adds nuance by citing the numerical dominance of men in agriculture and their understanding of what constitutes "evidence." She said:

> Most of the staff in agriculture are men and say, "Women are not involved—they want to be in [micro]finance, they want to be in the house." That attitude is there, so the women are not trained. The staff want to have

> an evidence-base—"Why do we need gender and technology? Technologies are neutral, why do you say 'female friendly technologies'?"…You need to have an evidence-base that says, "because of this and that," "because it decreases women's time from six hours to you know." Numbers, research, evidence—they [the staff] look for that.

Chungwa continues that gender-related outcomes are unknowable in advance, but that her coworkers are not comfortable operating without an "evidence base." Rosalie, a frontline gender advisor, highlights the importance of evidence to prove the worth of the gender advisor position, noting that qualitative descriptions of progress are undermined by accusations of limited representativeness or generalizability. She said:

> The reason I need quantitative thing in general is to say, "Ok this is our evidence." I don't want anyone to later say, "Yeah qualitatively some have progressed and some have not, so where is your result?" I want to be able to say, "Okay because of this intervention, it has moved from this to this much percentage" or whatever.

Rosalie needs numbers to proactively combat critiques about the effectiveness of her job, yet she went on to describe her preferred method of capturing project efficacy as diaries written by women participants. This suggests her interest in percentages is solely a *reaction* to coworker expectations of data: defined by them as quantitative. Her perceptions were confirmed by project leaders, one of whom bluntly stated "I don't want qualitative, it's too anecdotal."

These gender advisors perceive their coworkers as associating numbers with "evidence." Although the advisors desire qualitative data, they are drawn into a quantitative logic in an effort to "convince" their coworkers of the value and impact of engaging women. The imperative to convince is magnified by the structural location of gender mainstreaming; gender advisors have no choice but to meet the knowledge demands of their coworkers because gender mainstreaming is outside the core project and only successfully incorporated if coworkers are first convinced and then cooperate in mainstreaming gender through each project component.

Gender advisors, structurally marginalized by measurement practices and emotionally fatigued, search for expeditious strategies to garner support and cooperation. Bernadette, a former donor-based gender expert explains, "Gender is an integral part of the project, but lots of times it's language and not reality. It's on paper, but it doesn't happen. When we put down numbers, it's to force that into reality." When gender advisors discuss their desire for indicators, they present them not as meaningful measurement tools about the women development projects are meant to engage, but as *compliance mechanisms* for their coworkers. Gender advisors switch their sentence structure to position metrics as subjects capable of performing action: "numbers force" attention to gender, "targets hold people" accountable regardless of their attitude toward gender mainstreaming, and "indicators force" the inclusion of women. In other words, metrics are recast as neoliberal managerial tools meant to induce behavior toward the desired aim of gender mainstreaming.

Gender advisors' desire for more quantitative indicators, now understood as managerial, demonstrates a struggle for power and control is underway. Recasting

quantitative metrics as compliance mechanisms brings their power as potential tools for the gender advisors into focus, as Uwimana described:

> Always gender is considered an added-on responsibility. What is on paper holds people accountable. Some don't believe in targeting women. You hear that sometimes and it shows people's attitude. Even if they have excuses, targets hold people [responsible].

In Uwimana's formulation, indicator targets override coworkers' unsupportive attitudes about gender as an integral project element. This attitude ultimately leads to poor inclusion of women project participants in IR-specific programming, and sex-disaggregated targets then become a solution. Hamdiya illustrated the ability of targets to require IR teams to include women:

> The gender advisor before me, she faced a challenge: people in IR4 said, "We don't have indicators on gender, so nobody is going to measure us." It was a challenge to push, pushing and forcing them to focus.… In my observation it is only the indicator [targets] that forced them to include women.

This perspective confirms the ability of targets as tools to mandate staff attention to gender, while also highlighting that IR teams may use measurement practices to justify *not* working on gender mainstreaming. In doing so, they substantiate the structural marginalization of

measuring gender as sex-disaggregated targets. Still, gender advisors find that even the few, existing low-level indicator targets (see Table 1) help them make the case for engaging women. Because the IR team tries to dodge working on gender under the logic of "no measurement" (for them), this leads gender advisors to believe that *if* there were gender mainstreaming *indicators*, IR teams would comply. Rosalie agrees and advocates for developing such indicators. She said:

> We often say, "Gender is crosscutting, it doesn't need indicators." No! Crosscutting doesn't mean it doesn't require indicators—you need to have them! You need to have an indicator in there! Measure it then! Then they won't be forgotten!

By advocating for the measurement of gender at the indicator level (rather than as sex-disaggregated targets), Rosalie imagines she can leverage such numbers to better ensure that attention and resources are allocated to gender mainstreaming. Yet gender advisors' request for *more* higher-level indicators means they end up playing by the rules of a quantitative knowledge production game—one they do not actually want to play.

This is not to argue that strengthening evaluation systems is the solution, but rather to show that gender advisors strategically assess the tools available to help get their jobs done and see indicators as the most expeditious option. And with good reason: Every day, gender advisors observe coworkers and a project, organization, and donor oriented to performance metrics. The common mantra "what gets measured, gets done" is their lived reality. Thus, they imagine that greater inclusion in performance metrics is the best

solution, not for creating knowledge about how to enrich women's lives—their true desire—but for their pragmatic utility to overcome the emotional fatigue and structural marginalization they endure in the development workplace. From the gender advisor's perspective, quantified metrics are hard and fast bureaucratic tools that shift sole responsibility for "gender-inclusive development" away from the gender advisor to a shared responsibility with the technical IR teams. In other words, gender advisors become interested in trying to access and harness the reactive power of indicators as a manager would do. A gendered organizations lens draws our attention toward the contested workplaces in which gender advisors operate and reveals their strategic use of metrics to advance their own aims.

CONCLUSION

Although this case study speaks to an East African development landscape, the presence of quantitative indicators mandated by international development projects is global in reach. As I have shown here, development organizations are gendered, which leaves women gender advisors sidelined, struggling, and seeking ways to advance their goal of gender mainstreaming. In a bid to advance this agenda, gender advisors repurpose quantitative measures to assert more power in a gendered work context. This empirical case results in two theoretical contributions.

The first theoretical finding of this research is that gendered organizations theory is foundational to analyzing how professionals respond to the growing influence of performance metrics in the workplace. Much of the quantification scholarship views metrics as "engines of anxiety" (Espeland and Sauder 2016), yet these gender advisors

imagine performance metrics as potential mechanisms of *relief* through the devolution of responsibility. The gender advisors' assumption is that, if better included, gender-related metrics would devolve responsibility away from themselves to a collective staff responsibility, inspiring coworker compliance rather than complaint and dispute. In contrast to Levy (2016), who finds that masculinist identities prompt truckers' *resistance* to quantified performance metrics, I find that the emotional labor of women gender advisors, marginalized structurally and interactionally, prompts *endorsement* of metrics. Levy (2016) finds that gendered notions of self animate resistance to metrics, yet I argue that gendered work experiences can transform the gender advisors' self-preference for qualitative measures into a pragmatic advancement of quantitative metrics. Both truckers and gender advisors seek ways to optimize their feeling of control over their work, and these individual preferences interact with gendered organizational components to produce potentially opposite outcomes.

Incorporating gendered organizations as an analytical component is therefore critical to scholarship, especially as workplaces become increasingly quantified, digitized, and surveilled. Acker (2002) encourages research on the devolution of responsibility under globalization. Performance metrics offer an important, yet mundane pathway for such devolution. These gender advisors are not merely "acted upon" by globalization; they further the internationalization of quantified metrics by expressing agency in conditions not entirely of their making. Therefore, future research on professionals' reception of and reaction to performance metrics must account for intersectionality, inequality regimes, and/or gendered organizations.

From a knowledge standpoint, gender advisors want thick description about the "real change" in women's lives elicited by development projects but are caught up in a gendered organization that institutionalizes achieving quantitative targets for women's participation. Building on the feminist debate about quantitative and qualitative data (DeVault 1996), I show that attention should be less about whether one form is better at capturing women's lives, but about the interactions, practices, policies, and organizational relationships that encourage and value particular forms of knowledge over others. I demonstrate how the gendered nature of organizations can change potential resistors of quantified metrics into promoters because of their pragmatic utility.

This leads to my second theoretical contribution: When gendered organizations institutionalize quantified performance metrics, knowledge production shifts to match the context. In industries caught up within calls for "evidence-based" decision making, one would expect organizations to be interested in whether quantitative or qualitative data best capture the "real" situation of women and gender relations to inform efficacious programming, policies, training, marketing, and more. Yet this dismisses the reactive power of quantitative metrics to alter organizational behaviors: Development organizations instead assess the utility of data to communicate results to their funders. This case has implications for organizations more broadly, in which quantified indicators are used to manage from a distance, devolve responsibility, and externalize risk. Numbers remove context, quieting probing questions, and enable aggregation, standardization, and comparison among diverse geographies and cultures across time. In doing so, numbers gain authority as masculine: objective, trustworthy, and suitable as evidence for decision

making. Numbers are not necessarily superior measures, but are *perceived* as better, which increases their utility to organizations.

In a context where the allocation of funding requires demonstrable and uncontestable results, the value of quantified metrics is increased and reproduced. I coin the term "the paradox of quantified utility" to describe how, in contexts of institutionalized performance metrics, professionals re-orient to the metrics, not necessarily because they want to participate in this way, but because they are (or imagine they will be) *rewarded* for engaging in quantification and quantified knowledge production even at the cost of what they may recognize as the *substance* of their work. This paradox is likely felt most by marginalized people, such as gender advisors, and communities who engage in a betrayal of their own preferences to try to garner attention and resources. Yet given the very real material advantages that engaging in quantification may provide, their actions are strategic, albeit constrained.

In contexts that institutionalize "results," resources increasingly flow to easily quantifiable outputs or processes, and knowledge production becomes recursive. The pragmatic utility of quantitative metrics for organizational survival outweighs rich knowledge production. Quantitative indicators ignore diversities in gender relations, religions, sexualities, reproduction, and productive labor, among other areas—resulting in knowledge useful for policy makers who must demonstrate return on investment. Ultimately, development organizations, by relying on sex-disaggregated indicator data to capture "gender," produce knowledge that presents women as individuals to be acted upon, rather than as relational and dynamic community members grappling with gender norms as they respond to urbanization, migration, climate change, and capitalist development,

among others. I demonstrate how the pragmatic utility of performance metrics as bureaucratic tools recruits professionals and organizations into quantified knowledge production in ways that, paradoxically, may impede knowledge creation about the structural forces embedded in women's everyday lives around the globe—knowledge that may actually assist organizations to engage with women *on their own terms*. The importance of what constitutes knowledge production under the "results agenda," its relative salience, and source of value is a key theoretical finding of this research.

Practically, attention to organizational policies and pressures that steer knowledge production down particular paths is paramount. Professionals in development bureaucracies at the donor level have argued for engaging women because they produce a higher return on investment (i.e., the "business case"). My research suggests this action is likely explained by the utility of the business case to garner budget allocation in a gendered organization. Yet it also suggests a potential (partial) way out of the paradox: Feminists, practitioners, and scholars should labor to showcase the utility of qualitative data to provide rich insights for organizational decision making. As Irene noted in the opening epigraph, quantitative knowledge is biased against Global South knowledge. Grassroots organizations' ability to deliver meaningful and sustainable change in people's lives—substantive results—is often informed by knowledge that is situated, deeply contextual, and not standardizable. A movement to showcase the utility of qualitative ways of knowing would aid grassroots organizations to maintain their competitive advantage within the results landscape. Although this imagined solution may create more space for different ways of knowing more immediately, it still churns within a logic of utility, begging the question: At what cost in the long run? Conceptually, scholars must remain vigilant about

modes of knowledge production that devalue the relational and intersectional dimensions of social issues. Thus, gender scholars must rapidly engage the dynamic intersections of quantification and gendered organizations theory to analyze the variegated role of numbers in society and their impact on knowledge production.

Chapter 4

Women are "Counted as Present, but Quiet as Mice": Quantified Evaluation as a Logic in International Development

Abstract: Critical scholars have argued international development projects are unable to deliver on the feminist aim of transforming gender relations, yet to date the role of project measurement systems is underexplored. I argue that evaluation systems are the *logic*, and by extension, the *mechanism*, that reduces feminist intentions in the development sector into unmet promises. I utilize the empirical case of a standardized evaluation system to build a ground-up example of a logic in action and document its effects on gender-related efforts within a development bureaucracy. Institutional logics as an analytical framework draws upon the homogenizing forces of organizational structures while also accommodating for how logics may contend, conflict, hybridize, or dissemble through action. A large bilateral donor created a 20-country agricultural development initiative, with stated goals to empower women. I draw on interviews with sixty development professionals involved in the initiative, including project managers, gender advisors, and evaluation leads, and analysis of evaluation-related documents. I document how gender-related work is relegated to the sidelines, compounded by a lack of interest for incorporating gender by professionals who have clearer, more "valuable" tasks to complete. I argue that the structures and practices that comprise quantified evaluation in an organizational setting work in tandem to marginalize gender mainstreaming.

> It's much more about the numbers themselves. There's a huge difference between quantity and quality and it's very easy to focus on quantity: How many people showed up? How many people joined the cooperative? How many people are trained rather than actually participated? Oh — that's … amazing! Tick! Have focus groups that are half-and-half and the women sit there quiet as mice. But if we got them, the figures are there, everyone's very happy, and you tick the box 'done.' That's the sort of information hidden by the numbers. (Daniel, evaluation consultant)

> Development programs that are most precisely and easily measured are the least transformational, and those programs that are most transformational are the least measurable (Natsios 2010: 3)

As articulated by these quotes, monitoring and evaluation systems in international development compress diverse cultural realities into standardized formats, raising questions about the potential of measurable development programming to be transformational—a stated goal of many development initiatives. Monitoring and evaluation (M&E) systems produce numbers that travel space, time, culture, and epistemological perspectives. A wide array of scholars and practitioners (Eyben, Guijt, Roche, Shutt 2015; Natsios 2010; Hirschmann 2002) alike are concerned by what may happen as a result of performance metrics when they are applied to socially transformative

goals, such as women's empowerment, leadership, and democracy. Envisioning quantified evaluation as a logic allows us to consider its projection into new settings, its reception, and how it is embedded in organizational structures, norms, operating procedures, activities—ultimately congealing into coherent taken-for-granted practices that embody a particular cultural orientation to "doing" development. By articulating evaluation as a logic, we can see how women's empowerment fares under such a logic. I argue that efforts for gender equality in development programming and women's empowerment is consistently disadvantaged by quantified evaluation as a logic.

Bureaucratic donor policies export quantified evaluation to diverse locations around the world through a network of donor offices and those of the organizations contracted to implement development projects. Professionals from donor and recipient countries intermix and animate each office, inherently localizing and hybridizing logics. Yet the power of quantified evaluation is amplified by the bureaucratic policies and procedures that accompany it, resulting in an organizational contracting structure that forces compliance, disregarding unique local circumstances. A system of evaluative structures and practices reinforce quantified evaluation as the status quo of international development, driven by the threat of funding loss, despite a lack of shared meaning amongst professionals about the meaning and value of such practices. In this chapter, I explore the interlocking elements of quantified evaluation as a logic and document how gender-related efforts are sidelined, by what mechanisms, and with what effects.

While anthropologists have begun critical inquiry into the impacts and proliferating effects of performance metrics (Strathern 2000; Davis, Fisher, Kingsbury, Merry 2012; Rottenburg, Merry, Park, Mugler 2015), a growing body of sociological work focuses

explicitly on quantification (Espeland and Sauder 2016; Berman and Hirschman 2016; Hirschman, Berrey, Rose-Greenland 2016; Davis 2018). Feminist scholars have long raised questions about the ability of bureaucracies to promote transformational feminist aims (Staudt 1985; Staudt 1997). More recently, gender scholarship has criticized "smart economics" discourse—drawing attention to how the business case rationale for development investment has used econometric analyses to make women go to work for development, instead of ensuring development works *for* women (Collins 2016; Chant 2016; Chant and Sweetman 2012; Cornwall and Rivas 2015). This research centers the role of the evaluation system and quantitative indicators as the key *mechanism* by which feminist development aims are undermined through bureaucracy.

