ELSEVIER
www.elsevier.com/locate/worlddev

World Development Vol. 70, pp. 1–12, 2015

http://dx.doi.org/10.1016/j.worlddev.2014.12.019

"Show me the Numbers": Examining the Dynamics Between Evaluation and Government Performance in Developing Countries

LESLI HOEY*
University of Michigan, United States

Summary. — This paper examines the dynamics between monitoring and evaluation (M&E) and government performance in developing countries, where M&E systems are expanding rapidly. Findings in Bolivia suggest that approaches to M&E can lower staff morale, create burdensome paperwork, blind managers to operational problems and emerging innovations, and reinforce self-censorship, contributing to the very problem M&E is intended to solve. Crafted appropriately, M&E can instead become a tool to build practical judgment, increase staff motivation, and improve implementation incrementally. Ultimately, these findings contribute to efforts to design M&E that can support staff working under complex working conditions.

Key words — monitoring and evaluation, policy implementation, development planning, malnutrition, health policy, Bolivia

1. INTRODUCTION

This paper examines the dynamics between monitoring and evaluation (M&E) and government performance in developing countries, where M&E systems are expanding rapidly (Ernesto, Shand, Mackay, Rojas, & Saaverdra, 2006; EvalPartners, 2014). Rather than asking about the technical quality and rigor of M&E being used (Fukuda-Parr, Greenstein, & Stewart, 2013), my questions build on research about the conditions under which such systems improve or erode development policy implementation (Hood, 2012). Based on a case study of Bolivia's Zero Malnutrition (ZM) program, I suggest that mid-level managers may cling to collecting information about externally defined, quantitative indicators of staff performance as a *reaction* to complex social change processes—as coping mechanisms that give the allusion of controlling implementation. In these situations, evaluation can obscure operational issues, create burdensome paperwork, blind managers to emerging innovations, and reinforce self-censorship, contributing to the very problem M&E is intended to solve. On the other hand, where managers use M&E in ways that help build practical judgment about how to improve implementation, they create an environment conducive to learning—building motivation and trust—and engage more diverse actors and types of knowledge in analyzing problems and negotiating solutions. Ultimately, these findings contribute to literature aiming to reconsider how to design M&E to support staff working under complex conditions (Rogers & Fraser, 2014).

Efforts to institutionalize government-based monitoring and evaluation [1] (M&E) systems in developing countries have grown considerably over the past decade in response to the Millennium Development Goals (Savedoff, Levine, & Birdsall., 2006), Poverty Reduction Strategy Papers (Holvoet, Gildemyn, & Inberg, 2012), and the Paris and Accra Declarations (High Level Forum, 2008; OECD/DAC Organization for Economic Coordination and Development/Development Assistance Committee, 2005). Each of these initiatives calls for more "country-owned" development (Hyden, 2008) as well as monitoring and evaluation of donor investments and public policies (Thomas, 2010). In response, international institutions have launched numerous evaluation networks (IOEC, 2014) and initiatives to build development evaluation capacity (Mackay, 2006; Naidoo, 2013; Savedoff *et al.*, 2006). There is increasing evidence of "country-led—rather than donor-driven—efforts to institutionalize M&E" (May, Shand, Mackay, Rojas, & Saavedra, 2006, p. xi; Imas & Rist, 2009), and the demand for evaluators is growing; as of 2012, there were 138 national associations of professional evaluators representing 110 countries (EvalPartners, 2014), up from only five in 1990 when associations existed only in North America, Europe, and Australia (Donald, 2006).

At the heart of this exponential growth is a belief that M&E serves a variety of purposes: to hold actors accountable, identify policy options proven to work, and to improve the effectiveness of interventions during implementation (Hood, 2012; IOEC, 2014). While researchers—most notably those with the Abdul Latif Jameel Poverty Action Lab (J-PAL) (Kremer and Glennerster, 2012)—are showing that rigorous (e.g., randomized control trial) evaluations can help identify effective international development strategies, barriers still exist to support wider adoption of evidence-based policy during the planning process (Dhaliwal and Tulloch, 2011). Moreover, there is no standardized approach for building country-level capacity to mainstream M&E during the implementation phase (Goldberg and Bryant, 2012). One of the major debates is about whether certain evaluation models may contribute to the very problem they are intended to solve

* This study was supported primarily through an Inter-American Foundation Grassroots Development Fellowship as well as Field Research Grants awarded through Cornell University's Tinker Program in Latin American Studies, Einaudi Center for International Studies, and International Studies in Planning. The study's sponsors played no role in the design of the study, in the collection, analysis, and interpretation of the data, in the writing of the report, or in the decision to submit the paper for publication. I am grateful for the support, assistance, and participation of staff at national, departmental, and municipal levels of Bolivia's Ministry of Health and Zero Malnutrition Program and stakeholders from Bolivia's international nutrition policy community. I also thank John Forester, David Pelletier, and Mark Constas for their feedback on an early version of this report, as well as two anonymous referees for their valuable comments and suggestions. Final revision accepted: December 18, 2014.

(Hood, 2012; Rogers and Fraser, 2014; Westley, Zimmerman, & Patton, 2007). Non-governmental organization (NGO) staff and other development scholars have long argued that dominant forms of M&E systems focused on linear program designs, pre-determined, quantifiable indicators, and efficiency outcomes can discourage adaptation and innovation and encourage short-term and risk-averse projects (Chambers, 2010; Eyben, 2010; Natsios, 2010; Patton, 2010). Esser (2014) also raises concerns that the "country-ownership" and "aid harmonization" discourse supported by the Accra and Paris Declarations is moving the locus of accountability onto countries, so that they must take the responsibility for failures and successes, even as donors maintain their negotiating power over the types of interventions and M&E systems that countries must agree to in order to receive funding, what Esser calls a form of "expost-conditionality" (p. 51). The rapid rise in so-called "country-led" M&E systems, in this light, suggests that much of the trend is focused on pleasing donors (Eyben, 2010; Sjöstedt, 2013), rather than on improving government performance.

Regardless of the reasons, the growing evaluation agenda in developing countries suggests that a clearer understanding is needed about the ways in which M&E approaches sometimes complicate, rather than improve, complex interventions, even as we identify more appropriate M&E strategies for the types of problems development actors face. In what follows, I explain how different forms of M&E should be expected to affect policy implementation under different conditions, based on theories about complexity, bureaucracy, organizational change, and behavior economics. The remaining sections outline why Bolivia's ZM program offers a useful lens through which to explore these dynamics, the methods used in this study, and finally, how the findings offer lessons for re-designing development M&E systems.

2. THEORIES ABOUT M&E EFFECTS ON POLICY IMPLEMENTATION

At the heart of the many debates about the supposed benefits and drawbacks of evaluation are different understandings of the policy process that development actors face, with divergent implications for the form of M&E that should be appropriate. M&E proponents who believe that evaluation can inspire order and improve development results are often drawing on Max Weber's view of bureaucracy as a formal-rational system. Weber believed it was the normative appeal of rational–legal authority—the impersonal laws, procedures, and rules—that would compel employees to perform, because it offered a depoliticized, fair, stable, and predictable way of ordering society, rather than decision-making based on subjective beliefs, values, tradition, faith or "charismatic gifted persons" that was more common at the time (Gerth and Wright Mills, 1970, p. 199). Measurement systems play a key role in supporting this form of objective decision-making, allowing managers to track each worker to ensure results, as Weber wrote: "the performance of each individual worker is mathematically measured, each man becomes a little cog in the machine" (Weber in Mayer, 1956, p. 127) making "possible a particularly high degree of calculability of results for the heads of the organization" (Weber, 1978, p. 223).