Taking Espeland and Stevens (2008) "ethics of quantification" seriously, I demonstrate that we can no longer understand pathways toward equality without understanding the role of numbers. "Gender" as refracted through international development metrics becomes a primary arena in which dynamics of quantification and their effects occur. To understand efforts for gender equality in today's world, we must first understand metrics and the role they play in development interventions to better understand what spaces are foreclosed in their presence and how professionals respond to and reconstitute the power of metrics through their everyday labor. In this paper, I focus my argument on the bureaucratic requirements and practices of development professionals, documenting how sex-disaggregated indicator targets leave gender advocates with no leverage to enforce gender-related programming. Using the case of gender-related indicators, I demonstrate how the difficulty of rendering hard-to-measure concepts, like gender equality, into quantified evaluation results in a systemic marginalization of gender,

gender advisors, and gender-related programming. I argue that transnational quantified evaluation systems serve as a new form of governing power in the 21st century and are best understood an as an institutional logic which coordinates all professionals' behavior.

Transnational evaluation systems, an industry valued at $120 billion in 2009 (Ika and Lytvynov 2011), measure the 'success' of development projects through a process of downward-moving policies from funders and upward-moving data from the project site, aggregated to show project results for donor dollars. This bureaucratic system is only made possible through the joint effort of diverse development workers—from data collectors in rural areas in aid-recipient countries to evaluation directors in donor countries. The durability of quantified evaluation as an institutional logic becomes legible as analysis moves across these diverse sites. While feminist scholars cite the constant pressure of priorities, measurements, and focus on 'quick fix' solutions (Chant and Sweetman 2012; Collins 2016; Cornwall and Rivas 2015), this scholarship empirically focuses on the politics of international conferences and agenda-promoting discourse rather than the everyday interactions and practices within development organizations.

I utilize gender-related programming in a large agricultural development initiative, here called *Farming Sustenance*, as my case. Women's empowerment is recognized as a "driving force" within the initiative sponsored by a bilateral donor I call International Development Assistance Department (IDAD). The initiative is run in 20 countries around the world as part of a greater push for a second Green Revolution; I focus on an East African instance. Adopting a multi-sited ethnographic approach (Marcus 1995), I 'follow the thing' of performance metrics as they compel people to collect and aggregate data, produce charts and reports, and alter their work lives and identities. This article is based on

data from sixty interviews with gender advisors, project leaders, and evaluation leads who occupy different roles along the transnational evaluation system in both the recipient and donor countries.

In what follows, I first describe the growth of quantified evaluation and then introduce institutional logics as an analytical framework. Empirically, I then detail the structures of quantified evaluation that force the compliance of development professionals, and then detail a series of work practices that further entrench the marginalization of gender in development programming. In each section I demonstrate the *modus operandi* of IDAD and contracted implementing organizations and then show how and why gender is sidelined by such structures and practices. I conclude by highlighting the analytical value of understanding quantified evaluation as an institutional logic for both practitioners and for development scholars. I argue that understanding quantified evaluation as an institutional logic clarifies why and how gender equality is unattainable through the international development industry.

The Growth of Quantified Evaluation as a Logic

Beginning in the 1980s, financial accounting practices extended into other sectors, including education (Colyvas 2012, Davis 2018, Espeland and Sauder 2016), Institutional Review Board offices (Babb 2020), and management consulting (Shore and Wright 2015). The development sector was not immune to such global trends: Dar and Cooke (2008) described how a growing focus on neoliberal bureaucratic planning and measuring was universalized through the development organizational apparatus—a managerial style they

termed "new development management." Concomitant with other reforms, this means that funding decisions are increasingly tied to demonstration of performance (Strathern 2000).

Development practitioners have raised concerns about this new approach, documenting the pernicious effects of quantitative measurement on socially transformative processes, like empowerment, capacity building, and leadership (Eyben et al. 2015, Natsios 2010; Hirschmann 2002). The approach fosters linear understandings of development input and output—the "results agenda"—setting off a variety of mechanisms to ensure compliance. In extreme cases, donors withhold funds until the promised results have been achieved—a management method termed "payment-by-results" (Eyben et al. 2015). As the "results agenda" has gained momentum, what passes as acceptable measurement in the field of international development has shifted, with randomized control trials now considered the gold standard (Viterna and Robertson 2015) and demands for increasingly rigorous evaluation practices have become commonplace (Béné et al. 2017).

Development scholars have responded to the encroachment of quantification by studying it at different levels, including international discourse and target setting processes, organizational uses of evaluation, and the unintended consequences of the emphasis on results. At the international level, scholarship focuses on multilateral target setting and the creation of indices (Davis et al. 2010; Ilcan and Phillips 2010; Kabeer 2005) or the contested politics of United Nations conferences that result in adopting quantification (Fukuda-Parr and Hulme 2011; Leibowitz and Zwingel, 2014). Organizationally, scholarship focuses on the relationships created and reproduced by evaluative practices in the development sector: organizations use assessments of potential success in selecting project sites (Krause 2014), organizations, professionals, and beneficiaries labor together

to produce successful projects that ensure the continuation of mutually beneficial resources (Beck 2017), and through simple counting, overworked consultants, and performances of project efficacy, evaluations "create success" (Watkins et al. 2012). In terms of the unintended consequences of measurement within the development sector, scholarship has documented that the unquantifiable is left out (Merry and Wood 2015), including socially transformative processes that are the least measurable (Natsios 2010; Vallejo and Wehn 2016), or that developmental successes may be mistakenly perceived as failures (Easterly 2009). Furthermore, countries may begin rank-seeking practices to improve their access to international resources (Best 2017; Høyland et al 2012) and there is little acknowledgement that "results" require a level of measurement precision that is unlikely in the complex conditions development projects operate within (Hirschmann 2002). This focus on how evaluation sets macro-level targets, produces success, or creates unintended consequences has left out analysis of what happens inside organizations as professionals go about the work of quantified evaluation. This means that the organizational dynamics driven by evaluative structures and practices in the development sector are left underexplored.

Institutional Logics

"Institutional logics" was a response to the predominance of isomorphism in neo-institutionalism and rational choice theory in the 1980s and built upon recent advances in cognitive psychology and Douglas's (1986) *How Institutions Think* in which she argued that institutions have a hold on our ways of classifying and recognizing (Friedland and Alford 1991). Neo-institutionalism offered insight into how macro structures and culture shape organizations (Meyer and Rowan 1977; Thornton et al. 2012); this led to an

explosion of scholarship that analyzed cross-national datasets demonstrating the isomorphism of institutions and global scripts around the world. Institutional logics expands on this work, seeking to "bring society back in" by placing institutionalized principles in conversation with individual and organizational actions.

"Institutions are supraorganizational patterns of activity through which humans conduct their material life in time and space, and symbolic systems through which they categorize that activity and infuse it with meaning" Friedland and Alford (1991: 232). In other words, institutions are simple conventions that pattern human behavior (Douglas 1986) or patterns of beliefs and rules (Lammers 2011). The power of these institutions, or conventions, can be seen when people go about their work lives in organizations, engaging in action *because of* higher level orders that lie above the individual and the organization. Above the meso-level of organizations, institutional logics operate at the macro-level of societal norms and expectations. In the West, this means individuals and organizations grapple with and navigate the logics of democracy, individualism, the market, the state, and Christianity through their actions. Yet these logics are not universal and other institutional orders exist in other societies. Institutional logics matter because the principles, practices, and symbols associated with each logic outline the shared assumptions and values that determine sound reasoning and legitimacy (Thornton, Casio, and Lounsbury 2012). In other words, institutional logics are organizing principles that enable and constrain action in particular contexts—bringing beliefs, values, and ideas of individuals into dialogue with the material and symbolic concerns of organizations (Thornton et al. 2012). Although logics are observed in practice and discourse, they are hard to classify precisely because they become ingrained in ways of doing, making it

difficult to piece apart ideas, discourse, beliefs, values, and practices. Institutional logics are abstract in nature and exist as shared assumptions about what counts as accepted ways of reasoning and forms of legitimacy.

Institutional logics are ideas that are manifested in the practices in everyday life, occurring through nested layers individuals, interactions, organizations, fields, and society (Friedland and Alford 1991). An institutional logics analysis requires a multi-level approach: individuals themselves cannot have ontological status because they do not exist outside of society. Individual and organizational behaviors must be located in a societal context that makes sense of them. Therefore, analysis should occur at three levels: "individuals competing and negotiating, organizations in conflict and coordination, and institutions in contradiction and interdependency." (Friedland and Alford 1991, 240-241). Perspectives which place emphasis on any one level miss out on the analytical potential of a nested approach. While institutional logics can be studied at a meta-level engaging the central institutions of capitalism, family, religion, etc. (Thornton et al. 2012), I follow other scholars to bring institutional logics into organizations (Hammack and Heydemann 2009; Fejerskov 2017; Berman 2012; Hirschman et al. 2016) and bring a multi-level and multi-layered analytic to the complex circumstance of development contracting and delivery.

A global understanding of institutional logics outlines how Western ideas and practices are able to travel the world through organizational apparatuses built during colonialism and continued through international organizational structures: such as the United Nations, bilateral and multilateral aid and development organizations, and military reconstruction projects. Locating institutional logics in a time and a place mirrors the idea of development not as a truth emanating from the West, but rather "provincialized as a

particular dynamic cultural formation" (Mosse, 2013; 230). Words common in neo-institutionalism, such as "diffusion," normalize understandings that ideas shape practices in a direct and smooth pathway. "Diffusion" connotes a soft and unmanaged flow of ideas, yet ideas cannot diffuse without a receptive audience willing to put such ideas into practice (Heydemann and Hammack 2009). In response, Heydemann and Hammack (2009) suggest "project" as a verb is more appropriate for international development cases and highlights both intention and managerial aspirations. Swidler (2009) demonstrates that aid recipients can strategically appropriate transnational logics into local settings, hybridizing them into new meanings. In other words, reception is a crucial aspect of logics. Institutional logics demonstrates that while practices may be adopted, they are adopted in new ways, interacting with local cultural dynamics that will always create a version that is never the same, but a refracted version.

More recent work has focused on the projection of logics internationally through philanthropy (Heydemann and Hammack 2009), explain conflicting logics (Fejerskov 2017), how logics can gain traction without the presence of a promoting actor (Berman 2012), the "stickiness" of logics (Sauder and Chun 2018), and how logics can be undone (Hirshman et al. 2016). Within organizations, multiple logics can conflict, coexist, blend or hybridize depending upon their compatibility and centrality to the organizational context (Besharov and Smith 2014). Fejerskov (2017) presents the logics of "cost effectiveness" and "gender equality and women's empowerment" within the development sector as conflicting—when the adoption of one undermines the adoption of the other. The logic of "cost effectiveness" driving evidence-based policies and the rationalization of development policy and short-term investment and outcomes. In sharp contrast, the logic

of gender equality and women's empowerment is norms-driven and focused on social transformation and long-term change. This creates a continuous clash between logics. As the logic of cost effectiveness is dominant within the organizational culture, gender equality efforts are always subordinated to the logic of cost effectiveness (Fejerskov 2017). This is in contrast to cross-national research that demonstrates human rights normative frameworks can be leveraged over cost effectiveness, as seen in the case of universal primary education and school fees (Kim and Boyle 2012). Using institutional logics as an analytical frame provides a multi-level analysis that documents how and why some institutions become so powerful and successful alter accepted forms of reasoning and legitimacy, cultivating the actions of individuals and organizations as they go about their everyday lives.

Recognizing that multiple logics operate simultaneously in international development, I explore the logic of quantified evaluation under the "results agenda" in relation to the logic of gender-related programming. Hirschman et al. (2016: 268) noted "organizational attempts to rationalize the *reduction* of inequality are riddled with a tension between standardization (with its implicit logic of merit and neutrality) and redistribution (with its logic of justice and culturally-specific identification)" and that organizations address this tension by normalizing it into their routines of *evaluation*. This perspective invites detailed empirical analysis of evaluation policies and practices. As such, I use institutional logics as an analytical framework to document the current state of gender within large aid bureaucracies, exploring quantified evaluation's potential to deliver gender equality. Buttressed by calls for transparency and accountability, quantified evaluation has emerged as a non-negotiable mandate and delivering and demonstrated results has been

positioned as a moral endeavor. How do individuals and organizations make sense of these imperatives to measure, how do they manifest into organizational structures and practices, and with what effect on gender-related programming?

METHODS

In order to access the logics of the development sector around measurement, I 'followed the thing' (Burawoy et al. 2000; Marcus 1995) of performance metrics as they compel people to collect and aggregate data, produce charts and reports, and alter their work lives and identities. To do this, I completed a multi-sited ethnography of a bilateral donor evaluation system—interviewing the professionals who make evaluation happen beginning from development project offices and traveling "up" the aid chain to headquarters. The interview data presented here are from the professionals who work with and for a large bilateral donor and/or for organizations who are contracted by the donor to implement projects. Note that these "projects" now include both the typical development project—agricultural inputs, trainings, and linkages to markets—*and* the new evaluation-specific project—assistance to development organizations to complete evaluative activities and share knowledge.

Farming Sustenance's evaluation system serves as a meaningful case study of what happens to gender-related efforts in the presence of quantified evaluation practices. The donor stated an interest in inclusive agricultural development under a business case rationale for the inclusion of women: women are less productive farmers than men because of less access to land, credit, seeds, water, and trainings. If women had the same access, the world would benefit because women would increase crop yields by 20-30%. Although

this chapter is empirically based on a single country case study, the indicators and requirements at the heart of this case are mandated and standardized across all twenty countries in the initiative.

While this article presents data from a multi-sited study along an evaluation chain by a single bilateral donor, the phenomenon of evaluation and results-agenda is part of a greater shift bilateral funders have endorsed and promoted, including United States, Germany, United Kingdom, Sweden, and the Netherlands (Eyben et al. 2015). All names of professionals, projects, organizations, and the initiative are pseudonyms and identifying details of projects may have been altered slightly. Scholars who research evaluation have made similar decisions to anonymize or use pseudonyms extensively (Eyben et al. 2015; Rottenburg 2009) or have faced negative reactions (Mosse 2006). By using pseudonyms, attention remains on the cultures of evaluation, processes and practices of recruitment to this culture and its social reproduction, and the scope conditions under which quantified evaluation is established through the everyday labor of diverse professionals.

All participants were contacted due to their employing organization's role in *Farming Sustenance*, or their status as an expert. See Table 4-1 for the multi-sited interview frame. These professionals were primarily donor-country nationals (n=26) or from other common donor-countries (n=8) and the aid-recipient country (n=20). Three individuals completed multiple interviews, two of which were at the donor field office, allowing for deeper knowledge of donor evaluation practices. Verbal consent was obtained at the beginning of all interviews, and recordings were made when individuals consented. Each participant employed by IDAD took care to mention they did not represent the organization and were sharing personal opinions. Semi-structured interviews lasted anywhere from 1

hour to 2 hours and took place in English—the working language for these projects. A primary interview guide was used and tailored for each role and location. I found some professionals spoke clearly of evaluation processes and procedures as an object to be researched, while others "sat within" the system during the interview.

I began analysis by coding all transcripts by occupational role, organizational type, and location (headquarters or field). I then attempted to code for different measurement components, such as target, indicator, report, to understand professionals' relationship to such types of measurement. Yet I found professionals describe these practices not as singular engagements but as interlinked components that resist decomposition. Therefore, I began to look for similarities in experiences of measurement. I found commonalities in how professionals responded to the presence of evaluative structures and the practices they enact as they go about their everyday work. Adopting the perspective of structures and practices then allowed me to tease apart different elements that make up quantified evaluation as an institutional logic.