The Weberian model persists, and the form of disciplining M&E that goes along with it, because it works in many situations (Stacey, 1996). This view of bureaucracy should produce expected results, complexity theorists argue, when public problems are "simple"—like baking a cake—when there are clear and agreed solutions to the problem, and the solution can be perfected through repetition and strict adherence to a recipe (Westley *et al.,* 2007). Complexity theorists and organizational change scholars also agree that a Weberian bureaucracy can function well when problems are "complicated"—like sending a rocket to the moon—where confounding factors can be reduced with enough information and coordination through centralized decision-making (top-down bureaucracies that rely on expertise) and technical rationality (based on planning, evaluation targets that allow for quickly identifying and "fixing" weaknesses in inputs, and sanctions and incentives to "command and control" staff into a coherent system) (Elmore, 1980, p. 605; Glouberman and Zimmerman, 2002; Mazmanian and Sabatier, 1983, p. 20; Westley *et al.,* 2007).

Scholars argue, however, that problems arise when Weberian bureaucracy—with rigid rules and M&E used to control behavior—is used in a situation where problems are "complex" (Guijt, 2007; Westley *et al.,* 2007) or "wicked" (Rittel and Webber, 1973). These include problems where causes are multidimensional and dynamic, and a multiplicity of stakeholders have conflicting perspectives about solutions. Hodson, Martin, Lopez, and Roscigno (2012) contend that this can turn institutions into "Kafkaesk" bureaucracies, where the norm is "divergent goals, unwritten rules, patrimonialism" and "chronic states of contradiction and confusion" (Hodson *et al.,* 2012, p. 265). Their analysis of 160 institutional ethnographies showed that rule breaking occurred routinely in 60% of organizations, while 86% showed widespread evidence of at least one of the "Kafkaesk" bureaucracy characteristic, leaving only 14% operating entirely through Weber's formal-rational model (p. 265). The authors conclude that "mock bureaucracies", full of "confusion, deceit, conflict and personal power" (p. 257), or situations where management and staff informally agree to break the rules, should be expected as the norm, rather than the exception to the Weberian rule (p. 260). This is akin to the type of rule-breaking behavior Lipsky (1980) found among street-level bureaucrats, or Friedmann (1993) arguments that implementation is inevitably a political act, displacing some existing practice, resource, time, staff, decision-making power, and more.

Behavior economics and organizational change research reinforce these arguments, showing how incentives and fines can cause staff to do the opposite of what the supervisor intended under certain conditions (Bowles & Polania-Reyes, 2012; Osterloh, Bruno, & Homberg, 2007). This may be particularly true among civil servants who have been found to "have a greater interest in altruistic activities and socially desirable outcomes" (Osterloh *et al.,* 2007, p. 11). Such intrinsic motivation has been shown to foster creativity, speed learning, improve conceptual understanding of the problem and solution, and encourage a more holistic approach (Hodson *et al.,* 2012; Osterloh *et al.,* 2007). However, introducing rewards or sanctions—including negative or positive feedback from M&E—can lead a person to lose their intrinsic motivation and interest in the immediate goal (e.g., to deliver a service) and shift their "locus of causality" externally to do the activity on the basis of the reward or punishment, so that "you get what you measure" (Bowles and Polania-Reyes, 2012; Hodson *et al.,* 2012; Osterloh *et al.,* 2007, p. 6). Especially when evaluations are used to critique and punish, this can lead staff to blame negative evaluations on outside factors or defend their actions rather than learn from them (Argyris and Schön, 1996; Frey, 2010, p. 17). The meaning an employee attributes to the incentive also matters, so that if they view it as a form of control or if they believe the supervisor distrusts them, they may intentionally perform worse to exert their sense of autonomy (Bowles and Polania-Reyes, 2012).

Incentives have been shown to be particularly problematic when tasks are complex or ambiguous, such as ensuring "good health", in large part because identifying clear objectives, and measuring every aspect of the task is difficult (Osterloh *et al.,* 2007, p. 9). The result is that people tend to focus on the easiest, quantifiable, and immediate goals—and ignore what is not rewarded or punished—producing "stereotyped repetition" of what already works and a more "superficial" approach to the task (Ordoñez, Schweitzer, Galinsky, & Bazerman, 2009; Osterloh *et al.,* 2007, p. 8). In part, this may be because people confronted by too much information will tend to bracket the complexity as best they can and "settle on plausibility" (Ordoñez *et al.,* 2009, p. 419), similar to Lindblom (1959) idea of incremental decision-making and Simon (1957) discussion of "satisficing" in situations of bounded rationality (p. 129). People under stress may actually process fewer cues and information in their environment, and revert to familiar practices (Staw, Sandelands, & Dutton, 1981), or even "inertia, protection of the status quo and sometimes even inaction—the deer in the headlights syndrome" (Ancona, 2011, p. 12). A similar problem arises when a complex task requires collaboration and the sharing of tacit knowledge about how to achieve an outcome. In these cases, "the spirit of cooperation tends to be overridden by the spirit of competition when individual rewards are considered" (Osterloh *et al.,* 2007, p. 10).

Alternatives to the "punishment centered" Weberian bureaucracy, scholars theorize, can help avert these unintended outcomes when working in complex situations. Research suggests that staff are more willing to adopt complex programs when they can differentiate the essential components from "peripheral" aspects that can be adjusted or ignored without compromising the intention of the program (Dearing, 2008). Supervisors can also try to ensure that staff accept the legitimacy of rules and regulations by increasing trust and applying incentives fairly (Hodson *et al.,* 2012; Osterloh *et al.,* 2007). Frey and Osterloh (2010) further suggest focusing evaluations less on the individual—which can undermine cooperation—and more on the organization overall, similar to arguments about focusing evaluations on "learning goals" rather than "performance goals" (Locke and Latham, 2006), and "outcome fidelity"—getting to the ultimate goal—more than "process fidelity", when protocols may not be appropriate in all contexts (Dearing, 2008, p. 106).

These suggestions also align with Hodson *et al.* (2012) proposal for avoiding the Kafkaesk bureaucracy: a "representative bureaucracy" based on flexibility and transparency, where supervisors and staff "would be involved in ongoing dialogue about the directions and goals of the enterprise" (p. 261). This is similar to Patton's (2010) "developmental evaluation" and Eyben's (2010) argument for a more "relationalist" approach to working through wicked problems, based on an iterative process, decentralized decision making, debate and disagreement, not just consensus, different paths to finding a solution, not just a unified approach, and "messy partnerships" (Guijt, 2008) that bring different actors together to discuss their partial understanding of a larger system. Finally, research shows that institutional reform requires changes to the broader structures that affect management and shape the organizational culture, including active efforts to encourage teamwork, reward innovation, increase staff exchanges, and support learning from failure (Khaleghian and Das Gupta, 2005; Potter and Brough, 2004; Tendler, 1997), and equally important, the removal of old routines that may conflict with new strategies to build a culture of learning (Mahler, 1997).

The Bolivia case examined here demonstrates how these M&E and implementation dynamics can play out in practice, producing both a Kafkaesk bureaucratic scenario in some situations, and elements of a "representative bureaucracy" in others.

3. CASE STUDY TOPIC AND SITE

Several reasons led me to select Bolivia's ZM program as a lens through which to study the dynamics between M&E and government performance. First, the very problem ZM was tackling—malnutrition—is a classically wicked problem: no single solution works consistently or universally, nor do multiple stakeholders agree what to do because the causes of malnutrition vary and regularly change across contexts (Bryce, Coitinho, Darnton-Hill, & Pinstrup-Andersen, 2008; Pelletier and Pelto, 2013). Second, multiple interventions have been shown to reduce undernutrition in controlled, efficacy trials (Bhutta *et al.,* 2008), but increasing the uptake of evidence-based policies on a larger scale is still a challenge (Leroy, Habicht, Pelto, & Bertozzi, 2007; Mills, 2012; Shekar 2008), as is understanding what comes after the plan—the capacities and M&E systems that can support effective implementation and adaptation of these proven actions (Pelletier and Pelto, 2013).

Third, I argue that ZM is a "prototypical" case study (Hague, Harrop, & Breslin, 1998; Rose, 1991)—a case that is not yet representative of developing country programs, but which is expected to become a model from which late adopters could learn. Bolivia was one of the first countries to launch a malnutrition program of ZM's size and scope, but many other countries, particularly in Latin America, have begun to follow suit, often calling on ZM planners to offer advice (AAHM Alliance Against Hunger and Malnutrition, 2014; FAO Food, 2008; IICA Inter-American Institute for Cooperation on Agriculture, 2009). ZM program designers also set out to make ZM evidence- and results-based; they selected interventions (see CONAN Consejo Nacional de Alimentación y Nutrición, 2006) recommended by the World Health Organization and other international nutrition experts (as recommended by Bhutta *et al.,* 2008), carried out an extensive baseline study, and set in motion the collection of data based on annual targets to track across four goals and 17 indicators (CONAN, 2008, pp. 21–23, 32–33).

Finally, several years of action research that I had conducted with the ZM program prior to this study allowed me to overcome one of the most difficult barriers to conducting in-depth policy analysis in developing countries (Walt, Shiffman, Schneider, Murray, Brugha, & Gilson, 2008): a deep understanding of the history and context of the ZM program as well as access to the nutrition policy community from national to local levels, across both international aid and government stakeholders.

4. METHODOLOGY

An embedded case study design (Yin, 2003) allowed me to draw comparisons about M&E use across eight diverse ZM implementation sites and between staff at national, regional, and local levels. Fieldwork took place over a period of 13 months, between September 2010 and December 2011 during ZM's fourth and fifth years of operation. To select the implementation sites, key informants were first asked to list rural or urban sites implementing all four ZM interventions

Box 1. ZM program interventions of focus in this study

- **Micronutrient initiatives**: Free Vitamin A, Zinc, and iron pills, a multi-nutrient packet called *Chispitas* known as Sprinkles in other countries, a fortified peanut-butter paste called *Atlu* known as Plumpy'nut in other countries, and a free complementary food mix of dried milk and micronutrients for children 6 months to 2 years called *Nutribebe*.
- **Integrated Management of Childhood Illnesses** (IMCI in English, AIEPI-Nut in Spanish): A prevention-oriented protocol used in clinics, promoted by the World Health Organization and UNICEF; Bolivia's version prioritizes nutrition.
- **Integrated Nutrition Units (UNIs):** Nutrition promotion and prevention centers intended to monitor and facilitate the integration of many ZM interventions into local health systems, including IMCI, micronutrients, nutrition promotion, and referrals of acute cases to larger hospitals.
- **Bono Juana Azurduy Program (Bono):** A conditional cash transfer program that encourages expecting and recent mothers to complete health check-ups. The program offers mothers a total of 1,820 Bolivianos ($US 260) over the course of 33 months during pregnancy and until the child is two years old.

Box 1.

(see Box 1) that had been established longest, were expected to be implemented nation-wide, and were considered to be the most important ZM interventions for reducing malnutrition in Bolivia.

I selected eight sites using maximum variation sampling (Stake, 1995) based on the variety of contextual variables key informants believed could affect implementation apart from the role of M&E. These sites included five rural and three urban municipalities located across five of Bolivia's nine departments (like U.S. states) and all three geographic regions (highland, valley, and lowlands), with population sizes that ranged from under 10,000 to over 1.6 million. Site characteristics also varied in terms of malnutrition rates (4–61%), poverty rates (19–99%), and political orientations, including sites in strong support of the Evo Morales administration in power at the time and sites strongly opposed. Finally, the lead organization also varied at each site; six of the sites were run entirely by public health staff, while NGOs were collaborating with public health centers to improve implementation at two of the sites.

To reduce the risk of biased conclusions, I triangulated methods and data sources—including participant observation (e.g., of results-based management trainings and strategic planning meetings), document review of ZM reports, secondary data analysis of internal ZM evaluations and health data, and interviews with diverse stakeholders at local, regional, and national levels of the program. In total, I completed 128 semi-structured interviews, involving nine key informants at the national level, 23 regional health managers, 17 local supervisors, 27 staff from 13 NGOs involved with ZM, and 52 health center staff. I also discussed my emerging findings through "member checks" (Stake, 1995) with key informants during and after data collection. Analysis was conducted using the constant comparative method (Glaser and Strauss, 1967), and "process-tracing" to develop grounded theory about the causal mechanisms that explained how M&E systems were affecting staff performance (George and Bennett, 2005).

5. FINDINGS

When the Morales administration entered office in 2006, decades after structural readjustment policies had weakened state functions (Dickovick and Eaton, 2013; Fauget, 2014), the state health system was "fragmented and incoherent", made up of public, NGO-led and private health care systems (Farthing and Kohl, 2014, p. 108). The public health sector reached less than half the population, characterized by "administrative incompetence, an obsolete and contradictory legal framework, and profound dependency on international donors" (Farthing and Kohl, 2014, p. 108); one in four children was malnourished nation-wide, with as many as two thirds of children stunted in the highland region (CONAN, 2006). ZM was part of a larger government plan to reduce inequalities and improve access to basic public services like health, housing, and education (Farthing and Kohl, 2014). It was also part of a broader attempt to reclaim state-level policymaking power to set standards, regulate the private sector, establish national initiatives, and ultimately, to re-establish the government's "sovereignty" from the international aid community (MPD, 2007; MOH, 2010, p. 113). In this context, ZM became one of the government's star programs (Morales, Pando, & Johannsen, 2010) when it was officially launched in 2007, as nutrition champions convinced nine ministries and the broader Bolivian aid community to agree to a common agenda and to finance it as well (Pelletier, Frongillo, Frongillo, Gervais, Menon, & Ngo, 2011).

As ZM was implemented, coordinators were involved in a variety of efforts to improve program reflection and learning at the national level. Yearly, they organized program review meetings, and in the third and fourth years of the program, they held a series of regional Results-Based Management Trainings and Logic Modeling Workshops to support a strategic planning process to develop the next five year plan; one ZM intervention was also the focus for a workshop that piloted a new approach to national-level M&E called the Program Assessment Guide (Pelletier, Corsi, Hoey, Faillance, & Houston, 2011). As a participant observer, these events appeared to offer national decision-makers invaluable insight into local-level logistical bottlenecks, other implementation challenges and emerging innovations, while offering lower-level ZM staff a clearer understanding of national administrator constraints and goals.