Table 4-1 Multi-sited Interview Frame: Organizational Sites and Occupations of Interviewees

Organizational Sites	**Sample Interviewee Titles (n=number of individuals)**	**Evaluation Artefacts**
Large Project Field Offices (n=4)	Project Leader, Monitoring and Evaluation Specialist, Gender Advisor (n=13)	Monthly & quarterly reports Annual Reports Standardized Data Tools Performance Monitoring Plan Project Gender Analysis AGMIS (input)
Evaluation Organization Field Offices (n=5)	Project Leader, Deputy Project Leader, Senior Researcher, Research Fellow, External Evaluator, Field Manager, Research Consultant (n=15)	Impact Evaluations External Evaluations (review all project documentation)
Field Landscape	Chief of Party, Country Director, Agriculture & Gender Researcher, Gender Advisor, Policy Consultant, Evaluation Lead (n=10)	--
Donor Field Office (n=1)	Agriculture Officer, Project Officer, Responsible Representative, Evaluation Advisor, Gender Specialist (n=7)	Contracts & Agreements Indicator Justifications Indicators Handbook Country Strategy Gender Strategy & Plan AGMIS (review)
Large Project HQs (n=2)	Director, Program Performance & Quality, Senior Evaluation Director, Director, Gender & Social Inclusion (n=4)	New Business Proposals Performance Monitoring Plan Definition of Indicators
HQ Landscape	Consultant, Senior Advisor & Thought Leadership, Evaluation Director, Learning and Knowledge Management Consultant (n=6)	--
Donor HQ (n=1)	Director, Office of Learning, Evaluation, and Research; Gender Advisor, Africa Department (n=4)	Request for Proposal (RFP) Compliance Rules & Regulations AGMIS (review & publicize)

Notes: HQ=Headquarters; AGMIS=Agricultural Monitoring Information System
"Landscape" includes expert interviews with professionals serving smaller Farming Sustenance projects or closely related agricultural development projects. To check that the evaluation system was mandated and standardized as reported in donor documents and observed in the primary case, I conducted 2 interviews with Evaluation Directors serving large projects under Farming Sustenance in another East African country. A total of 62 interviews were conducted with 59 professionals.

FINDINGS

This section documents quantified evaluation as a logic in international development bureaucracies and juxtaposes how efforts for gender equality in development programming are handled. The manifestations of quantified evaluation occur within and between different office sites nested within the greater aid chain. Institutional logics are unseen forces that manifest in how individuals make sense of the array of options for action. Analyzing structures and practices then brings an institutional logic into empirical view.

Donor commitments to particular development efforts are turned into evaluative structures—evaluation-related artefacts—that demarcate professional action: the contract, scope of work, project design, and more. These documents determine the parameters of projects and outline priorities. Evaluation systems take shape around these documents and people organize behavior to accomplish what is articulated in the plans. Therefore, I separate quantified evaluation into two layers for analysis: structures and practices. See Table 4-2 for the structures and practices described. Evaluative structures serve as key levers that reverberate throughout development work and thus are discussed first. I then document the practices that development professionals put into action, making the logic of quantified evaluation visible in the development workplace. Empirically, I document how the structures and practices of the logic of quantified evaluation marginalizes gender, providing a systemic look at how gender remains on the outside of agricultural development programming. In the conclusion, I discuss potential possibilities for this to change.

Table 4-2 Quantified Evaluation as an Institutional Logic that Marginalizes Gender

	Manifestation of Logic	Incorporation of gender
Evaluative Structures	Evaluation Organizations	Do not include gender
	Statement of Work (SOW) & Terms of Reference (TOR)	Include one question on gender, allowing the compartmentalization of gender in the project. Often stymied by lack of data from M&E systems
	Monitoring & Evaluation (M&E)	Gender is included at the lowest level, as targets for only 2 indicators out of 52
	Web-based Monitoring: Agriculture Monitoring Information System (AGMIS)	Digital reporting leaves out unquantifiable & normalizes attention to indicators
Practices	Comply with bureaucratic mandates	"Tack gender on" to existing projects rather than during new project design phase
	Focus on agricultural development	Gender is perceived as peripheral to agricultural development
	Communicate quantitatively	Difficulty of quantitative measurement reduces gender to simple measures that do not capture gender
	Focus on project items that are capable to be marked "done"	Gender is unquantifiable, undermining an ability to document completion
	Hit targets	Hit quantified targets without supporting the intention of the target

EVALUATIVE STRUCTURES

Evaluative structures exist at every level across the aid chain, outlining how staff time, attention, and resources should be allocated. I demonstrate below that quantified evaluation is buttressed by 4 key evaluative structures: evaluation-specific organizations and contracts, Statement of Work (SOW) or Terms of Reference (TOR), the monitoring and evaluation system (M&E), and electronic and digital forms of monitoring. At the same time, gender-related efforts are marginalized through these evaluative structures: the

absence of gender is reified by a new organizational actor, excluded in SOWs and TORs or undermined by a lack of data, not counted by M&E systems, and the possibilities for more complex measurements are upended by the focus on only quantifiable project aspects in the web-based AGMIS reporting system.

Organizational Commitment to Evaluation

A new organizational actor in development has emerged: donor calls for evaluation-specific contracts has resulted in evaluation-specific organizations. The role of these organizations is increasingly important in development work. With new actors comes new potentials, yet as I demonstrate below this new organizational actor amplifies and reifies the current marginalization of gender with development work.

First, an organizational reconfiguration is present. As observed in other institutional sectors, such as higher education (Sauder and Chun 2019) and the Institutional Review Board (Babb 2020), evaluation in international development is only growing. In order to manage these growing interests and requirements, there has been an organizational shift. The most common relationship in development is between the donor and the implementer. Utilizing materialist analyses of Cooley and Ron (2005) and Krause (2014), I propose terminology which draws attention to this reconfiguration and focuses on the underlying dynamics present in development relationships. I term donor organizations "buy it" and implementing organizations as "do it" organizations. My fieldwork captured an organizational reconfiguration in the existence of an undocumented organizational element in the development landscape: a new parallel group of contractors and projects with the sole purpose of supporting evaluation within and between projects. The main function of

these is “to check” on the “do it” organizations. These are evaluation-specific contracts in which the donor has purchased evaluation-related results. While evaluation organizations, in this view, could classify as “do it” organizations, the substantive content of the project goals is not to bring about development, but to bring higher quality project evaluative practices, sharing and learning across “do it” organizations, and to complete evaluations of “do it” organizations. In other words, “check it” organizations (evaluation-specific contracts) exist to help “do it” organizations handle the rising demands for compliance and evaluation and allow “buy it” organizations to further outsource and professionalize demands for rigorous evaluation.

In order to understand this new organizational actor, we must first understand what they are contracted to deliver. And, like all development projects, tasks and requests morph to suit client needs. The original idea behind Contracts 1 and 2 was to increase learning, collaboration, and effective programming of the initiative (see Table 4.3). However, during interviews, professionals reported that the actual activities associated with each contract turned into a stronger focus on performing external evaluations. This act of evaluation places the “check it” organizations as *overseers*, which upended their ability to then later serve as facilitators of learning, prompting “do it” organizations to refuse to share what was or was not working during project implementation. And although the third contract was meant to provide robust evidence of the outcomes of the entire initiative, the sheer cost of completing statistically significant randomized surveys meant that it required a ten-percentage point drop in poverty levels in order to appear significant. Researchers on this project felt the donor had “shot themselves in the foot” by setting such an ambitious goal. At the time of my data collection (and from internet searches in 2019), this report has not

been made public. In practice, these contracts were turned into *evaluative* contracts, rather than addressing learning or capacity building as originally intended. This demonstrates that the bureaucratic demand for evaluation overrides other concerns and the privatization of data by the donor of a publicly funded initiative demonstrates the high-stakes context of poor results.

Table 4-3 Evaluation "Check it" Contracts in Support of Farming Sustenance

Evaluation Contract	**Project Objective** (source: official documentation)	**In practice** (source: interviews)
CONTRACT 1 ($9.3 million)	Facilitate learning and knowledge sharing across *Farming Sustenance* implementing organizations	Completed external evaluations of *Farming Sustenance* implementing organizations. This undermined the Evaluation Organization's ability to build knowledge across the projects in a non-evaluative sense
CONTRACT 2 ($16 million)	Evaluate projects, capacity building of implementing organizations and local organizations in M&E, enhance data quality, promote collaborations across implementation organizations	Assist donor field office in strategic visioning, complete external evaluations of projects. Capacity building for IOs took a backseat to assisting donor field office
CONTRACT 3 ($2 million)	Perform statistically significant randomized survey of *Farming Sustenance* initiative and analyze results	Perform statistically randomized survey of *Farming Sustenance* initiative and analyze results. Midline report was not made public due to poor performance

I turn now how gender-related efforts fare in the presence of this new organizational actor. In each "check it" organization, the projects were run by men, and predominantly staffed by men. Only one of the three contracts had an assigned "gender person." This was a man who held the title consultant and amongst his many responsibilities he was also the "gender focal point." In telling me how he came to occupy the position, he said that at the

time the contract started his boss approached him saying "You are the gender focal person." He responded, laughing, "What?! You already decided that?" The boss replied, "Yes. I just looked at the team profile and there are no women on the team." The consultant went on to say that not only women should be involved in gender and throughout our conversation it became clear that he approached analysis with a gendered lens, however this was due to him *as an individual* rather than these organizations making concerted space and positions related to gender (a common bureaucratic response).

The incorporation in gender in the other two contracts was driven fully by donor interests as documented in the Statement of Work or Terms of Reference. The large-scale survey contract collected and produced data on women's nutrition activities and began writing research reports, yet this was because they are contracted to deliver such reports. And the other organization hired a gender consultant to complete a gender-specific assessment of donor activities. She confided in me that she didn't know where to put what she really wanted to say in her final report—that the donor needs to pay attention internally to gender politics—because it was not an answer to a question in her Statement of Work. In other words, what happens within these evaluation-specific organizations and contracts is determined by another evaluative structure: that of Statements of Work and Terms of Reference.

SOWs and TORs: Documents that Structure Attention

"Check it" organizations are required to complete external evaluations of the initiative's core projects. Each evaluation has a separate contract which outlines the questions the evaluation report should answer. In this manner, the Statement of Work

(SOW) or Terms of Reference (TOR) is a key lever in determining attention to women's empowerment and gender equality.

In order to assess the questions in an evaluation SOW/TOR, consultants turn to a variety of sources. These typically include interviews and focus group discussions with staff and beneficiaries and a review of existing project documentation and the project's performance against their contracted indicators and targets. In the midterm review of *Farming Sustenance* projects, for example, there was a question on how projects were advancing "women's empowerment," "gender issues" or nutrition. For example, if the SOW included 7 questions that must be addressed, 1 question would ask about "women's empowerment." Yet because the quantified evaluation system offered little information on women's participation and outcomes, there was little data to assess:

> Being in the SOW facilitates that it [gender] will not be overlooked because there is a specific question on gender. Problem is that at the design stage [of a project]. Our finding is that at the beginning of the project, they didn't have specific indicators on gender. At the design stage! … If it's not in the design, it's not even budgeted.

This results in a one step forward and two steps back situation. Evaluators, too, are under contract to deliver the client, the donor, what they asked for in a report. In this manner external evaluations often promote the idea that "gender work" can only be captured in anecdotal forms. If there is a question about gender in an evaluation SOW, the final report often concludes that there is too little data to answer the question or will include a narrative

summary about a particular training that was conducted, highlighting the number of women who attended. Across all of the midterm and final evaluations I was able to access, there were two reports which did document that gender had not been properly accounted for, budgeted for, or even tracked during the course of the project.

Daniel, an external evaluator, noted that women's participation had been a strong thesis in the initiative. Upon asking where he got this information, he replied, "Within the scope of work, in the program [initiative] design, all of that says 'Yes, we're going to get women participating more.' So, it's driven from the top down—the idea that you've got to have this female participation—which is great, but at the same time: how do we achieve it? And you end up looking at indicators that are just not clear." Despite stated intentions to empower women and garner their participation in project activities, Daniel points out that these ideas need to be properly operationalized in indicators. I now turn to these lower level systems and explore how gender is incorporated monitoring and evaluation systems.

Quantified Monitoring and Evaluation Systems

Given the focus in the development sector on demonstrating results, then the incorporation of gender efforts into project measurement systems is a proxy for its relative importance to other project activities. Irrespective of the funding mechanism, the International Development Assistance Department (IDAD) mandates sex-disaggregated data collection and reporting for all people-related indicators. For all largescale initiatives, indicator handbooks detail each indicator. The *Farming Sustenance* Indicator Handbook (Updated 2016) is 135 pages long and includes 52 different indicators. An indicator is essentially a unit of measure or "a quantifiable measure of a characteristic or condition of people,

institutions, systems, or processes that may change over time." All projects under IDAD are required to have a variety of indicators attached to project designs and strategy documents. 'Performance indicators' track the output, outcome, or impact of IDAD investments. All implementing organizations are required to perform based on a pre-defined set of indicators and associated targets. In Table 4-4, the only two *Farming Sustenance* indicators that engage sex-disaggregated data are shown. These indicators are tracked by the "do it" organizations or implementing organizations. Other indicators include "value of sales from farms in project" or "hectares under improved management."

Table 4-4 Gender-Related Indicators from Farming Sustenance

Type	Title
Output	Number of individuals who have received short-term agricultural training *Ex. 25 Men trained & 13 Women trained*
Outcome	Number of farmers and others who have applied improved technologies *Ex. 17 Men & 10 Women using new technology*

Although *Farming Sustenance* promotional materials highlight the role of women and gender equality as critical to achieving global food security, this stated interest is not reflected by the indicators selected to measure. The presence of women in the indicator titles is generally seen as biological and in their connection to a childbearing and care giving role under nutrition. Since these are people-related indicators the donor mandates data be disaggregable by sex. The attention to women's role in agricultural and gender equality is seen only in project indicator targets when disaggregated by sex, such as a performance target of 45% women trained or 30% women using a new technology (targets are set at a country level).

While IDAD describes *Farming Sustenance* as an "inclusive agricultural development program" this is not currently reflected in the monitoring and evaluation system as mandated by IDAD. The mandated M&E system does not include attention to gender or women other than sex-disaggregated indicator *targets*—the lowest level of the M&E system (Springer, 2019). To the development professionals who run these projects, these two indicators exist within a greater set of 15-30 indicators that their project must deliver to the donor. And they track their performance against annual targets for performance in a digitized web-based monitoring system.

Digital Evaluation Structures

Evaluative structures are now in a process of being digitized, rendering items which are not easily quantifiable and aggregable to the sidelines. In this context, richer, more contextualized data about gender and gender relationships in projects are unlikely to be incorporated and the possibilities for such inclusion are in a state of decline as organizations invest in digital tracking methods.

The Agricultural Monitoring Information System (here called AGMIS) is a custom-built system for *Farming Sustenance* and a global data collection platform that opens once a year for reporting. From the perspective of the implementing organizations, AGMIS serves as a data entry portal into which they are contractually obligated to enter their performance on mandated indicators. Prior to the opening of the web portal, evaluation directors at implementing organizations tabulate and aggregate data from all of their rural implementing organizations (in most cases the local IOs were 3-4). Once the portal opens, the evaluation director is presented with a user interface that presents the indicators they

are contracted to deliver with a column that charts their previous performance (targets already achieved) and the annual target for the reporting year. This is followed by a column of to-date performance, which evaluation directors enter. If the amount entered is above or below 10% of the expected performance, another narrative justification field will open. In this open text field, evaluation directors enter narrative text as to why they have under or over performance for that indicator. Not only does AGMIS demonstrate donor priorities to accomplish annual targets, but by requiring staff to justify performance irregularities, they coercively keep implementing organizations on track. This system is the core mode of reporting.

After the individual implementing organizations enter their data, it is routed to the donor field office, where their responsible representative (RR) will review the data. Because AGMIS will auto-lock at a set date, IOs must input their data in enough time for their RR to review, comment, and ask for any justifications. With the click of a button, the RR approves, and the performance is routed to the donor headquarters for review. At donor headquarters staff log into the AGMIS and review each file, commenting on performance and asking questions. If donor headquarters staff want additional explanation, they will reroute the form back to the implementing organization and unlock access. For example, one project was running a women's-only business training program (note: it is extremely rare for projects to run activities only for women) and entered into AGMIS sex-disaggregated data, which showed zero men beneficiaries. Donor staff wrote and wanted to know why there were no men beneficiaries despite the predominance of men in all other activities.

All evaluation directors brought up the AGMIS system in the context of deadlines and stress. The system draws the evaluation directors down into the minutiae of indicators in what feels like a high-stakes process. It should be noted that AGMIS is not the only digital system: implementing organizations must enter their data into a separate system which aggregates data slightly differently. Evaluator directors attend webinars which train them on how to complete this data entry, or if there are any changes to how indicators should be tabulated. Digital systems are useful for aggregation across implementing organizations, field offices, countries, and entire initiatives. Yet numerous professionals questioned if the data entered into this system was ever truly reviewed or utilized. Although donor staff did not mention this, consultants mentioned the sheer volume of data and noted that no one seemed to truly be looking at the data from the headquarters side:

> Here's the thing, how do you use something like that? You know how it's *supposed* to work, maybe you can search it or drag out the higher-level indicators. But can you honestly use something of this size with any sort of ease? It's hard enough to just scroll through. Let alone… you haven't gotten to the bottom yet! [I scroll down to the very bottom and we both laugh] *Line 5,489*. And you're still looking at just this country's indicators. Now somehow a lot of them are blank. At the same time, in my mind, this is a totally unreal management tool. I don't see how they could really use that.

In the context of digital evaluation systems as "efficient" aggregating systems across global portfolios of development projects, the trend toward quantification is only

accelerating. Transformative processes—like women's empowerment, capacity building, leadership, and others—that are difficult to quantify in meaningful ways will be increasingly pushed out as project priorities. As such, the trend towards quantifying social processes is expected to continue and grow as is utilizing the count of attendees to trainings, despite the questionable worth of this indicator. Digitized evaluation systems compound these trends by raising the expectation for automatically aggregable data.

These four evaluative structures serve to systematically marginalize gender-related efforts in development programming by placing parameters on what is considered reasonable and legitimate action in the development workplace. These structures work in tandem: evaluation organizations and the contracts they execute include gender only if it is contracted by the donor or the personal prerogative of an employee; statements of work pay marginal attention to gender; monitoring and evaluation systems do not incorporate gender-related indicators; and digital tracking systems have no field for anything other than simple aggregable data—meaning "gender" must be measured in quantifiable forms in order to be tracked.

PRACTICES

Practices are elements in organizations that have become commonplace, regular, and taken-for-granted ways of doing things. Mundane activities coalesce into work practices—accepted ways of acting that elicit cooperation—in the development workplace. Below I demonstrate several practices in development that work in concert to sideline and marginalize authentic attention to gender-related development programming. In each

section, I describe the standard mode of operation and then document what happens to gender in this context.

The everyday social work of development was described by the participants as one of overwork, stress, deadlines, and a general sense of "too much to do." IDAD staff would respond to "How are you?" with "Good. [slight pause] Okay. It's reporting season you know, so… long days." Implementing organization staff similarly marked their days in relation to project activities: "just got back from the field" or "working on a quarterly report." With overwork as context, development professionals found themselves reduced to firefighting rather than time to proactively address the project. In order to handle this high-speed work context, a series of practices emerge, each of which marginalizes gender-related work or further entrenches its marginalization.

Comply with Bureaucratic Mandates and Tack Gender on as an Afterthought

Dealing with bureaucracy is a daily feature of anyone associated with a large donor office, especially IDAD. Since bureaucratic requirements are ubiquitous, understanding what requirements exists and how gender is, or is not, included in them provides a meaningful baseline of practices. I demonstrate here that professionals and organizations live and breathe IDAD bureaucracy, complying with its mandates. This common practice results in gender being "tacked on" to already existing projects in order to appeal to bureaucratic requests, but not legal requirements.

Complying with the rules and regulations for running IDAD projects takes a disproportionate amount of time. Martin, a project leader of a large project, describes the rules for contract oversight and its impact on the everyday project activities:

> I think we are using 30% of the time in the budget the project just for bureaucracy. I produce a 145-page work plan every year and to implement that work plan I still need 120 approvals in the twelve months after that. An *approved* work plan! Every person I need to hire for a few days needs to be approved and that means approval on scope of work, experience, salary, everything, 'Is that the right candidate?' or 'Doesn't have to make that kind of salary' sort of thing [mimicking a donor responsible representative] those are the main things. Every procurement over $500 needs to be approved it's a $41 million project so there are quite a few of those.

Martin draws attention to the sheer volume of tasks and time required to deal with the bureaucratic requirements of the donor. Another project leader joked about "our overlords at the mission" while describing how oversight has been increasing over the last several years.

Bureaucratic procedures are a mainstay of IDAD's functioning and management of field offices around the world. To operate within or as a vendor to IDAD, professionals must successfully navigate and incorporate IDAD operating policies into their everyday work lives. All operational policies are documented in a large repository, including rules and procedures. For example, operating procedures are outlined in hundreds of pages that cover numerous topics, such as how project designs should be conceptualized, requirements for proposal bids, and requirements for project implementation, required

documentation, etc. These procedures are legally binding, and thus successfully elicit compliance from organizations and professionals around the world.

Despite the fact that IDAD has released policies on how gender should be fully incorporated—into the conceptualization, description, M&E system, reporting, staffing, and scoring criteria for the selection process—it is often not included. How could this be? Donor responsible representatives are in charge of ensuring compliance with bureaucratic policies, yet with so many indicators to follow up on, sex disaggregated data and women's involvement never seems to be a priority. Unlike the other main cross cutting issue, climate change, gender advisors at IDAD headquarters do not hold clearance authorization, meaning they are unable to legally require compliance. In other words, the incorporation of gender is strongly encouraged, but not legally required. Gender advisors, in both the donor country and the field, confirmed that strongly encouraged was not enough to get gender incorporated in meaningful ways.

The lack of incorporation and legal power to require it, creates a circumstance where the incorporation of gender is always sliding to the next tier in the aid chain. IDAD staff are meant to be responsible for ensuring solicitations are written to include gender (but they often do not). Which means implementing organizations respond to only what is requested in the solicitation. If gender is not mentioned in the RFP, the HQ-based gender advisors work with the New Business Development teams at their organizations to try to "get gender in." If gender does not "get in" to the winning proposal, the gender advisors in both the donor country and the field will have a hard time advocating for gender to be included in the project activities. Here we can see the power of the solicitation to magnify the inclusion or exclusion of gender throughout the rest of the project cycle. The absence

of legal power within a large bureaucracy to incorporate gender creates a waterfall effect that leaves gender advisors and gender-related project activities on the outside of development projects.

I now shift gears to describe how professionals deal with gender within bureaucracy. The practice of compliance with bureaucratic requirements creates a secondary, downstream effect: tack gender on. One gender advisor stated it plainly: "When I first started in it, I felt like I was annoying people. Gender was always the thing that you worked on last and forgot about and then stuck in a few words." This action of "add it on" was reported by so many different professionals at different levels throughout the aid chain, it emerges as a practice in development.

In any bureaucracy, the tendency towards compliance without buy in, or compliance without understanding, can be seen when everyday workers 'check the box.' 'Checking the box' was mentioned by numerous participants in a variety of professional roles—from IDAD staff to implementing organization staff. Numerous participants, while motioning with their index finger, literally said 'Tick! Tick! Tick!' in relation to supposedly meeting some IDAD requirement for their project, including the staff who work at IDAD headquarters. One IDAD professional based in the donor country agreed "Yes, gender is 'check the box' but *everything* is check the box." This statement demonstrates the daily grind of working within a bureaucracy and also perhaps the realities of international development, which requires the coordination of professionals and documentation across vast temporal, geographic, and cultural dimensions. This was confirmed by an external evaluator on an IDAD project:

> Certainly some [indicators] are just easier to achieve than others and so I think that there was a little bit of a sense of 'let's try to meet as many as possible and if we miss one, even if it's the most important for our actual project, it's more important to at least check as many boxes as possible. (Beth, evaluation consultant)

With "checking boxes" as a predominant practice, adding gender on as an afterthought allows busy, stressed professionals who aren't quite sure *what* to do with gender a bureaucratic way out. An evaluation consultant at a "check it" organization explains:

> So what happened is that they came to add [gender] on after they'd started. So it really reduced their effectiveness in that regard because it wasn't well thought through from the start. But they've done okay. For example, they have promoted messages for men and women, cooking demonstrations, gardens, they've done reasonably well with men and women, but when you don't have at design stage. You haven't thought about it. You just came to it in the middle. Which we said as evaluators. They accepted it. This is the importance of being in the terms of reference.

In other words, projects are designed without attention to gender because of the lack of legal veto power to prevent a project design from moving forward described above. By not including gender as an integral part of the project at the beginning there are only two options: leave it completely out or to tack it on. Given the stated interest of the initiative in empowering women, it would merely be poor form to leave it out completely. One of the ways IDAD was able to push responsibility for gender (despite not writing it into the

Request for Proposals) is by including the single women's empowerment question in external evaluations. This allows IDAD documentation to show an interest in gender and then the IOs look as if they are not performing. However, donor field office staff noted that often, even if a question is in the external evaluation SOW it is often removed because of budget concerns and a focus on the more technical (meaning agricultural) project components.

Many of the project leaders were a bit baffled when I asked about the donor emphasis on gender, one even going so far as to ask in retort, "*Do* they care about gender?" Geoff, an external evaluator, explains how gender, though a stated core component of the initiative, was secondary to the main agricultural development framing of hunger and poverty:

> The focus was on improving hunger and poverty and not necessarily as it impacted women. But in fairness, the design of the project was like that as well. In designing these projects, there is always a gender equality aspect that comes in at some point, you know you have to do something about this, and you sort of tag it on at the end.

Interviews with staff serving evaluation-specific projects noted a common theme: gender was "tacked on" as an afterthought. This was true for all projects except for one. The project which had incorporated gender from the beginning had done so out of the organization's own volition and even, at times, felt stymied by the donor. These statements were reinforced by donor staff, many of whom noted that gender was largely a "check the box"

activity. One of the leaders within the donor field office said, “My sense is that it was a checkbox in the past. Maybe it still is, I’d like to think people take it more seriously.” Despite the donor’s stated interest in women’s empowerment, the presentation of gender as a “check it off” item was reconfirmed by other donor staff:

> Depends on personalities, depends on how much they value gender integration. Everyone pays lip service. But it’s often viewed as a layer of bureaucracy, more ‘We have to do this to get our SOW [statement of work] through the procurement office’ rather than saying ‘Okay we know this is important and how can we meaningful address this and what does it look like?’ (Milly)

This donor staff highlights a mediating variable on top of mere bureaucratic requirements: personality or individual commitment. If gender is predominantly understood as something to “check off,” how does one garner attention to gender mainstreaming? It is in this context that Gender Advisors and experts, in the donor country and the field alike, highlighted the role of leadership and champions. And yet, each role mentioned the importance of the person in the position above them. While they were perhaps frustrated with their coworkers, a strong project leader could make things happen on a project. For project leaders, they needed their responsible representatives at the donor office to be amenable to changes in activities or budget allotment to support women’s engagement. And so on and so forth.

While most professionals spoke to the importance of their manager's leadership, numerous HQ-based Gender Advisors (nearly all women themselves) commented on the level *below* them, noting 'the good old boys' club of project leaders as a key barrier to greater incorporation of gender in projects. They noted that uptake was slow because the managers of these massive projects around the world were generally managed by old, white guys. The "aggies" (agriculture specialists) were highlighted as the worst offenders. Although this perspective was shared by nearly every donor-country citizen, the gender advisors in the field spoke about the importance of their project leaders in making gender an agenda item at the level of the project. The acknowledgement that gender is predicated on organizational power dynamics, which are gendered, leads to the next common practice in development: outsourcing "gender work" to gender advisors.

Focus on the "real" project, the gender advisor will do the gender

The offloading of gender work within everyday project work has been documented in Chapter 3 (Springer 2019). Here, I want to highlight men development professionals' unwillingness to take up gender as a core element of the project. This was a dynamic reported primarily in headquarters by gender advisors working for implementing organizations. These advisors are less interested in compliance and more interested in helping support implementation of gender-related work within the projects.

The leftover residual of "check the box" mentality carried into the minds of these so-called aggies. One donor-based professional commented that it was strange the aggies would be resistant to gender work because across their decades-long careers they traveled and have witnessed firsthand exactly who is doing the agricultural labor in developing

countries. Danielle, a gender and evaluation specialist at headquarters, connected workplace resistance to the more intimate nature of gender:

> Gender is personal. Agriculture isn't personal. People get mad about the pest destroying crops, but that's a universal feeling. Gender is really personal. So learning around gender is personal… A good example of this: they [men] do trainings for staff for understanding gender better, and they say, 'But *I* don't do gender!' They don't understand is that it's just part of their life and they do it.

This draws attention to how men execute gender-related events or participate in some way but appear unable to use a gendered lens to view the world—which would imbricate them into relations of power in their everyday lives. Another gender advisor in the field, Rosalie, agreed:

> You don't necessarily have to depend on the gender advisor only. The way they [men] make it sound it's like we are holding the key for nuclear physics! A passport or something like that—we don't! It's not! It's not that complicated! You just need to pay attention to it!

These gender advisors are asking their men counterparts to understand the social world, and specifically development projects, through a gendered lens. Gender advisors mentioned banal elements in order to approach the project with a gendered lens. For

example, if the project is dropping off agricultural inputs to bag them in smaller bags less than 110lbs so they are easier for women to carry.

Gender advisors were careful to note that not all women understand how to operationalize a gendered lens, but planned events specifically to create a safe space for men to grapple with issues around gender and help men gain a gendered perspective. However, men do not attend these events. Gender advisors reported that even the act of garnering men's participation in these events that should have be mandatory fell onto their plate. One gender advisor explained:

> One of the dynamics today, and I've gotten more angry about it, and I don't have to be as diplomatic because now I'm a consultant. The idea that the burden is on gender specialists to engage men when we have gender working groups. Like it's our own fault that men aren't there. We need to make them feel welcome. Not many men come, and we need to do a better job of making them. And in the beginning, I sort of believed that but now I think its bullshit.

This was a common feature mentioned: make sure men coworkers are engaged in gender, even if they don't self-select for participation. Another headquarters-based advisor told a story about how she had worked so hard to get men into the room and watched as all men sat silently despite the fact the event was called "Getting men engaged in gender." When she confronted a coworker she was close to, he responded that he was not going to say a single thing during the event because if he said one thing wrong he would be bashed. The

gender advisor described feeling exasperated that even her best attempts to engage men were stymied.

Gender advisors mentioned other ways in which they had to fight for respect in the workplace beyond project activities. One gender advisor said, "You have to prove yourself constantly, no one recognizes you as a professional like they would if you can say 'I have ten years of experience in education.'" Gender advisors in the recipient country spoke about their status as women and occupation of gender advisor as key reasons for not being taken seriously by their coworkers. Whereas, in the donor country, white gender advisors spoke about age as a barrier to being taken seriously while women of color brought up race and nation as categories upon which they are "discredited."

The numerous practices of bureaucracy and checking the box are amplified by the interpersonal dynamics between men and women who are in the gender advisor role. Gender advisors in all locations mentioned the sheer difficulty of their jobs, constantly trying to advocate for gender and being the ones to "go knock on doors." They are expected to "do it all," when, the concept of gender mainstreaming is a team effort, requiring all staff members engage in ensuring gender is incorporated throughout the project. Ellen, a donor-country gender advisor, explained:

> If it hadn't been in proposal, we'd ask the project leaders what they were willing to take on. Sometimes just 1 indicator to humor me… I guess that buy in from the top was missing. It needs to be top down not just gender folks squawking about it. You really need an insistence from senior management.

Ellen, like her gender advisor counterparts, highlights the importance of buy-in from senior project management. Because gender has not been incorporated into the structures of development projects, the incorporation of gender into projects comes down to the willingness of individual project leaders—a space that invites resistance due to the personal nature of gender politics.