However useful these nationally-led workshops were as stand-alone activities, some key informants wondered if "the workshop culture" that was common through the Ministry of Health (MOH) was improving the day-to-day implementation realities of staff at the front-lines. While the workshops involved diverse stakeholders connected to the program—including international nutrition experts, mid-level ZM managers, and increasingly, local level ZM staff—most of the thousands of health center staff and local managers that were part of the existing health system could not feasibly participate in such events. Furthermore, such meetings tended to focus on ZM operations, rarely addressing M&E issues tied to the entire health sector's information management system that arose as the MOH implemented multiple, new programs that funneled down onto the same staff person at the front-line. Some key informants described this as ZM's Achilles heel; while coordinators could rely on ZM-specific staff to facilitate many actions, they still needed to work with existing, overburdened staff to institutionalize a preventive, nutrition-orientation throughout the health system, even as they juggled ZM responsibilities alongside other MOH initiatives, each of which often came with their own, separate M&E systems.

A focus at the front-line of ZM implementation—at the M&E experiences of newly hired ZM staff (e.g., UNI nutritionists) as well existing public health staff implementing the four interventions noted in Box 1—revealed two scenarios. In four of the implementation sites, managers were using M&E as a disciplining tool with negative repercussions for implementation, much like what is expected from Weberian forms of M&E implemented in a Kafkaesk institutional environment. At the other four sites, however, supervisors were using novel forms of evaluation that reconceived the types of actors, reflection processes, and knowledge useful for complex problem solving. Differences in geographic location, departmental affiliations, political orientation, or malnutrition and poverty rates did not explain this breakdown. Two of the four sites in the latter group had significant NGO support, suggesting that the resources NGOs had and the relative independence NGOs often have to try novel approaches likely played a role. However, in both cases, the NGOs were intentionally trying to build models that could be institutionalized in the public health sector, and were actively trying to work within the administrative, regulatory, staffing, and political constraints that health staff would normally face. Both also happened to be located in the two largest cities, but this did not seem to play a role as both NGOs indicated that they were implementing the same models in small towns and rural areas with similar effects.

(a) *Weberian evaluation*

What became quickly apparent in four of the study sites was that health managers appeared to rely on the collection of ZM indicators and data as a crutch, largely because they were unsure how else to manage ZM's ambitious change process and many other responsibilities. In these sites—three rural and one urban, all run by public health staff—M&E was weakening the ability of staff to implement ZM interventions, lowering staff morale, reducing the time staff had to deliver quality services, and reinforcing self-censorship among staff, so that many managers ultimately became blind to emerging innovations and heard little about how to improve implementation from those tasked with carrying out the program.

Front-line staff at these sites often discussed how their supervisors insisted "show me the numbers" in an effort to "control" them and "impose everything", leaving staff feeling like "we have to comply", "we must obey", and "we can't make adjustments because they're orders and we have to complete them". Behind this authoritarian veneer, however, supervisors were often struggling. In numerous conversations, managers confided that they were learning how to implement ZM as they went. One regional health administrator described, for instance, "Sometimes it seems the Ministry of Health does things just to do them, without thinking about how they will work…" In another regional health office, a Nutrition Coordinator admitted that she was losing a sense of the bigger picture because of the numerous elements she had to manage simultaneously, saying "We sometimes don't know what we're looking for because there are so many components…all with their own models and policies". Asked about one supervisor known for inciting fear, one ZM employee explained, "I guarantee you that [he] is more afraid of everyone else than they are of him. He's terrified that people will realize that he doesn't always know what he's doing…". This punishment-based approach to M&E that managers hid behind, however, had a profound effect on implementation.

First, many front-line staff explained that they felt defeated because of their supervisors' constant criticism, little interest in operational problems and failure to recognize what looked on paper like an insignificant victory. In one regional health office, administrators knew that they should "show staff that you know their reality" and that, "a firm hand isn't always best", but ultimately, they rationalized that they had to keep staff from "deceiving" supervisors and that staff needed to "learn how to obey" by using M&E to "see staff deficiencies, the bad, so that we can correct them". At sites like these, ZM managers distrusted staff, believing that M&E, especially quantitative indicators, were "lie-proof" disciplining tools that would scare staff into doing their job.

Staff who worked under supervisors like these expressed that they wanted to have the opportunity to explain in rich detail what their working conditions were like, and most of all, share the ideas they had about how to improve ZM interventions. One doctor felt supervisors' focus on numbers made managers forget "the simple things" that would improve staff morale:

> Supervisors think that because they experienced the same stress, we have to experience the same…I've never heard them say "congratulations"—the incentive could even be symbolic. But no, they just give us more work. It's stressful. The Head Doctor just comes to look at our numbers. Regional health supervisors have forgotten about simple things. They think more about their work. They view us as machines, not as people. We're also our own world. Every day, we fight (to create change), and our bosses don't know…

This mention of "machines" is similar to the way Weber talked of bureaucracies, where staff become cogs in a health sector machine that feels dehumanizing, to the point that supervisors do not see that "We're also our own world", full of stress and desires to be treated fairly and to be appreciated for their commitment to "fighting" for social change. In the absence of such recognition of "sacrifices", doctors like this interpreted supervisors' behavior as a ritualistic hazing—because I went through this, you do too—exerting their power through M&E and acting as if they have forgotten about the challenges of working in the field.

Second, supervisors' focus on "numbers" and the paperwork this required weakened performance by taking considerable time away from the very work M&E systems were intended to improve, leaving many health staff to conclude that supervisors were concerned with "quality paperwork" more than "quality health care". One Health Network administrator described how centers turned in 20 reports each month, in addition to 24 reports every 3 months. In rural areas, staff reported that paperwork was done by hand if health centers had no computer, taking as much as a week each month to complete—a fourth of staff time. In another rural center, one nurse worked each night until 10p.m. filling out forms, and sometimes until 1a.m. One NGO study, where a participant observer documented the daily routines of a rural health nurse for a month in 2006—even before the MOH added ZM and other programs to their workload—estimated that as much as 40% of the nurse's time was devoted to paperwork and other administrative tasks (Ramiro Llanque, personal communication, July 10, 2011). Some managers actually recognized that the paperwork was a burden on staff time, as one administrator explained: "Staff have to focus on filling out form after more forms…". But not knowing how else to "get results", she assumed this paperwork was their best option, asking, "But if we don't get staff to fill out these forms, how do we *control* them and see whether we're achieving our goals or not?"

The paradox was that this paperwork and combination of mounting programs created a situation where staff admitted making mistakes, did not carry out many of the steps or chose

not to implement particular programs, or simply decided not to fill out the required paperwork. In one city, a doctor described how, "The directors of each program require more, more, more—we're saturated...It's so much paperwork, it limits our ability to do some things correctly...We do what we can, but we can't do everything that we're required to do...The paperwork doesn't allow us make progress!" In thinking about how such paperwork was used, doctors often asked, "So these reports—for what? They make us work, fill out forms, for nothing!", demonstrating how staff began to see M&E systems as something that gets in the way of their job, serving only as a reporting tool to ensure compliance, because they never saw the information used.

Finally, one of the most pernicious effects of the management-by-numbers approach was that it often overlooked—and contributed to—the way staff self-censored their novel ideas and feedback. Rather than encourage learning, innovations, and problem solving, managers were often unknowingly fostering an atmosphere of secrecy and competition, reducing the chances of ideas being shared in much the same way behavior economics and organizational change literatures predict that M&E systems can backfire. Several staff wanted to suggest ways to improve operational problems, but worried that they would be punished for doing so. One regional health manager shared the same fear as lower level staff, noting, "I didn't want to tell [supervisor] my ideas or observations because I didn't want to lose my job!" Another nutritionist explained that some managers had earned reputations for punishing staff who questioned particular interventions, like the UNIs, "I want to tell [supervisor] about the problem we have with the UNIs, but ...the moment you start talking to him about the problems, he'll stop helping you, they've warned me." To her, this sent a message that "I think [supervisor] doesn't want to look at operational issues." These staff, like others, were describing how M&E was too often one-sided, requiring them to report on their activities, without a chance to offer their own feedback.