Communicate with quantified data, irrespective of its worth

Communication between implementing organizations, donor field offices, evaluation organizations, and headquarters occurs through documents—but it is the numbers on the pages that most professionals are invested in. After the midterm evaluations of the large "do it" projects were completed, the external evaluators described donor field staff as anxious about their performance: "There are a lot of reputations and, frankly, careers on the stake of meeting those targets." Although quarterly and annual reports have narrative description of projects activities, performance against the project objectives and targets are of paramount importance. Nearly every professional described some level of distrust of numbers, many sharing stories of potentially falsified data. Despite this mistrust, professionals are situated within interlinked bureaucracies that demand quantitative data. Professionals at every location described a yearning for more meaningful indicators, as a donor representative, called Aiden, shared:

> Reporting demands, that's the problem. You try to have indicators to measure, but then you end up with just requirements on the number of

> women. What's the real change if anything? You do a field visit and they [implementing organizations] say, 'We'll introduce you to one of our women owners!' and you go and you ask a question and the husband answers everything. I mean c'mon—who is the owner here?

While implementing organizations often deride donor staff for not understanding, this donor representative, like his counterparts, shared in the skepticism about quantitative indicators about gender-related topics. However, he notes they are "on the hook" due to reporting demands. The focus on quantitative data appears to come from a single external party: governing bodies that allot funding. Here Tiffany, a gender advisor at headquarters, described:

> If we don't count it, it doesn't count, and I think that's absolutely true. We do need data, we absolutely need data and quantified data to make policy decisions and to get the political will, so in that sense it's important.

The power of this single external audience to drive reporting demands ricochets around the globe to development organizations and professionals through IDAD's bureaucracy and reporting requirements. In describing their relationship to headquarters, another donor rep, Mariam, in the field office said:

> It's data that we provide, they need data to go to politicians. Maybe on job creation: 'this many new jobs.' We basically play with data. This is the

> means of communication, the data, we just have the privilege to go out and also see at the project site with the implementing partners. This is our privilege.

In this configuration, reporting demands for *quantitative* data are turned into the main mode of communication across the tiers of the development aid chain. Mariam goes on to say that she does not trust numbers and would rather go out to see projects in the field, which she calls her privilege. In other words, it is not central to her work, but peripheral. Miriam explicitly ties this with the immeasurability of gender while describing a more meaningful indicator of women's empowerment. She and her coworker, Rahel, have a conversation:

> Mariam: If you train 100 people it doesn't actually mean anything. We want to know where they were and where did they go. But you just cannot measure gender.
>
> Rahel: We went out and saw women beneficiaries. We've seen how they experience the project. That was my way of monitoring and evaluating.
>
> Mariam: One time we called all of the ladies and there's this one rural woman from the north and she was nursing. Her husband even came with her to be with the baby and was babysitting for her. So she gets to be at the training and she kept leaving to nurse the baby. *This* is an indicator.

Although these donor reps have found what they feel are more meaningful indicators of changing gender relations, they too are drawn into a quantitative logic for communicating up and down the aid chain, saying "it would be hard to aggregate." Despite the reliance on quantitative data for communication, with the AGMIS system as a key example, professionals in all organizations in the field spoke about one-off requests from donor headquarters as taking up much of their time, maybe it's to share a positive story, perhaps to share details on a particular piece of the project, etc. Theory predicts that numbers are an efficient way to communicate and yet the donor has needs beyond just the quantified systems they have developed and mandated. But quantification offers organizations a very valuable tool: organizations can claim they are "done" and have finished a quantified activity.

Work on what is "doable," sideline what cannot be "checked off"

Organizations implement development projects on a set timeline with a set budget using the staff they are able to hire. Their ability to execute the project and achieve the results, as contracted by their client (the donor), impacts their reputation and ability to win future business. In this circumstance, organizations have strategic reasons to focus on programming that is achievable and demonstrable. Under this *modus operandi*, however, gender faces additional barriers, as one donor representative, Milly, said, "Everyone knows it's important, but when on a tight timeline to get an activity awarded and in place, we don't want to do more analysis, we just want to move." The very context of everyday development work as rushed and high stakes leaves that which cannot be easily "done" pushed to the sidelines.

If the daily work orientation is toward what is accomplishable (and then able to be demonstrated to the funder), what does this imply for gender-related programming? The tough-to-quantify area of gender and women's empowerment gets de-prioritized, especially in comparison to more quantifiable aspects. Future funding and do-ability become intertwined, as Bernadette, a former headquarters-based donor staff member, describes:

> The whole issue with gender is that it's not tangible. For people, even those who understand the importance of gender, it's hard to show what you've achieved. It's very hard. You'd get more buy in if you could show it. Over time people are moving towards scientific explanations: 10 seedlings and 1000 sweet potatoes. Village and nutrition. If you can measure it, then funders are more likely to fund. But if you can't say that, then people are less likely to believe it. You may have done that, but maybe not.

Bernadette demonstrates that fundability drives what occurs within projects. Indeed, gender advisors at the project level spoke about the immensurability of gender. This drove them to understand and even *sympathize* with the sidelining of gender. This very act of accommodating the struggle of the project demonstrates the sector's focus on results:

> But with millions of orders, deadlines, issues I cannot blame managers for this [sidelining gender]. Too many priorities. And they want to pick something which is clearer and in-the-face, something clear, something that

> has already set standards on how to do it. If you want to construct a latrine there's a rule for how to construct a latrine, you know how much input that you need, you know when to construct and, so they give priority to something that's just doable so you can do it and tick! It's done! (Rosalie field-based Gender Expert)

The context in which development organizations operate is at odds with the inclusion of gender programming, which is understood as longer, harder, and with ambiguous outcomes. Although the donor gender advisor understood that gender was difficult to measure—saying "If it's the budget, it's not a problem 1+1=2. But with gender 1+1+1=4 or 5"—the realities for implementing organizations are to be attentive to their client. Rosalie continued, explaining how this dynamic undermines working on gender social norm change before it even begins:

> But social norms? Way complicated! It requires time for thinking and analyzing and for you to sometimes say, 'Okay, we don't know where we're going.' Sometimes you could be going—change doesn't come only in the way you want, it could go the other way. You don't know the change, it could go somewhere else! But tracking that, following that, is impossible when your donor is also saying 'Did you reach the targets? How much did you reach?' Success rate, baseline, base value, 'how was the base value and where is it now' kind of thing. And it's ambitious, very ambitious.

The everyday life of professionals serving development projects socialize them into

focusing on what is do-able and demonstrable to the donor. Although gender advisors, at implementing organizations and the donor alike, realize that simple counts of women trained does not provide them with meaningful information, they agree to operate, and act, aligned with a dominant focus on the achievable—which means they accept a focus on simple counts as the only way to measure women.

Hit the target, but don't support the intention of the target

When implementing organizations set up a field office, hire staff, and begin programming, they bring project management practices and tools into new social contexts, interacting with diverse peoples. These practices, tools, and project goals interact with local gender dynamics, spinning out unknown outcomes. The local circumstance at the site of any development project will be different. In this context, there are vast discrepancies between men and women with respect to literacy, numeracy, formal education obtainment, access to money, decision-making authority, to name but a few areas. When gender-related programming is delivered by projects, it intersects and interacts with these dynamics, affecting the ability of organizations to reach their indicator targets for trainings. Professionals at all offices noted the difficulty in getting women to attend trainings and described a workplace unable to slow down to address women's differential needs.

A project leader started by describing the lack of interest from the donor, "Nobody at IDAD says 'Hey listen guys, we want you to do that [engage more women], and we understand that it does take more time and money.'" He continued, noting the reasons why the project couldn't easily hit the targets of women's attendance, "because if you want to get 25% women, and even for many activities we didn't find women. Not because they're

not there, but because they didn't have the required pre-education to enter." One response to this would be to reach out to women to build their skills and ability to engage in the project. The project leader followed with, "So if I want to get them there than I need to pre-invest in them and that takes more time away from the project, not from the project but it takes time from the targets that are not designed for gender." He described a willingness and an interest to do this extra work, but then returned to the donor interests as a barrier:

> So as long as IDAD doesn't really translate the gender policy into how that relates to targets you will always have the problem of 'Okay let's make sure the activity goes on because I can't afford to wait'. We don't have 25% women and that's even planning for 50% women, but I can't wait because I need to reach my targets.

In this case, engaging women takes the backseat to hitting *other* requirements because those targets are more directly linked to the overall project objective according to the donor. Several professionals described how women's participation was undermined by low literacy, numeracy, or having participation requirements as too difficult for women. Numerous midline and endline evaluations of *Farming Sustenance* projects state that women's participation was limited due to "social norms barriers." This understanding was confirmed by gender advisors in headquarter offices. Blaine, an IDAD Gender Expert, noted even the more mundane aspects of improving women's involvement in programs, "And the thing is that with gender, the issue is often something very banal like the logistics of 'we need a babysitter,' the women need an advance for the cost of transportation, you

need to start by asking the husband, and we're not asking that so then we give up." Blaine mentioned these examples in the context of endorsing targets, yet recognizing that their achievement is not always supported:

> It's always good to have targets! The difficulty is balancing the low hanging fruit with more ambitious items and those are the items that we need support for. So for example I was in the field and there was an implementing partner and they got dinged because they didn't hit their targets and their targets were not very ambitious, it was just showing up… So then why are women not showing up? Well, we found out that they needed childcare so the implementing partner went and asked the mission for extra money to pay for child care and the mission said No! So clearly, we expect a result but we are not supporting that result.

Field-based gender advisors did state that project activities occurred with a gender lens in mind, i.e. which days of the week were easier for women, what times of day, some projects did provide childcare. In one circumstance an implementing partner utilized gifts of backpacks to help get women to come to trainings so that the project met its sex-disaggregated target of women's participation - Number of individuals who have received short-term agricultural training. Beth, an external evaluation consultant highlights how targets can serve as a double-edged sword, and encourage reductive thinking about women's engagement:

> Targets obviously can help articulate priorities so they can send clear signals of where priorities lie which is where they can be really helpful. The problem lies when they get reduced down to just targets that can be counted and then very complex dynamics get reduced to how many people show up to a meeting. I think the example of providing backpacks to women so you hit the target of 25%. It's a great example, I mean you've met your target but have you really understood the underlying dynamics of what is drawing people to participate in a program and what is actually going to incentivize them to make the kind of changes that you're hoping for?

Other headquarter-based professionals noted that projects "count women because they're told to, not because they understand it." Above we see targets encouraging quick and fast methods to get women in the door, but targets can also be misunderstood to suppress women's involvement. The issue of translation was often raised by donor country professionals, while the professionals in the field spoke about confusion and being unsure of what and how they should work on items. The ways in which indicators can create confusion was described by Tiffany, a donor country gender professional previously employed by an implementing partner. She described a project situation where local female participation was very high in a Middle Eastern country but when the local implementing partner heard the target for women's participation, they told women to stop coming in order to decrease their numbers to match the target. When the gender advisor visited the country, the local women accused the project of holding back women:

> That is the best example of why you can't have targets all by themselves, you have to explain to the people implementing, they need to understand the targets, why there are targets, and be based in some context. And we need to support it so that filling the quota isn't just checking a box. So often times we talk about women's leadership in cooperatives and we need to support them to *become* leaders and if that means that we need to think about training, people around them to support them, changing the systems or structures so they're able to be leaders, provide child care, land ownership, etc.

The observation that development projects work to hit targets, without supporting the underlying idea that the indicator was meant to capture was confirmed by many. A frontline gender advisor raised her voice as we spoke, "Having the target is very much important! Because we've been trying to go that way and even evaluate ourselves, but '20% women leaders' is mandated… We failed to do that! We failed to do that! And that is given by IDAD for us to make 20% female leadership." She slammed her fist down on the table. This gender advisor wanted to keep the target, knowing that they had not accomplished it, but she was looking for something deeper, "We go, give them training, they elect females and then after some time when we check she's not there because she has so many things to do." The preponderance of hitting a target, or striving to hit a target, without supporting the target was commonplace.

I have demonstrated a consistent skepticism, from project-based gender advisors and managers to IDAD staff both abroad and in headquarters, about the role targets play around

gender programming. There is a strong sense of disconnect between IDAD's stated interest in investing in women's participation which does not appear properly supported at the project level, undermined by a focus on merely achieving indicators and targets, without addressing the underlying gender dynamics that would result in more sustainable development.

CONCLUSION

I add value to studies of measurement and its effects by documenting quantified evaluation as an institutional logic operating within development organizations. This logic serves to demarcate development professional and organizational actions in ways that foreclose attempts at incorporating a gendered lens into development programming. Despite the emergence of randomized control trials forming the new gold standard in development measurement, the international community has come to accept simple ways of measurement under the "results agenda." This "dataism" has created a self-fulfilling imperative: "create indicators that are measurable and then require that social justice work be directed, even pigeonholed, to achieve progress on said indicators." (Liebowitz and Zwingel 2014: 363). More sophisticated ways of measuring have similarly come under scholarly critique: Kabeer (1994) argues that development economics renders women's experiences unintelligible by using supposedly neutral measurements such as cost-benefit ratios and Bose (2015) argues a composite index to capture gender inequality in the United States would never be accepted as worthy as it would obscure and homogenize inequalities across race, class, and other variables.

This chapter documents the structures and practices that interlock to create an organizational reality that marginalizes gender-related programming at every level. Although gender has been enshrined into IDAD bureaucratic policies, the professionals responsible for its incorporation have no legal authority to force compliance. This sidelining means that gender becomes negotiated every day between development professionals—engaging interpersonal gender dynamics. This circumstance means that an organizational imperative—"to incorporate gender"—is *individualized* and left to negotiation by staff members. The odd combination of "strongly encouraged incorporate of gender" with no legal authority to require its inclusion within *evaluative structures* results in *practices* that development professionals cobble together: tack gender on, attempt to engage men counterparts, dismiss the validity of quantitative indicators but comply with reporting requirements, acquiesce that gender is "undoable" in a way that can be counted, and focus on simple quotas without supporting the intention of the quota.

Throughout the aid chain, gender programming simply does not *matter* to the majority of development professionals. Although gender policies and occupational positions have been created and allotted, there is no legal requirement to include gender. This lack of authority means that gender-related efforts are left to the personal prerogative of development professionals at all levels of the development aid chain. I demonstrate that the quantified evaluation system, set within a bureaucracy, is not a single source of marginalization, but spins out practices that at every point in the aid chain serve to marginalization gender programming. Professionals interested in gender efforts are left with no legitimate pathways to make claims for such investment unless they link into the power of the quantified evaluation system. One IDAD headquarter based professional, a

woman and a gender specialist, linked the absurdity of requiring "more data" to justify an investment in gender efforts in development programming, quipped dryly, "As if 50% of the population isn't enough." Professionals who are personally committed to gender-related efforts *feel like* activists because they go to work and actively advocate for the inclusion of gender in a system that marginalizes gender through structures and practices that reflect the logic of quantified evaluation.

Future directions for analysis could include deeper engagement with organizations scholarship. This could include assessing the moral dimensions of quantified evaluation (Cloutier and Langley 2013), analyzing the messages between development organizations (Lammers 2011), or searching for development cases where non-quantified evaluation practices were allowed to thrive (Hirschman et al. 2016). Archival research on evaluation-specific contracts or additional interviewing with professionals responsible for new contracts would provide a richer picture of quantified evaluation as a logic. Another avenue for potential exploration is that of decoupling. Development professionals engage in evaluative activities despite their acknowledgement that time could be better allocated elsewhere. Therefore, a decoupling lens that utilized both a policy-practice (Meyer and Rowan 1977) and means-ends (Bromley and Powell 2012) framework could assist in understanding professional's compliance. For the purposes of this chapter, I narrowed my focus to the structures and practices that, ultimately, marginalize gender programming, yet the above could be fruitful pathways for additional analysis.

Perceiving of quantified evaluation as an institutional logic highlights how difficult it is for professionals and organizations to enact alternative practices that are not valued by others in their milieu. But institutions must be maintained in order to manifest as logics

into structures and practices that help professionals make sense of the world around them. And with social reproduction comes the possibility for new or evolved institutional logics. Although development professionals substantiate quantified evaluation as a logic through the execution of such practices, its dominance as a logic is not certain. Professionals speak openly about their frustration, sometimes growing into anger, about simple measures. This research captured the strength of the practices that reify quantified evaluation as a logic, but simultaneously captured the frustration and lack of buy-in amongst professionals around the very indicators that they themselves calculate and report. The most exciting line of additional inquiry would be to consider the power inherent in development professionals lack of shared meaning and contested validity of these indicators.

This chapter documents why gender-related efforts—and other transformative social processes that are difficult to meaningfully measure—stand to lose under a focus on quantitative metrics. Interrogating institutional logics documents why professionals and organizations behave in particular ways. By going inside organizations to the heart of their evaluation practices, we reemerge with a better understanding of the data that development knowledge-making rests upon. Simply put, the ability to demonstrate results, even simple counts, increasingly matter for development organizations vying for reputation and the ability to win the next contract. Analyzing *inside* development organizations highlights the difficulty of incorporating gender efforts under the logic of quantified evaluation. However, institutional logics operate at higher and higher levels. Understanding development organizations as in relation to one another in a competitive field, oriented toward quantified evaluation, demonstrates that feminist attention must not be on "getting gender in" to quantified evaluation systems, but to question the very foundations of

quantified evaluation as an institution. This turns our analytical eye towards the underpinnings of quantitative ways of knowing and why the act of evaluation has become so powerful in modern organizational life. Further, although quantified evaluation may marginalize efforts for gender equality, what might occur inside development organizations if gender, itself a higher order structure, did not pattern the beliefs and behaviors of individuals?

Chapter 5 Conclusion

The major contribution of this research is to connect scholarly advancements in quantification, commensuration, and reactivity studies to that of development studies. Rather than analyze these processes at the international level, I sought to create an understanding of these processes from the "shop floor" up: centering the mandatory measurement systems of development projects, implemented by professionals, serving organizations, contracted by donors, who are accountable to legislative bodies. Bringing the sociology of organizations literature into conversation with development studies provides a sharper focus on the institutional constraints that development bureaucracies face in their efforts to externalize uncertainty, including constraints on knowledge production, the realities of accountability mechanisms across distance, and the unusual position of organizations tasked with achieving complex social outcomes (Mowles, 2007; Swidler and Arditi, 1994; Watkins, Swidler & Hannan, 2012).

Through three empirical chapters developed as article manuscripts, I used theories associated with the sociology of organizations (Acker 1990; 2002; 2006; Espeland and Stevens 1998; Espeland and Sauder 2007; Friedland and Alford 1991) to advance scholarly understandings of development organizations and their inclusion, or lack thereof, of gender-related efforts in the greater context of the "results agenda." In summary, I argue that quantified evaluation has become an inescapable and coordinating logic toward which all professional labor is oriented. In the words of one evaluation director I spoke to: "We're not running projects for the project anymore, we run the project for the results!" The orientation to results is cultivated by the greater system of donor resource allocation which

requires the demonstration of prior delivery performance to build organizational reputations.

Most notably, utilizing a sociology of organizations approach more sharply identifies the underlying logics of development actors, including the distinct roles and motivations of each actor type, be it "buy it," do it," and "check it" organizations and tightly focuses on the survival imperatives faced by development organizations. Development scholarship often focuses on projects, but if we turn our analytical eye inside development organizations, we see an organizational milieu far less concerned with the stated goal of development project efficacy and instead a milieu preoccupied with organizational survival. Recasting development organizations in this light demonstrates that "development" is not the goal of organizations—they work to secure their organizational ability to win future funding in a competitive field of multiple actors.

Within this context of survival imperatives, sociology of organizations sharpens analytical attention to the role of knowledge-making practices in the development sector's reproduction. Sociology of organizations demonstrates that metrics help solve the principal-agent dilemma and, in the 21st century, provide mechanisms of managerial control across distance through a new form of governance unleashed by the growing societal demands for organizations to provide evidence of their activities: that of governance-by-indicator. The ability to deliver such evidence has become a core component of organizational reputation-making. By looking at how knowledge production is bound up inside organizational relationships and the survival imperative, we see there is no escaping these dynamics. Such dynamics provide the overarching context that knowledge production takes place within, shaping and reshaping knowledge production to

fit organizational needs. Understanding the role of reactivity in development knowledge-making calls the bluff on M&E systems as capable of producing evidence of "what works" in the development sector—raising ethical and moral dilemmas about who is the true beneficiary of the development industry. Evaluation systems serve as export routes to new locations around the globe, representing the organizational infrastructures of which they are a part and the global relationships that they signify. By ensuring organizational survival through knowledge-making divorced from project realities, these global relationships are made and remade anew failing to address the global world order and instead reproducing it. Further, evaluation systems foster the adoption of performance metrics around the globe by diverse professionals and promote the idea that "measurability" is somehow more worthy of investment.

In this context, development knowledge production, discursively presented as an empirical exploration of project efficacy with an intent to improve, is better understood as an edifice not connected to project realities. Any critical questioning of development knowledge production is easily met with attacks against the morality or intention of the questioner: Oh, so you want to be sloppy with taxpayer dollars? You don't care about figuring out what is the best way to deliver development or find what actually works to help people? But this line of attack leverages data with the presumption that development knowledge production is somehow accurate and empirical, ignoring the influence of organizational survival imperatives on what is accepted as "knowledge." In response, I document that reactivity—professional and organizational reactions to measurement—is present within the international development sector, providing explanatory power to *how* and *why* development measurement documents success while projects continue to fail

around the world. Scholarship has noted that development knowledge-making "creates success" (Watkins et al 2012) yet the mechanism by which this occurs has been lacking: reactivity is the social process that drives development organizations to "create success" using measurement practices.

In addition to the broad insights gained through a sociology of organizations perspective noted above, this research demonstrates the utility of organizations scholarship to development studies. In Chapter 2, I established reactivity as a phenomenon embedded in development organizations, discussed above. Further, I document how social life within organizations takes shape around performance metrics; harnessing professional agency toward the accomplishment of project performance metrics in support of relationship-building with the managerial tier above. At the same time, metrics serve as comparative pressures, negatively sanctioning any professional who attempts to break out of the shared understanding that evaluation data is a knowledge-production edifice. In Chapter 3, I use gendered organizations to document the contentious and gendered nature of the development workplace. In this space, gender advisors promote gender mainstreaming yet do so as interactional foot soldiers due to their structural marginalization in the workplace. This occupational role was created by Global North organizations to facilitate their goals of "gender mainstreaming" yet this intention is not shared by the "technical" development professionals from Global North and South countries alike. And, in Chapter 4, I demonstrate that hard-to-measure processes, such as women's empowerment, leadership, or capacity building, remain peripheral to development projects precisely because of their difficulty to be rendered into a bureaucratic measurement logic. Using institutional logics as a framework clarifies how bureaucratic inclusion can result in persistent marginalization

due to conflicting logics, and documents how socially transformative projects are doomed to fail before they even begin. In short, organizations are unlikely to risk their reputations by agreeing to work on transformative processes that are not easily demonstrated as "done" to donors. Sociology of organizations theories sharpen attention to the underlying survival imperatives faced by development organizations, perhaps historically obscured because of understanding development organizations as moral and values driven. Sociology of organizations theories helps pull back the mask of morality by positioning development organizations as self-interested.

This research also builds sociology of organizations using the case of complex organizations in the international development sector. In Chapter 2, I documented the existence of reactivity—reactions to measurement—where extant theory predicted it would not be found (Davis et al 2012; Rottenburg et al 2015; Espeland and Sauder 2007): as reactions to simple counts like "number of people trained." Understanding that even simple counts, when nested within comparative regimes, elicit reactivity broadens the empirical terrain of extant theory and demonstrates reactivity will occur at more meso- and micro-sites than previously predicted. This finding invites additional research on understanding how reactivity operates cross-culturally, eliciting decoupling or ceremonial adoption in cultures that are less attentive to bureaucratically proving organizational performance.

In Chapter 3, I demonstrate that broader studies of quantification and rankings must attend to gender as an analytical category and inequality regimes within organizations. I document how professional placement within inequality regimes impacts their reception of metrics, effecting professional resistance and adoption of metrics in the workplace. I show that professionals see and understand the power of metrics in contexts that prioritize them.

With this recognition, professionals then attempt to wield metrics as bureaucratic tools to achieve their aims. This recasts metrics in a starring role in future organizations research on workplace politics. This finding invites much more nuanced research about how metrics impact which professionals, why, in what ways, and with what effect on constellations of workplace power. It also presents metrics as a core form of managerial power, one that is likely to grow as workplaces become increasingly digitized and surveilled.

In Chapter 4, I document the conflict between organizational gender frames and personal gender frames (Ridgeway 1997; 2009) by examining the bureaucratic inclusion, or lack thereof, of gender-related development efforts. The bureaucratic mandate of gender mainstreaming does not actually result in gender mainstreaming as an organizational imperative, which leaves gender-related efforts to be considered and resolved between development professionals on an individual basis. Bureaucratic mandates around quantitative measurement have been powerful forces within organizations, as seen in the other chapters, yet gender-related metrics do not pulse with this same power. This invites future research on organizational attempts at equalities, such as affirmative action or gender mainstreaming, in an era of evidence-based decision and policy making. Lastly, I demonstrate how institutional logics can be embedded at the interactional register within organizations, inviting case-based research on institutional logics.

But the future of evaluation is poised for disruption by technological advances—disruptions that return to the fundamental question of trust between people who do not know each other yet interact across geographic distance.

The Future of Evaluation

In 2017, as I walked into an interview with an evaluation specialist in an East African country, he motioned for me to sit down in a chair in front of several stacked boxes. Only one box had been opened, revealing neat rows of smaller boxes. I peered inside: tablets. Not medical tablets to address Vitamin A deficiency, which is common in nutrition projects looking for quick wins, but portable, electronic, touchscreen tablets. The evaluation specialist told me there were 400 tablets in all—to be rolled out across the country to collect M&E data electronically. I thought back to another interview with staff of an organization sub-contracted for data collection where the staff described their "coding guy" in India and how quick and fast he is able to deliver digitized data collection tools. The organization sends him specifications on a data collection tool and he builds the instrument into "pages" on tablets that enumerators move through in real time as they complete interviews. These "pages" force validation on particular fields. For example, "household type" field is standardized, presenting the enumerator with drop downs for male-headed or female-headed, obscuring any household type that was not coded. Technology has become an omnipresent part of measurement in the development sector. These new efforts seek to perfect development measurement under the assumption that inadequacies are technical in nature. Once evaluation systems have been digitized, new possibilities for automation become possible.

Just as evaluation systems were built up and created massive structures of trust, or what served to *suffice* as trust to manage the principle-agent problem, they morphed into structures of reputation, cultivating the motivations of multiple development actors to engage in "lies" because of material benefits. Similarly, blockchain is a new technology that has generated new organizational actors in the forms of "innovation labs" opened or

funded by the largest development donors, including the World Bank, USAID, GIZ, DfID, Gates Foundation, UNICEF, and United Nations Operations. Promoters say the technology itself is poised to transform the locus of trust the development sector (Shreves 2017). The creation of such "innovation labs" within international development organizations suggests something big is afoot. But what is blockchain? How does it work? And why are development organizations interested in this technology? How do they envision using these technologies for social good?

Blockchain technology as the "trust machine"

In 2015, *The Economist* headlined an issue with a deeply sociological statement: calling blockchain a "trust machine." This was preceded by a white paper in 2009 by avatar Satoshi Nakamoto titled "Bitcoin: A peer to peer electronic cash system." In it, Nakamoto outlined how to use cryptography and math to build trust without intermediaries in a virtual online environment. Using the internet to connect individuals and code to execute record keeping through an addition-only system known as blockchain, people could build trust without a central authority to validate identities or transactions. Nakamoto's main use case was a digital currency system, commonly known as cryptocurrency, that no longer requires the use of banks, credit card companies, and entities like Western Union and PayPal. In his proposed virtual world, third party intermediaries become *dismediated*. In the words of *The Economist* (2015): "The real innovation is not the digital coins themselves, but *the trust machine that mints them*—and which promises much more besides" (emphasis mine). This "much more" is what excites the development sector. The United Nations Blockchain Group (n.d.) defined it as:

> A chain of ledgers distributed over a network that is robust in its unstructured simplicity, that is computationally impractical for an attacker to change, that no one can control but everyone can view, and that can connect an individual with another individual without giving up trust to a third party or a central authority.

In smaller societies people were able to trade or barter directly because they trusted the interaction and knew they would be paid in kind. However, as distance in trade grew, we invented institutions to help us trade and exchange, building entire global economies. The most recent iterations of these institutions, such as Uber, Airbnb, and Amazon, are digital marketplaces and platforms that help us exchange a form of value (WIRED 2017). Businesses today, such as the diamond trade, are interested in "blockchaining" their supply chains to help provide stronger mechanisms of certification to customers (Calvão 2019). Put simply, blockchain technologies offer a new way of doing business: instead of a *company* in the middle helping someone make a transaction, there is *code*. Therefore, blockchains create consensus amongst parties that do not need to trust each other but need to trust the code that sits between them.

Yet in order to trust code, do we need to understand it? As a society we interact continuously with technology yet do not understand how or why it works. Where does our trust *come from*? Do we trust something because people we trust start to use it? Or because a lot of people start to use it? Or because we understand the code itself? Trust requires not only an initial buy in, but an ongoing maintenance of that trust. Just as trust can be

generated, it can also degrade. Trusted intermediaries will be *dismediated* by blockchain technologies *if* we place our trust in code. However, are we just pushing mediation elsewhere and, in doing so, pushing trust into new spaces? In this new digital utopia, the people who understand and write the code are best poised to become new intermediaries. Laypeople would need to buy technologists' services to either do the work of initiating us into these new technologies or building lay skillsets to engage in this new virtual world. Although technology may reconfigure power, it does not dismantle its existence.

The Development Sector and Blockchain

> To the extent that international NGOs function as guarantors of trust–trust that the funds donated will be used for an appropriate purpose, trust that the aid has been given to the right beneficiaries, trust that the development work that was contracted for was done on time and as specified–then NGOs too are poised for disruption (Shreves 2017).

Many development organizations act *as intermediaries*. If implementing organizations merely distribute aid and manage that process, they themselves could be dismediated. But by creating "innovation labs," development organizations are busy working to ensure they stay ahead of the blockchain race. Instead of being dismediated, organizations now work to showcase to donors their ability to utilize and execute blockchain technologies in aid of global development. But digitalization merely becomes another layer of how "haves" are able to utilize resources to widen the gap between them and global "have nots" (Schia 2018).

The Stanford Center for Blockchain Research manages a spreadsheet which records "Blockchain for Social Good" projects. As of November 2019, they have catalogued 240 different projects. These idea of "social good" being generated by blockchain technology is showcased by headlines such as "Blockchain Against Hunger: Harnessing Technology in Support of Syrian Refugees" and claims that blockchain can address "the roots of hunger" (World Food Programme 2017). In 2018, an academic, blockchain coder, and blockchain user teamed up to create a special issue on blockchain *for* global development (Lubin, Anderson, and Thomason 2018) demonstrating a growing perception that blockchain can be "harnessed" for social good. Kshetri (2019) stated the core use cases to utilize blockchain to address poverty in the Global South are: promoting transparency, reducing costs of property registration, promoting efficiency in business to business trade, reducing costs in international payment systems, insurance and risk management, and banking. For example, BanQu, a blockchain startup, has created a platform for the "unbanked" allowing them to record their transaction in digital currency, purchase goods, and "prove their existence in global supply chains" (Galen et al 2018). A world in which the unbanked skip over state-backed currency in favor of digital currency is increasingly envisioned. Farmers have historically engaged in the required paper-based forms of certifying trust (Seshia Galvin 2018), will the rural poor adopt blockchain technologies in the manner envisioned by donor organizations and promised by implementing organizations?

But inequalities are persistently present at the same time blockchain technologies are under development. Gender inequalities are not only stark in the greater tech world (Wachter-Boettcher 2017), but blockchain annual meetings have been held at strip clubs

and few women are engaged in research and development (Bowles 2018; Koffman and Rosenzweig 2018). Early blockchain rhetoric positions poor rural women as a key "winner" of blockchain adoption. In other words, development organizations position poor rural women as the upcoming beneficiary of blockchain-based development projects. Blockchain is presented as the missing technical solution capable of cutting out existing social relations:

> Money is distributed *directly* into user's accounts which empowers women to use their own accounts, not their husbands, fathers, or brothers. And studies show that when women manage money, they are more likely than men to pay for food, healthcare, and education for their families (Dunleavy 2016).