This focus on quantitative indicators also prevented supervisors from hearing about innovative ideas or strategies staff were already trying. This occurred because the "official" space to collectively analyze health indicators and to regularly adjust implementation plans—referred to as Information Analysis Committees (CAIs)—often did not encourage staff to discuss their work outside of a set of prescribed, quantitative indicators. CAIs meetings were held with all health staff and open to community members as frequently as every month, at departmental health office to village levels of the health system. They had been part of the public health system operations for years prior to ZM's launch, but many staff explained that CAIs were not geared to generating strategies for resolving weaknesses in health program operations. They described CAIs as "informative" rather than strategic planning sessions, noting how "Each center or program presents for just a few minutes. There's no time to talk, analyze, think about ideas". A national MOH administrator admitted that part of the problem may have been the way national-level program designers were introducing new data to local health systems:

> CAIs have become very institutionalized but in an unhelpful way. It tends to be very technical, involving only health staff, and only looking at coverage rates. They might say, "huh, our coverage is low there or there, we should do something", but no real discussion about why that is, or how it happens, and almost never any follow-through—so it's not useful...We want this data to be the basis for decision-making and priority setting, but we haven't gotten very far on coming up with concrete ideas of how to do that.

This national administrator recognized, then, that simply throwing more data at staff was not helpful and that they needed to learn what to do with this information. But, she also admitted that the entrenched culture of how CAIs functioned might make this retooling quite difficult.

In the most egregious example of how such spaces excluded discussions beyond the official indicators, a nurse in a rural health center described a long list of unique strategies she had been implementing, some for years, including participatory workshops, trainings, and surveys she had developed with mothers and community health promoters. She explained, however, "A lot of these [activities] I don't report to my supervisors—there's nowhere to report them". Not being asked, and not seeing a space within CAIs to talk about innovative practices that did not fit the "official" indicators meant she carried out her ideas in virtual obscurity. And when she did speak up once in a CAI, supervisors and colleagues thought she was lying when she explained that she had organized women's clubs years before the MOH suggested them.

Similarly, supervisors were often unaware of—or chose not to mediate—situations where health staff self-censored or suppressed each other's ideas out of "jealousy and egoism", what one regional UNI coordinator described as a "mafia against sharing experiences". Several staff told stories of other colleagues stealing their ideas or taking an innovative suggestion as "an affront, and as suggesting that they aren't doing something *right* to begin with". This behavior among colleagues led one health center doctor to conclude: "We're destructive, not constructive. We should be working as a family in the municipality, to generate ideas, like a team. ...we all have the same objectives. So that we don't reinvent the wheel, we've got to share experiences. ...The theory is something, but practice another—it's so important to learn from practice...". His reflections showed how the "mafia against sharing experiences" prevented staff from learning from each other's tacit knowledge and to work collectively toward the same objective—to improve public health.

One M&E tool ZM administrators introduced several years after ZM was launched—the SVIN-C (Sistema de Vigilancia Nutricional Comunitario)—was intended to improve local-level data-based decision-making and increase ZM support and involvement of other health staff and community-level leaders through a participatory survey and analysis of mothers' changing nutrition knowledge, household practices, and use of micronutrient products. While several staff interviewed explained that the SVIN-C had done just as national coordinators had hoped—assisting them in prioritizing their work and building a shared sense of purpose in the municipality—the experience of one UNI nutritionist suggested that the survey might tell staff *where* in the municipality to increase micronutrient distribution or even *which* nutrition messages to prioritize, but gave no new answers about *how* to convince mothers to use the micronutrients or change their behavior. Particularly because ZM coordinators expected staff to implement SVIN-C two to four times a year, this nurse described how she was concerned that "It takes time. With just one person in the UNI it's as if all we're doing is dedicating ourselves to the SVIN-C without any new ideas of what to do differently". She also faced a major barrier not apparent in places where the SVIN-C worked—supervisors who were unsupportive, if not hostile, to the nutrition work she was trying to lead. The mixed experience with SVIN-C suggests that simply putting new M&E tools in the hands of staff was not working in all cases to resolve the many issues just described, particularly when staff felt inadequately trained to translate the information into more innovative actions and did not already have

the political support of mid-level management in the broader health system.

(b) *Indigenous evaluation*

In the other four sites—two rural and two urban sites—supervisors were not simply introducing new M&E tools, but creating a culture of learning and mutual trust to encourage staff to learn from each others' tacit knowledge and to problem solve together. Indigenous evaluation strategies they were using—Mobile Brigades, Forum for Debate, Case Study Reviews, and Quality Assurance Management—suggested options for preventing many of the pathologies that were emerging where ZM actors were using disciplining forms of M&E.

(i) *Mobile Brigades*

In the two rural examples, staff launched "Mobile Brigades", which involved municipal administrators and heads of each of the key health programs traveling together to a different village on a monthly or bi-monthly basis to conduct "multiprogrammatic" visits (e.g., to distribute micronutrients, vaccinate, follow up with tuberculosis patients and more). Staff described how the Brigades helped build allies for the ZM program and reach more distant communities with comprehensive health services. From an M&E perspective, these Brigade trips created an intimate and intensive space for staff to learn about each other's programs and think through challenges they collectively faced to reduce malnutrition. Where before, ZM and other health staff had been working in isolation of each other, after initiating the Brigades, ZM staff began to develop joint schedules, joint budgets, and innovative food security initiatives with other health staff and municipal employees. They described about how they "spoke the same language" about their collective goal to reduce malnutrition and how the Brigades motivated them—and gave them the knowledge—to step in to support other ZM programs if one intervention began to falter.

(ii) *Forum for Debate*

An additional evaluation strategy used in one of these rural municipalities, which led to the idea of a Mobile Brigade, was something staff referred to as a "Forum for Debate". After a drop in enrollment rates in one ZM initiative, health administrators decided to invite all health staff and village authorities in the municipality to collectively analyze the situation during a day-long event. After reviewing health data and reports from health staff about the challenges they were facing, an "open debate" allowed participants to "talk about everything". The Mobile Brigade was suggested as a way to address the enrollment problem but also other weaknesses identified. During this dialog and regular staff meetings, one nutritionist noted that a key reason they were able to work out the challenges they faced collectively was because, "Everyone gives their opinion... Criticism is welcome. We see that it all helps improve health. We share a trust among us".

(iii) *Case Study Reviews*

In a third example, an NGO in the city of El Alto started by intentionally building staff dedication and teamwork, before implementing innovative M&E approaches. One manager described how, "I think working with each individual has been fundamental...listening to their expressions, ideas, emotions and feelings". She actively talked about building staff discretion as well, by "respecting the manner in which each person works", allowing each person to develop their own approach to achieve results, only suggesting *necessary* corrections along the way, noting how:

> When someone says to you, "I have confidence in you", it means that that person knows it's not necessary to control your work schedule nor stand behind you to know that you're doing your work, because that person assumes you are a responsible person, a professional. We work based on results. So if you're so good that you meet your results in half the time that it normally requires, that's fine...It's a way of respecting the manner in which each person works...obviously correcting what isn't done well, but it's about communicating a level of confidence and respect towards their work that I believe is fundamental...

Rather than tracking the minutiae of staff work, this manager built confidence by trusting that different approaches could get the same results. Field staff reiterated the respect they felt, noting how the NGO "has listened" as they described how: "We're always invited to meetings. Especially in the beginning, we talked about the values of (the NGO) and our own vision. This helps us become committed because we feel like part of the institution, as if we also are helping to construct it." Building confidence and commitment, however, did not rest entirely on pep talks or inclusion in decision-making, but was tightly connected to data-based decisions and reflection. Staff described how the variety of M&E activities—including a baseline study, impromptu meetings, trimester and yearly program reviews, field supervisions and community-based updates—"helps us reflect on what's missing, what we haven't completed" and "helps to adjust goals".

The key way staff engaged data was through Case Study Reviews. During two sessions I observed, staff displayed data about a family they found particularly interesting or challenging on a poster and discussed how the family got to El Alto, who lived in the home, their access to water and sanitation, positive practices that could be reinforced, the greatest risks for children becoming or staying malnourished. They also explained how they had attempted to establish a good first impression, how they reacted to unexpected events, and other strategies they had implemented so far. They then opened the floor to a group discussion about what colleagues might have done differently based on their experience, questions to ask the family next time, family members they might engage more fully, and issues they should prioritize as they continue working with the family. By making explicit how staff thought through a problem and strategies staff were using, the learning space improved staff judgment to react strategically when they were with families and to prioritize the array of factors to which they could pay attention. As one person described, "sometimes I think a certain issue is urgent, but then other staff help me see that I should focus on something else first". Staff also learned as much about each other as they did about the case in these Reviews, important for team building and developing a shared understanding about the task at hand. One staffer described her realization coming out of these sessions that, "Sometimes we have different perceptions within the team, but those with more experience, 7, 24 years, help us understand. We all have distinct experiences." In the end, unlike CAIs that focused on aggregate numbers, this space allowed staff to debate, challenge, and reinforce actions, so that the next time, a staffer had a more concrete sense of what to do to "improve the numbers".

(iv) *The Quality Assurance Model*

The final example comes from another NGO working in the city of Santa Cruz, funded by USAID to increase implementation of the malnutrition prevention methodology AIEPI-Nut (Box 1) using USAID's globally-implemented, but locally adapted Quality Assurance Model (QAM). Staff involved

believed QAM had not resolved all their implementation issues, but considering that no one was implementing AIEPI-Nut in the beginning of the project—despite the fact that the MOH had launched it and initiated trainings nationally at least two years prior to the start of the USAID project—it was notable that managers and staff alike cited that up to 70% of health staff in the health network were implementing the protocols by the end of the project. This was especially impressive in a city where many health staff were energized by anti-government sentiments and saw new initiatives coming out of the Morales administration as "of the MAS", Morales' *Movimiento Al Socialism* political party. The NGO accomplished these changes by: (a) using intentional motivational strategies, (b) reinforcing skills and knowledge, (c) focusing on simple, doable and immediate changes, and (d) encouraging more diverse stakeholders to participate in "Learning Sessions".

First, one supervisor described that the main reason medical staff eventually adopted AIEPI-Nut was because they became motivated, "If you tell someone to do something, they won't do it. Right? So we worked with motivational aspects a lot. We talked about what their responsibilities were as professionals, to society and legally... That's where we define with them 'what's quality health care'. We defined it from below, not a definition dictated to them". This idea, that "if you tell someone to do something, they won't do it" was fundamental, ensuring that staff came to a decision that this new practice was aligned with their interests, spoke to their responsibilities "as professionals, to society and legally", and that requirements dictated by the MOH for improved "quality care" could also reflect their own values. Part of the motivational strategy also involved offering symbolic "prizes", like putting staff photos on a project poster or offering an electric water boiler for centers implementing AIEPI-Nut most effectively.

Second, the program actively reinforced staff skills and motivation throughout the project, based on the idea that "one training isn't enough". This involved supplying staff with a baseline of their skill level and monthly tests to monitor their own progress, meetings each month, ongoing peer supervisions, and visit exchanges to other departments to do "crossed-monitoring". The third element—and the heart of QAM—links data to changes that are doable, small, and immediate, actions that would not overwhelm already overextended staff, using already available resources, knowledge, and time. This meant thinking of specific and concrete solutions to problems that could make work easier—rather than abstract notions of improving "coordination" or "communication"—and also using one simple indicator to monitor whether such changes were making any difference (i.e., whether a child was still malnourished). If the action had made little difference after 30 days, health center teams proposed a different strategy the next time, and if the first effort had led to improvements, they moved onto the next, doable problem they could tackle. Some of the small changes that emerged out of this process included making children's health cards larger so that mothers could see them better and using colored tabs on clinic histories to identify malnourished cases rapidly. Ultimately, doing something fast was also part of the project's aim to motivate staff—assuming that breaking a complex program like AIEPI-Nut into smaller components, and accomplishing even small actions quickly, sets the pace to gradually make enough changes to fully adopt the program.

One crucial element of this last step also involved thinking differently about the types of actors needed in the room to improve problem solving. Importantly, the program coordinator called the analysis meetings "Learning Sessions" to differentiate them from CAIs, symbolizing that the space was intended to bring many actors together to analyze quantitative indicators but also to test, debate, suggest, and share ideas and experiences. The NGO coordinator noted that part of the importance of involving everyone connected to the health center—including doctors, nurses, administrators, and even the doormen, receptionists and local mothers—was to "deliver the same message", recognizing that, "If they weren't involved in these processes, then at some point, they could become an obstacle."

6. DISCUSSION

When there is an overload of information and multiple, even conflicting priorities, it only seems rational that many ZM supervisors focused on meeting short-term tasks and documenting simple performance indicators, bracketing out other, tangible cues that suggested program implementation was faltering precisely *because* they were managing an ambitious program with multiple components alongside other MOH responsibilities. But the lowered morale among staff, self-censorship, and rule-breaking that occurred under these supervisors should have also been predictable (Bowles and Polania-Reyes, 2012; Osterloh *et al.*, 2007). Knowing that mangers may have relied on simplified M&E indicators in response to an ambiguous work environment, however, offers some hope that many would have been open to more guidance about how to re-think their M&E approach. Some of the more frustrated managers asked for ways to resolve the M&E challenges they faced, noting how "we have to learn a culture of learning". The indigenous evaluation approaches that emerged in the other sites offered approaches that could have averted the pathologies of implementing Weberian forms of punishment M&E in more Kafkaesk circumstances (Eyben, 2010; Frey and Osterloh, 2010; Hodson *et al.*, 2012; Patton, 2010).

First, supervisors in the initial four sites overlooked questions about what was being done, and how to fix implementation on a day-to-day basis, in part, because they were looking at multiple, abstract indicators of implementation progress while rarely involving staff in productive analysis or follow-up. Alternatively, by focusing on simple, doable, and immediate actions, the last four sites did what Dearing (2008) suggested, differentiating "essential" from "peripheral" program components. This not only made a complicated task—to reduce and prevent malnutrition—more feasible, but also motivated staff to keep improving. The immediacy of seeing tangible changes in their work environment—and often, changes in children's health—sparked more ideas about other strategies they could try or routines they could change. In the Forum for Debate, the start of the conversation was enrollment rates, but this led to a strategy that could address multiple public health interventions. Case Study Reviews helped staff focus on the few key issues that might make the greatest difference in preventing malnutrition for a specific family—not on every determinant for which they had information. In Santa Cruz too, coordinators boiled down the entire purpose of AIEPI-Nut to one key indicator—whether each child's nutrition was improving each month. In each of these cases, focusing on essential indicators also gave staff the license to rethink protocols, even simple protocols like the nutrition card mothers read, or questions to ask a family during the next visit, reflecting Dearing's (2008) idea of emphasizing "outcome fidelity" over "process fidelity".