This rhetoric continues the dominant trend to rationalize investment in women as "smart economics" (Calkin 2015, Collins 2016, Chant and Sweetman 2012), while ignoring how and why gender dynamics will impact blockchain adoption (gender differences in mobile phone ownership as an example). Like the envisioned scenario above, the World Economic Forum's "What is Blockchain" video (2016) visually presents women farmers as the core beneficiaries of this breakdown—featuring women standing in rural areas holding mobile phones receiving digital money from donors and family members abroad.

Blockchain promoters focus on the validity of blockchain records internally, yet blockchain technology does nothing to address external validity: "garbage-in garbage-out" remains unsolved. Nonetheless, donors are increasingly funding projects which utilize such

technologies, meaning that the development organizations which are best able to utilize current resources to capture this new realm of donor investment stand to advance their project portfolios. Today, the development sector promotes blockchain with the imagery of engaging women without a recognition of how current gender inequalities will shape the future of blockchain around the globe. In response, my future research will continue with a gendered analysis of the development sector, engaging the shifting forms of trust and reputation in an era premised on the demonstration of results.

Quantification and Perceptions of Justice

In closing I want to return to what Espeland and Stevens (2008) termed "an ethics of quantification." Since numbers are perceived of as rational and university, quantification facilitates the appearance of "existence" as that which is *measurable*. In this historical moment, fairness has become understood through the application of "objective" numbers. Digitizing these quantified systems only makes them appear more objective. Resources, status, and reputation increasingly accrue to those who can best collect data and transform it into aesthetically pleasing charts and graphs—all of which require large capital investments and technical resources (Espeland and Stevens 2008). The power of numbers is best seen when rendered through bureaucracies which allocate funds, increasingly occurring in digital forms: the Chinese government is rolling out the Social Credit Score (Liu 2019), workplaces track the habits of their employees, sanctioning "unhealthy behaviors" and offering customized health care plans (Ajunwa, Crawford, and Schultz 2017), algorithms are used to determine welfare recipients (Eubanks 2018; O'Neil 2016) and big data informs policing tactics (Brayne 2017). As our social world is increasingly

bound up in numbers, scholarship must help frame the public debate about what numbers are *doing* as they enter organizational contexts. In agreement with Espeland and Stevens (2008), we can no longer understand efforts for justice without understanding the role of numbers. As such, scholarship in the 21st century must attend to how the use of numbers in diverse forms—from aggregating websites to encoding trust in blockchains—reconfigures gender relations specifically and global relationships more broadly.

Bibliography

Acker, Joan. 1990. Hierarchies, jobs, bodies: A theory of gendered organizations. Gender & Society 4(2): 139–58.

Acker, Joan. 2002. The Future of 'gender and organizations': Connections and boundaries. Gender, Work & Organization 5(4): 195–206.

Acker, Joan. 2006. Inequality regimes: Gender, class, and race in organizations. Gender & Society 20(4): 441–64.

Acosta-Belén, Edna and Chrisine Bose. 1990. "From Structural Subordination to Empowerment: Women and Development in Third World Contexts." *Gender & Society* 4(3):299–320.

Adams, Vincanne. 2016. Metrics: What Counts in Global Health. North Carolina: Duke University Press.

Ajunwa, Ifeoma, Kate Crawford, and Jason Schultz. 2017. "Limitless Worker Surveillance." California Law Review 105:735–76.

Babb, Sarah. 2020. *Regulating Human Research: IRBs from Peer Review to Compliance Bureaucracy*. Stanford University Press.

Baines, Donna. 2010. Gender mainstreaming in a development project: Intersectionality in a post-colonial un-doing? Gender, Work & Organization 17(2): 119–49.

Banks, Nicola, David Hulme, and Michael Edwards. 2015. "NGOs, States, and Donors Revisited: Still Too Close for Comfort?" World Development 66:707–18.

Barman, Emily. 2016. Varieties of field theory and the sociology of the non-profit sector. Sociology Compass 10(6): 442–58.

Bebbington, Anthony. 2005. "Donor–NGO Relations and Representations of Livelihood in Nongovernmental Aid Chains." World Development 33(6):937–50.

Beck, Erin. 2017. How Development Projects Persist: Everyday Negotiations with Guatemalan NGOs. Durham: Duke University Press Books.

Béné, Christophe, Fahim S. Chowdhury, Mamun Rashid, Sabbir A. Dhali, and Ferdous Jahan. 2017. "Squaring the Circle: Reconciling the Need for Rigor with the Reality on the Ground in Resilience Impact Assessment." World Development 97:212–31.

Benschop, Yvonne, and Mieke Verloo. 2006. Sisyphus' sisters: Can gender mainstreaming escape the genderedness of organizations? Journal of Gender Studies 15(1): 19–33.

Berman, Elizabeth Popp, and Daniel Hirschman. 2018. The sociology of quantification: Where are we now? Contemporary Sociology 47(3): 257–66.

Berman, Elizabeth Popp. 2012. "Explaining the Move toward the Market in US Academic Science: How Institutional Logics Can Change without Institutional Entrepreneurs." *Theory and Society* 41(3):261–99.

Besharov, Marya L. and Wendy K. Smith. 2014. "Multiple Institutional Logics in Organizations: Explaining Their Varied Nature and Implications." *Academy of Management Review* 39(3):364–81.

Best, Jacqueline. 2017. "The Rise of Measurement-Driven Governance: The Case of International Development." *Global Governance* 23(2):163–81.

Bose, Christine E. 2011. “Eastern Sociological Society Presidential Address: Globalizing Gender Issues: Many Voices, Different Choices1.” *Sociological Forum* 26(4):739–53.

Bose, Christine. 2015. Patterns of global gender inequalities and regional gender regimes. *Gender & Society* 29(6): 767–91.

Bowles, Nellie. 2018. “Women in Cryptocurrencies Push Back Against ‘Blockchain Bros.’” *The New York Times*, February 28.

Bowman, Nicholas A. and Michael N. Bastedo. 2009. “Getting on the Front Page: Organizational Reputation, Status Signals, and the Impact of U.S. News and World Report on Student Decisions.” Research in Higher Education 50(5):415–36.

Brandtner, Christof. 2017. “Putting the World in Orders: Plurality in Organizational Evaluation.” Sociological Theory 35(3):200–227.

Brass, Jennifer N. 2012. “Why Do NGOs Go Where They Go? Evidence from Kenya.” World Development 40(2):387–401.

Brayne, Sarah. 2017. “Big Data Surveillance: The Case of Policing.” *American Sociological Review* 82(5):977–1008.

Bromley, Patricia and Walter W. Powell. 2012. “From Smoke and Mirrors to Walking the Talk: Decoupling in the Contemporary World.” *Academy of Management Annals* 6(1):483–530.

Burawoy, Michael. 2000. Global Ethnography: Forces, Connections, and Imaginations in a Postmodern World. Berkeley: University of California Press.

Calkin, Sydney. 2015. “Feminism, Interrupted? Gender and Development in the Era of ‘Smart Economics.’” *Progress in Development Studies* 15(4):295–307.

Calvão, Filipe. 2019. “Crypto-Miners: Digital Labor and the Power of Blockchain Technology.” *Economic Anthropology* 6(1):123–34.

Campbell, Marie L. and Katherine Teghtsoonian. 2010. “Aid Effectiveness and Women’s Empowerment: Practices of Governance in the Funding of International Development.” *Signs* 36(1):177–202.

Chant, Sylvia and Caroline Sweetman. 2012. “Fixing Women or Fixing the World? ‘Smart Economics’, Efficiency Approaches, and Gender Equality in Development.” *Gender & Development* 20(3):517–29.

Chant, Sylvia. 2016. “Galvanizing Girls for Development? Critiquing the Shift from ‘Smart’ to ‘Smarter Economics.’” *Progress in Development Studies* 16(4):314–28.

Cloutier, Charlotte and Ann Langley. 2013. “The Logic of Institutional Logics: Insights From French Pragmatist Sociology.” *Journal of Management Inquiry* 22(4):360–80.

Cole, Georgia. 2018. “How Friends Become Foes: Exploring the Role of Documents in Shaping UNHCR’s Behaviour.” *Third World Quarterly* 0(0):1–17.

Collins, Andrea M. 2016. “‘Empowerment’ as Efficiency and Participation: Gender in Responsible Agricultural Investment Principles.” *International Feminist Journal of Politics* 18(4):559–73.

Colyvas, Jeannette. 2012. “Performance Metrics as Formal Structures and through the Lens of Social Mechanisms: When Do They Work and How Do They Influence?” *American Journal of Education* 118(2):167–97.

Connell, Raewyn, Barbara Fawcett, and Gabrielle Meagher. 2009. Neoliberalism, new public management and the human service professions: Introduction to the special issue. *Journal of Sociology* 45(4): 331–38.

Cooley, Alexander and James Ron. 2002. "The NGO Scramble: Organizational Insecurity and the Political Economy of Transnational Action." *International Security* 27(1):5–39.

Cornwall, Andrea and Althea-Maria Rivas. 2015. "From 'Gender Equality and 'Women's Empowerment' to Global Justice: Reclaiming a Transformative Agenda for Gender and Development." *Third World Quarterly* 36(2):396–415.

Cornwall, Andrea. 2000. "Making a Difference? Gender and Participatory Development." *IDS Discussion Paper* 378.

Dar, Sadhvi and Bill Cooke, eds. 2008. The New Development Management: Critiquing the Dual Modernization. New York: Zed Books.

Davis, Alexander. 2018. "Toward Exclusion through Inclusion: Engendering Reputation with Gender-Inclusive Facilities at Colleges and Universities in the United States, 2001-2013." *Gender & Society* 32(3):321–47.

Davis, Kevin E, Fisher, Kingsbury, and Merry. 2012. *Governance by Indicators*. Oxford University Press.

Davis, Kevin, Angela Fisher, Benedict Kingsbury, and Sally Engle Merry, eds. 2012. Governance by indicators: Global power through classification and rankings. New York: Oxford University Press.

Desai, Manisha. 2007. "The Messy Relationship Between Feminisms and Globalizations." *Gender & Society* 21(6):797–803.

DeVault, Marjorie. 1996. Talking back to sociology: Distinctive contributions of feminist methodology. Annual Review of Sociology 22(1): 29–50.

DFID. 2013. "DFID Evaluation Policy 2013." GOV.UK. (https://www.gov.uk/government/publications/dfid-evaluation-policy-2013).

DiMaggio, Paul and Helmut Anheier. 1990. "The Sociology of Nonprofit Organizations and Sectors." *Annual Review of Sociology* 16(1):137–59.

Douglas, Mary. 1986. *How Institutions Think*. 1st ed. Syracuse, N.Y.: Syracuse University Press.

Easterly, William. 2009. "How the Millennium Development Goals Are Unfair to Africa." World Development 37(1):26–35.

Ebrahim, Alnoor. 2002. "Information Struggles: The Role of Information in the Reproduction of NGO-Funder Relationships." Nonprofit and Voluntary Sector Quarterly 31:84–114.

Ebrahim, Alnoor. 2003. "Accountability In Practice: Mechanisms for NGOs." World Development 31(5):813–29.

Ebrahim, Alnoor. 2005. "Accountability Myopia: Losing Sight of Organizational Learning." Nonprofit and Voluntary Sector Quarterly 34(1):56–87.

Espeland, Wendy and Michael Sauder. 2007. "Rankings and Reactivity: How Public Measures Recreate Social Worlds." American Journal of Sociology 113(1):1–40.

Espeland, Wendy and Michael Sauder. 2016. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. New York: Russell Sage Foundation.

Espeland, Wendy and Mitchell Stevens. 1998. "Commensuration as a Social Process." *Annual Review of Sociology* 24(1):313–43.

Espeland, Wendy and Mitchell Stevens. 2008. “A Sociology of Quantification.” *European Journal of Sociology* 49(3):401–36.

Eubanks, Virginia. 2018. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York, NY: St. Martin’s Press.

European Commission. 2014. “Evaluation Matters - The Evaluation Policy for European Union Development Co-Operation.” International Cooperation and Development - European Commission. (https://ec.europa.eu/europeaid/evaluation-matters-evaluation-policy-european-union-development-co-operation-0_en).

Eyben, Rosalind, Irene Guijt, Chris Roche, and Cathy Shutt. 2015. *The Politics of Evidence and Results in International Development: Playing the Game to Change the Rules?* Practical Action Publishing.

Fejerskov, Adam Moe. 2017. “Contending Logics of Action in Development Cooperation: The Bill and Melinda Gates Foundation’s Work on Gender Equality.” *The European Journal of Development Research* 29(2):441–56.

Ferguson, Lucy. 2015. “‘This Is Our Gender Person: The messy business of working as a gender expert in international development.’” *International Feminist Journal of Politics* 17(3):380–97.

Friedland, Roger and Robert Alford. 1991. “Bringing Society Back In: Symbols, Practices, and Institutional Contradictions.” Pp. 232–63 in *The New institutionalism in organizational analysis*, edited by W. W. Powell and P. DiMaggio. Chicago: University of Chicago Press.

Friedland, Roger, John W. Mohr, Henk Roose, and Paolo Gardinali. 2014. "The Institutional Logics of Love: Measuring Intimate Life." *Theory and Society* 43(3–4):333–70.

Fukuda-Parr, Sakiko, Joshua Greenstein, and David Stewart. 2013. How should MDG success and failure be judged: Faster progress or achieving the targets? World Development 41: 19–30.

Galen, Doug, Nikki Brand, Lyndsey Boucherle, Rose Davis, Natalie Do, Ben El-Baz, Isadora Kimura, Kate Warton, and Jay Lee. 2018. *Blockchain for Social Impact: Moving beyond the Hype*. Stanford Center for Social Innovation.

GIZ. 2017. Evaluation Report 2017: Knowing What Works. Bonn and Eschborn, Germany: Deutsche Gesellschaft fur FIX U Internationale Zusammenarbeit (GIZ) GmbH.

Goldman, Michael. 2005. *Imperial Nature: The World Bank and Struggles for Justice in the Age of Globalization*. New Haven, Ct: Yale University Press.

Hammack, David and Steven Heydemann, eds. 2009. *Globalization, Philanthropy, and Civil Society: Projecting Institutional Logics Abroad*. Indiana University Press.

Harris, Michael and Bill Tayler. 2019. "Don't Let Metrics Undermine Your Business." *Harvard Business Review*, September 1.

Hesse-Biber, Sharlene. 2012. Feminist approaches to triangulation: Uncovering subjugated knowledge and fostering social change in mixed methods research. Journal of Mixed Methods Research 6(2):137–46.

Hirschman, Daniel, Ellen Berrey, and Fiona Rose-Greenland. 2016. “Dequantifying Diversity: Affirmative Action and Admissions at the University of Michigan.” *Theory and Society* 45(3):265–301.

Hirschmann, David. 2002a. “‘Implementing an Indicator’: Operationalising USAID’s ‘Advocacy Index’ in Zimbabwe.” *Development in Practice* 12(1):20–32.

Hirschmann, David. 2002b. “Thermometer or Sauna?: Performance Measurement and Democratic Assistance in the United States Agency for International Development (USAID).” *Public Administration* 80(2):235–55.

Hodgson, Dorothy. 2017. *Gender, Justice, and the Problem of Culture*. Bloomington & Indianapolis: Indiana University Press.

Hoey, Lesli. 2015. “‘Show Me the Numbers’: Examining the Dynamics Between Evaluation and Government Performance in Developing Countries.” *World Development* 70:1–12.

Høyland, Bjørn, Karl Moene, and Fredrik Willumsen. 2012. “The Tyranny of International Index Rankings.” *Journal of Development Economics* 97(1):1–14.

Hwang, Hokyu and Patricia Bromley. 2015. “Internal and External Determinants of Formal Plans in the Nonprofit Sector.” *International Public Management Journal* 18(4):568–88.