Second, supervisors in the first four sites spoke of needing to "teach them to obey" or "control" staff, relying on indicators

and measurement while ignoring individual emotions, struggles, and concerns. This epitomizes how Weber thought bureaucracies should function "the more perfectly" they are developed, noting how "the more it is 'dehumanized', the more completely it succeeds in eliminating from official business love, hatred and all purely personal, irrational and emotional elements which escape calculation" (Weber, 1946, p. 216). Yet, ZM staff interpreted such dehumanizing actions as caring more about "quality paperwork" than "quality healthcare", leading some staff to outright defy orders. Alternatively, supervisors in the last four sites appeared to build on suggestions to improve trust (Osterloh *et al.,* 2007) and a more "relational" managerial approach as Eyben (2010) suggests. These supervisors used terms like "listen", "friend", and "connection" to describe their interactions with staff. In using the term "friend", supervisors were not implying that they were letting staff do whatever they liked. Rather—they were speaking of changing the hierarchical relationship of a supervisor holding staff to account or "monitoring" based on indicators, to supervision based on being attentive to the challenges staff faced. They were implying that the manager could listen to "emotions and feelings", as a way to show respect, but also as a way to understand what was motivating or impairing staff work. They asked staff about their experiences and opinions to establish trust, offered low-cost incentives, but more importantly, symbolic recognition, and got out from behind their desks to show up personally in the clinics and villages where staff worked tirelessly. Forester (1999) similarly described urban planners who were adept at facilitating change processes in the face of contentious relationships, calling them "critical friends", or planners "who care enough to listen for more than what has been said, who care enough to wonder about what has been missed, who are engaged and collaborative enough to help, yet detached and independent enough to carry forward their own projects" (p. 196).

Finally, the problem solving approaches used in the last four examples also revealed a different way of understanding the type of knowledge that can be useful for evaluating a situation or coming up with a solution. A focus only on quantitative indicators as occurs in many CAIs trusts the systems of accountability, the data collectors, the standards set by program designers and the experts, more than the individuals involved in implementation (Freeman, 2002). Such faith in indicators reduces the need to involve diverse actors in the analysis, like the receptionist in Santa Cruz. Indicators also encourage a focus on accountability—to answer questions about who got their numbers this month—rather than learning. The experiences of the last four sites, on the other hand, suggest that learning how to improve a complex implementation process requires an interrogation of the indicators with multiple stakeholders—to not simply look at them for their own sake, but to set them alongside experiential and tacit knowledge, values, priorities, and operational constraints, and only then deciding what to do—incorporating the "messy partnerships", engaging disagreement, and other aspects of a "representative bureaucracy" Eyben (2010) and Hodson *et al.* (2012) suggested. Ultimately, the Case Study Reviews, Forum for Debate and Learning Sessions, showed how much more goes on in "implementation" than abstract argumentation through data analysis could ever capture, requiring as well (a) a dialog to understand the problems and obstacles staff were facing, (b) debates about better and worse solutions, and (c) negotiations to agree on a course of action, all steps that Forester (2009) and Weick and Sutcliffe (2005) described as useful for collectively understanding what to do next in a complex situation.

7. CONCLUSION

Policy actors likely persevere with a Weberian bureaucratic model because of the presumed *possibility* that a public problem will be resolved if only a more stable, highly efficient, coordinated system can be developed, with a disciplinary form of M&E as its basis. The first four implementation sites in this study, however, were emblematic of Hodson *et al.* (2012) prediction, that: "A failure to fully recognize the darker side of bureaucracy allows problems to go unaddressed or to be interpreted as resulting from individual malfeasance rather than being a predictable part of organizational functioning" (Hodson *et al.,* 2012, p. 274). Where evaluation advocates expect M&E systems to improve the efficiency of government services (Haily, 2000; Lindenberg and Bryant, 2001), the ZM case showed that reporting requirements can instead overwhelm staff time, distracting them from the real work they are intended to do. And where M&E is expected to improve government effectiveness (OED Operations Evaluation Department, 2004; Savedoff *et al.,* 2006), the disciplining form of evaluation systems used in ZM obscured critical operational issues, blinded managers to emerging innovations, and reinforced self-censorship, ultimately weakening, not strengthening implementation.

The indigenous evaluation strategies that emerged in this study suggest the opposite—that crafted appropriately, M&E can instead become a tool to build practical judgment about what to do in a complex task situation, avoiding the pathologies of "punishment centered" Weberian forms of M&E (Dearing, 2008; Frey and Osterloh, 2010; Hodson *et al.,* 2012; Osterloh *et al.,* 2007; Weick and Sutcliffe, 2005). Even if certain approaches that emerged may not be cost-effective to maintain after external NGO funding and facilitation ends (Franco and Marquez, 2010), key concepts shared across the QAM, the Forum for Debate, Mobile Brigades and Case Study Reviews could still form the basis of rethinking M&E systems in complex problem solving situations, including: (a) focusing monitoring activities on identifying and testing simple, doable, and immediate actions to improve implementation incrementally, (b) supervising staff performance more as compassionate "critical friends" by not just looking at staff numbers but also listening and recognizing staff work, even symbolically, and (c) involving more diverse actors in scrutinizing quantitative indicators alongside experience and tacit knowledge to enhance learning and strengthen staff members' practical judgment.

The experience at these sites also suggests that participatory approaches can add additional insights into how to improve implementation, though stakeholders other than staff were only nominally included in the Forum for Debate and Learning Sessions. In other studies, more formal community-based accountability systems have been successful in improving the provision of infrastructure services in Bolivia (Yanez-Pagans and Michacado-Salas, 2014). Bolivia in particular, with its long history of participatory planning and decentralized government (Fauget, 2014) is a place where such approaches would be expected to work in a variety of sectors. But studies also show that community-level accountability systems work best when the community itself values the public service, public officials have an incentive to engage in monitoring, and there is a formal process for the public to voice complaints and hold providers accountable (Yanez-Pagans and Michacado-Salas, 2014). None of these conditions were true in this case; ZM focused on a public service that the public, local officials, and even the medical community often failed to recognize was needed because the issue itself—malnutrition—is so "hidden".

In such cases, or until the public is mobilized sufficiently to care about malnutrition, improving the types of staff-based M&E studied here are all the more important.

Despite the many challenges ZM implementers faced, initial evaluations suggest that the number of stunted children under two that attend health establishments fell during the time of the ZM program from 18.5% to 13.5% during 2008–11 (Laforce and Silva, 2013, p. 39). I argue, nonetheless, that ZM could have achieved even greater impact, faster, if administrators had reflected further on their assumptions about the role they expected local-level M&E to play in implementation, asked about the impact these reporting practices were actually having on staff performance, and taken the time to learn about the more innovative forms of indigenous evaluation that were emerging. While this was happening on a limited scale—as with the SVIN-C surveys—one consultant who visited over fifty of ZM's priority municipalities believed that some of the indigenous M&E practices he too observed were rare and "spontaneous...not *created* by the national level". The yearly ZM program reviews, PAG, Logic Modeling Workshops, and Results-Based Management Workshops demonstrated that ZM coordinators were regularly thinking about how they might improve M&E throughout the ZM program, but the findings of this study suggest that there was a disconnect between these one-off events and the daily reality ZM actors faced on the ground.