Ika, Lavagnon A. and Vasyl Lytvynov. 2011. “The ‘Management-Per-Result’ Approach to International Development Project Design.” *Project Management Journal* 42(4):87–104.

Ilcan, Suzan and Phillips. 2010. “Developmentalities and Calculative Practices: The Millennium Development Goals.” *Antipode* 42(4):844–74.

Johansen, Christina Berg and Susanne Boch Waldorff. 2017. “What Are Institutional Logics – and Where Is the Perspective Taking Us?” Pp. 51–76 in *New Themes in Institutional Analysis*, edited by G. Krücken, C. Mazza, R. E. Meyer, and P. Walgenbach.

Kabeer, Naila. 2005. “Gender Equality and Women’s Empowerment: A Critical Analysis of the Third Millennium Development Goal 1.” *Gender & Development* 13(1):13–24.

Kabeer, Naila. 2015. “Tracking the Gender Politics of the Millennium Development Goals: Struggles for Interpretive Power in the International Development Agenda.” *Third World Quarterly* 36(2):377–95.

Kanter, Rosabeth Moss. 1977. Men and women of the corporation. New York: Basic Books.

Kaplan, Robert S. and David P. Norton. 1992. “The Balanced Scorecard--Measures That Drive Performance.” *Harvard Business Review* 70(1):71–79.

Kim, Minzee and Elizabeth Heger Boyle. 2012. “Neoliberalism, Transnational Education Norms, and Education Spending in the Developing World, 19832004.” *Law & Social Inquiry* 37(2):367–94.

Kipnis, Andrew. 2008. “Audit Cultures: Neoliberal Governmentality, Socialist Legacy, or Technologies of Governing?” *American Ethnologist* 35(2):275–89.

Koffman, Tatiana and Jay Rosenzweig. 2018. “Blockchain: Mind the Gender Gap.” *Thrive Global: Community*. (https://www.thriveglobal.com/stories/39295-blockchain-gender-gap).

Kothari, Uma. 2005. “Authority and Expertise: The Professionalisation of International Development and the Ordering of Dissent.” *Antipode* 37(3):425–46.

Krause, Monika. 2014. *The Good Project: The Field of Humanitarian Relief NGOs and the Fragmentation of Reason*. Chicago: The University of Chicago Press.

Kshetri, Nir. 2017. “Will Blockchain Emerge as a Tool to Break the Poverty Chain in the Global South?” *Third World Quarterly* 38(8):1710–32.

Lammers, John C. 2011. “How Institutions Communicate: Institutional Messages, Institutional Logics, and Organizational Communication.” *Management Communication Quarterly* 25(1):154–82.

Lamont, Michèle. 2012. “Toward a Comparative Sociology of Valuation and Evaluation.” *Annual Review of Sociology* 38(1):201–21.

Leaders. 2015. “The Trust Machine.” *The Economist*, October 31.

Levy, Karen and Solon Barocas. 2018. “Refractive Surveillance: Monitoring Customers to Manage Workers.” *International Journal of Communication* 12(0):23.

Levy, Karen. 2016. Digital surveillance in the hypermasculine workplace. Feminist Media Studies 16(2): 361–65.

Lewis, David and David Mosse. 2006. Development Brokers and Translators: The Ethnography of Aid and Agencies. Bloomfield, CT: Kumarian Press.

Liebowitz, Debra and Susanne Zwingel. 2014. “Gender Equality Oversimplified: Using CEDAW to Counter the Measurement Obsession.” *International Studies Review* 16(3):362–89.

Lipsky, Michael. 2010. Streel-level bureaucracy: Dilemmas of the individual in public services. 30th anniversary expanded ed. New York: Russell Sage Foundation.

Liu, Chuncheng. 2019. Multiple Social Credit Systems in China. *economic sociology: the european electronic newsletter* 21(1):22-32.

Lubin, Joseph, Mally Anderson, and Bobbi Thomason. 2018. “Blockchain for Global Development.” *Innovations: Technology, Governance, Globalization* 12(1–2):10–17.

Marcus, George E. 1995. “Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography.” Annual Review of Anthropology 24(1):95–117.

Martin, Patricia Yancey. 2003. “Said and done" versus "saying and doing": Gendering practices, practicing gender at work.” Gender & Society 17(3): 342–66.

Merry, Sally Engle and Summer Wood. 2015. “Quantification and the Paradox of Measurement: Translating Children’s Rights in Tanzania.” Current Anthropology 56(2):205–29.

Merry, Sally Engle. 2011. “Measuring the World: Indicators, Human Rights, and Global Governance.” *Current Anthropology* 52(S3):83–95.

Merry, Sally Engle. 2016. The seductions of quantification: Measuring human rights, gender violence, and sex trafficking. Chicago: University of Chicago Press.

Meyer, John W. and Brian Rowan. 1977. “Institutionalized Organizations: Formal Structure as Myth and Ceremony.” *American Journal of Sociology* 83(2):340–63.

Miller, Peter. 2001. “Governing by Numbers: Why Calculative Practices Matter.” *Social Research* 68(2):379–96.

Moser, Caroline O. N. 1989. “Gender Planning in the Third World: Meeting Practical and Strategic Gender Needs.” *World Development* 17(11):1799–1825.

Mosse, David and Lewis, David. 2006. “Theoretical Approaches to Brokerage and Translation in Development.” Pp. 1–26 in *Development Brokers and Translators: The Ethnography of Aid and Agencies*. Kumarian Press.

Mosse, David. 1994. “Authority, Gender and Knowledge: Theoretical Reflections on the Practice of Participatory Rural Appraisal.” *Development and Change* 25(3):497–526.

Mosse, David. 2005. “Global Governance and the Ethnography of International Aid.” in *The Aid Effect: Ethnographies of Development Practice and Neo-liberal Reform*. Pluto Press.

Mosse, David. 2006. “Anti-Social Anthropology? Objectivity, Objection, and the Ethnography of Public Policy and Professional Communities.” *The Journal of the Royal Anthropological Institute* 12(4):935–56.

Mosse, David. 2011. “Introduction: The Anthropology of Expertise and Professionals in International Development.” Pp. 1–31 in *Adventures in Aidland: The Anthropology of Professionals in International Development*. Berghahn Books.

Mosse, David. 2013. “The Anthropology of International Development.” *Annual Review of Anthropology* 42(1):227–46.

Mowles, Chris, Ralph Stacey, and Douglas Griffin. 2008. “What Contribution Can Insights from the Complexity Sciences Make to the Theory and Practice of Development Management?” *Journal of International Development* 20(6):804–20.

Mowles, Chris. 2007. “Promises of Transformation: Just How Different Are International Development NGOs?” *Journal of International Development* 19(3):401–11.

Mowles, Chris. 2008. “Values in International Development Organisations: Negotiating Non-Negotiables.” *Development in Practice* 18(1):5–16.

Mowles, Chris. 2010a. “Post-Foundational Development Management—Power, Politics and Complexity. *Public Administration and Development.*

Mowles, Chris. 2010b. “Successful or Not? Evidence, Emergence, and Development Management.” *Development in Practice* 20(7):757–70.

Mowles, Chris. 2012. “Keeping Means and Ends in View—Linking Practical Judgement, Ethics and Emergence.” *Journal of International Development* 24(5):544–55.

Muller, Jerry. 2018. The tyranny of metrics. Princeton: Princeton University Press.

Nagaraj, Vijay Kumar. 2015. “‘Beltway Bandits’ and ‘Poverty Barons’: For-Profit International Development Contracting and the Military-Development Assemblage.” *Development & Change* 46(4):585–617.

Narayanaswamy, Lata. 2016. “Whose Feminism Counts? Gender(Ed) Knowledge and Professionalisation in Development.” *Third World Quarterly* 37(12):2156–75.

Natsios, Andrew. 2010. The Clash of the Counter-Bureaucracy and Development. Center for Global Development.

Nelson, Paul. 2018. *Primer on Blockchain: How to Assess the Relevance of Distributed Ledger Technology to International Development*. USAID.

Nogueira, Roberto Martinez. 1987. “Life Cycle and Learning in Grassroots Development Organizations.” *World Development* 15:169–77.

O’Neil, Cathy. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown.

Owczarzak, Jill, Michelle Broaddus, and Steven Pinkerton. 2016. “Audit Culture: Unintended Consequences of Accountability Practices in Evidence-Based Programs.” *American Journal of Evaluation* 37(3):326–43.

Parpart, Jane L. 2014. “Exploring the Transformative Potential of Gender Mainstreaming in International Development Institutions.” *Journal of International Development* 26(3):382–95.

Power, Michael. 1997. The audit society: Rituals of verification. Oxford: Clarendon Press.

Ransom, Elizabeth and Carmen Bain. 2011. “Gendering Agricultural Aid: An Analysis of Whether International Development Assistance Targets Women and Gender.” *Gender & Society* 25(1):48–74.

Ravallion, Martin. 2012. “Mashup Indices of Development.” The World Bank Research Observer 27(1):1–32.

Ridgeway, Cecilia L. 1997. Interaction and the conservation of gender inequality: Considering employment. *American Sociological Review* 62(2): 218–35.

Ridgeway, Cecilia L. 2009. Framed before we know it: How gender shapes social relations. *Gender & Society* 23(2): 145–60.

Ridgeway, Cecilia L., and Shelley J. Correll. 2004. Unpacking the gender system: A theoretical perspective on gender beliefs and social relations. *Gender & Society* 18(4): 510–31.

Roberts, Adrienne and Susanne Soederberg. 2012. “Gender Equality as Smart Economics? A Critique of the 2012 World Development Report.” *Third World Quarterly* 33(5):949–68.

Roberts, Susan M. 2014. “Development Capital: USAID and the Rise of Development Contractors.” *Annals of the Association of American Geographers* 104(5):1030–51.

Rottenburg, Richard, Sally Merry, Sung-Joon Park, and Johanna Mugler, eds. 2015. The World of Indicators: The Making of Governmental Knowledge through Quantification. Cambridge University Press.

Rottenburg, Richard. 2009. *Far-Fetched Facts: A Parable of Development Aid*. Cambridge: MIT Press.

Sauder, Michael and Hyunsik Chun. 2018. “How Logics Become Embedded: The Case of Quantification.” University of Iowa. Unpublished manuscript.

Sauder, Michael, Freda Lynn, and Joel M. Podolny. 2012. “Status: Insights from Organizational Sociology.” *Annual Review of Sociology* 38(1):267–83.

Schurman, Rachel. 2018. “Micro(Soft) Managing a ‘Green Revolution’ for Africa: The New Donor Culture and International Agricultural Development.” *World Development* 112:180–92.

Scott, James C. 1998. Seeing like a state: How certain schemes to improve the human condition have failed. New Haven, CT: Yale University Press.

Schia, Niels Nagelhus. 2018. “The Cyber Frontier and Digital Pitfalls in the Global South.” *Third World Quarterly* 39(5):821–37.

Sen, Gita and Avanti Mukherjee. 2014. “No Empowerment without Rights, No Rights without Politics: Gender-Equality, MDGs and the Post-2015 Development Agenda.” *Journal of Human Development and Capabilities* 15(2–3):188–202.

Seshia Galvin, Shaila. 2018. “The Farming of Trust.” *American Ethnologist* 45(4):495–507.

Shore, Cris and Susan Wright. 1997. “Policy: A New Field of Anthropology.” Pp. 3–39 in Anthropology of Policy: Critical perspectives on governance and power, edited by C. Shore and S. Wright. New York: Routledge.

Shore, Cris and Susan Wright. 2015. “Governing by Numbers: Audit Culture, Rankings and the New World Order.” *Social Anthropology* 23(1):22–28.

Shreves, Ric. 2017. “A Revolution in Trust: Distributed Ledger Technology in Relief & Development.” *Mercy Corps*. (https://www.mercycorps.org/research/revolution-trust-distributed-ledger-technology-relief-development).

Staudt, Kathleen. 1985. *Women, Foreign Assistance, and Advocacy Administration*. New York: Praeger Publishers.

Staudt, Kathleen. 1997. *Women, International Development, and Politics: The Bureaucratic Mire*. 2nd ed. Philadelphia: Temple University Press.

Strathern, Marilyn. 2000. *Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy*. New York: Routledge.

Swidler, Ann. 2009. “Dialectics of Patronage: Logics of Accountability at the African AIDS-NGO Interface.” Pp. 192–222 in *Globalization, Philanthropy, and Civil Society: Projecting Institutional Logics Abroad*, edited by D. Hammack and S. Heydemann. Indiana University Press.

Thomas, Vinod. 2010. “Evaluation Systems, Ethics, and Development Evaluation.” *American Journal of Evaluation* 31(4):540–48.

Thornton, Patricia H., William Ocasio, and Michael Lounsbury. 2012. *The Institutional Logics Perspective: A New Approach to Culture, Structure, and Process*. Oxford: University Press.

Timmermans, Stefan and Steven Epstein. 2010. A world of standards but not a standard world: Toward a sociology of standards and standardization. *Annual Review of Sociology* 36(1): 69–89.

Turco, Catherine. 2010. Cultural foundations of tokenism: Evidence from the leveraged buyout industry. *American Sociological Review* 75(6): 894–913.

UN General Assembly. 2015. Transforming our world: the 2030 Agenda for Sustainable Development, A/RES/70/1.

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

United Nations Operations. n.d. "UN Blockchain: Multi-UN Agency Platform." *UN Blockchain: Multi-UN Agency Platform*. Retrieved (https://un-blockchain.org/).

Urueña, René. 2012. "Internally Displaced Population in Colombia: A Case Study on the Domestic Aspects of Indicators as Technologies of Governance." Pp. 249–80 in *Governance by Indicators: Global Power through Quantification and Rankings*, *Law and Global Governance*, edited by K. E. Davis, A. Fisher, B. Kingsbury, and S. E. Merry. Oxford University Press.

USAID. 2011. "Evaluation Policy." (https://www.usaid.gov/evaluation/policy).

Vallejo, Bertha and Uta Wehn. 2016. "Capacity Development Evaluation: The Challenge of the Results Agenda and Measuring Return on Investment in the Global South." *World Development* 79:1–13.

van Eerdewijk, Anouka. 2014. The micropolitics of evaporation: Gender mainstreaming instruments in practice. *Journal of International Development* 26(3): 345–55.

Van Oort, Madison. 2018. The emotional labor of Surveillance: Digital control in fast fashion retail. *Critical Sociology* 1–13.

van Zyl, Winston Hyman, Frederik Claeyé, and Véronique Flambard. 2018. "Money, People or Mission? Accountability in Local and Non-Local NGOs." *Third World Quarterly* 1–25.

Venugopal, Rajesh. 2018. "Ineptitude, Ignorance, or Intent: The Social Construction of Failure in Development." *World Development* 106:238–47.

Viterna, Jocelyn and Cassandra Robertson. 2015. "New Directions for the Sociology of Development." *Annual Review of Sociology* 41:243–69.

Wallace, Tina, Lisa Bornstein, and Jennifer Chapman. 2006. The Aid Chain: Coercion and Commitment in Development NGOs. Rugby: ITDG Publishing.

Watkins, Susan Cotts, Ann Swidler, and Thomas Hannan. 2012. "Outsourcing Social Transformation: Development NGOs as Organizations." *Annual Review of Sociology* 38(1):285–315.

Weber, Max. 1978 [1922]. Economy and society: An outline of interpretive sociology. Berkeley: University of California Press.

WIRED. 2017. *Blockchain Expert Explains One Concept in 5 Levels of Difficulty*.

World Bank Independent Evaluation Group. 2013. "Evaluations | Independent Evaluation Group." (http://ieg.worldbankgroup.org/evaluations).

World Economic Forum. 2016. "What Is Blockchain?" *World Economic Forum*. Retrieved November 18, 2019 (https://www.weforum.org/videos/what-is-blockchain).

World Food Programme. 2017. “Blockchain Against Hunger: Harnessing Technology In Support Of Syrian Refugees.” (https://www.wfp.org/news/blockchain-against-hunger-harnessing-technology-support-syrian-refugees).

Yoder, Janice D. 1991. Rethinking tokenism: Looking beyond numbers. *Gender & Society* 5(2): 178–92.