There was also some indication that Bolivia's National Health Information System (SNIS) authorities were beginning to reconsider the national health data system that had grown unchecked since it was launched in 1990, discussing a plan for creating a "unified" information system. The SNIS plan, however, still said little about the frequency with which staff would be expected to collect data, who really needed the data, if at all, and most importantly, how to facilitate effective use of the data (SNIS Sistema Nacional de Informacion en Salud., 2010). As this study showed, the problem was not simply about duplicate data sources; the major issue was that staff embroiled in these systems saw little personal use for such data, even as it was used as a disciplining tool against them and took up inordinate time. If anything, this case study suggests that the process of changing the broader institutional cultures and removal of old routines that can get in the way of creating a learning culture (Khaleghian and Das Gupta 2005; Mahler 1997; Potter and Brough, 2004) is a slow process. Realistically, institutionalizing and sustaining such change through a top-down approach may never work, but at a minimum, higher level administrators would have to support the shift toward "learning goals" and "outcome fidelity" to encourage and support cross-exchanges of the sorts of place-based indigenous approaches that were emerging here.

Ultimately, the ZM experience suggests that if the aid community is indeed committed to supporting more effective development outcomes, as well as country-owned and country-led efforts, it cannot simply encourage, demand, or offer more training in M&E, but must also think more critically about the *form* of M&E they advocate. Certainly, there is still a need for rigorous impact evaluations of development interventions and rational approaches to M&E in situations where problems are simple and complicated, but the norm will likely need to move toward supporting more adaptive and relational M&E approaches during the implementation of interventions in the complex change situations that are more common. The ZM case also suggests the need for system-wide meta-evaluation—an assessment that considers the impact and interactive effects of multiple, overlapping evaluation systems that cross government sectors, sector-based programs, and simultaneous donor and NGO measurements systems. Without considering how existing and new M&E systems ultimately affect day-to-day work on the ground, lofty plans to build more transparent, accountable, efficient, and effective results-based management may in fact do the opposite, as the ZM case suggests.

NOTES

1. The Development Assistance Committee (DAC) Working Party on Aid Evaluation (2002) defines evaluation, which incorporates ongoing "monitoring" of implementation, as "The systematic and objective assessment of an on-going or completed project, program or policy, its design, implementation and results...to determine the relevance and fulfillment of objectives, development efficiency, effectiveness, impact and sustainability (and to enable)...the incorporation of lessons learned into the decision-making process" (21-22).

REFERENCES

AAHM (Alliance Against Hunger and Malnutrition), (2014). Retrieved February 14, 2014: <http://www.theaahm.org/>.

Ancona, D. (2011). Sensemaking: Framing and acting in the unknown. In S. Spook, N. Nohria, & R. Khurana (Eds.), *The handbook for teaching leadership: Knowing, doing and being* (pp. 1–18). Thousand Oaks, CA: SAGE Publications.

Argyris, C., & Schön, D. (1996). *Organizational learning II: Theory, method, and practice*. Reading, MA: Addison-Wesley.

Bhutta, Z. A., Ahmed, T., Black, R. E., Cousens, S., Dewey, K., Giugliani, E., *et al.* (2008). What works interventions for maternal and child undernutrition and survival. *Lancet, 371*, 417–440.

Bowles, S., & Polania-Reyes, S. (2012). Economic incentives and social preferences: Substitutes or complements?. *Journal of Economic Literature, 50*(2), 368–425.

Bryce, J., Coitinho, D., Darnton-Hill, I., Pelletier, D., & Pinstrup-Andersen, P. (2008). Maternal and child undernutrition 4 material and child undernutrition: Effective action at the national level. *The Lancet, 371*(9611), 510–526.

Chambers, R. (2010). *Paradigms, poverty and adaptive pluralism. IDS working paper 344*. Brighton, UK: Institute of Development Studies.

CONAN (Consejo Nacional de Alimentación y Nutrición), (2006). *Desnutrición Cero al 2010: Compromiso multisectorial*. La Paz, Bolivia: Author.

CONAN, (2008). *Plan Estrategico 2007-2011 del Programa Sectorial de Desnutricion Cero*. La Paz, Bolivia: Author.

Dearing, J. (2008). Evolution of diffusion and dissemination theory. *Journal of Public Health Management Practice, 14*(2), 99–108.

Development Assistance Committee (DAC) Working Party on Aid Evaluation (2002). *Glossary of key terms in evaluation and results-based management*. Paris: OECD.

Dhaliwal, I., & Tulloch, C. (2011). From research to policy. Using evidence to inform development policy. In R. Puttck (Ed.), *Using*

evidence to improve social policy and practice (pp. 92–134). London: NESTA.
Dickovick, J., & Eaton, K. (2013). Latin America's resurgent centre: National government strategies after decentralization. *The Journal of Development Studies, 49*(11), 1453–1466.
Donald, C., (2006). Consequences of contemporary evaluation practice on the future of the profession: Are we heading toward disaster or paradise? Paper presented at the American Evaluation Association conference, Portland, Oregon, November 1–4, 2006.
Ernesto, M. Shand, D. Mackay, K. Rojas, F., & Saaverdra, J., (2006). Towards the Institutionalization of Monitoring and Evaluation Systems in Latin America and the Caribbean: Proceedings of a World Bank/Inter- American Development Bank Conference pp. 29–35. Washington, DC: World Bank.
Esser, D. (2014). Global public health: An international journal for research, policy and practice. *Global Public Health, 9*(1–2), 43–56.
EvalPartners, (2014). International mapping of evaluation associations launched. Retrieved August 29, 2014: <http://www.mymande.org/evalpartners/international-mapping-of-evaluation>.
Eyben, R. (2010). Hiding relations: The irony of 'effective aid'. *European Journal of Development Research, 22*, 382–397.
FAO (Food and Agriculture Organization), (2008). Panorama del Hambre en el America Latina y el Caribe. Rome: Author.
Farthing, L., & Kohl, B. (2014). *Evo's Bolivia, Continuity and change*. Austin, TX: Univ. of Texas Press.
Fauget, J.-P. (2014). Can subnational autonomy strengthen democracy in Bolivia?. *Publius: The Journal of Federalism, 44*(1), 51–81.
Forester, J. (1999). *The deliberative practitioner*. Cambridge, MA: MIT Press.
Forester, J. (2009). *Dealing with differences*. New York: Oxford University Press.
Franco, L., & Marquez, L. (2010). *The learning system to support health care improvement*. Washington, DC: USAID.
Freeman, T. (2002). Using performance indicators to improve health care quality in the public sector: A review of the literature. *Health Services Management Research, 15*, 126–137.
Friedmann, J. (1993). Toward a non-euclidian mode of planning. *JAPA, 59*, 482–485.
Frey & Osterloh (2010). Evaluations: Hidden costs, questionable benefits, and superior alternatives. In T. Jansen, G. van den Brink, & J. Kole (Eds.). *Professional pride. A powerful force* (pp. 175–196). Boom, Amsterdam: Eleven International Publishing.
Fukuda-Parr, S., Greenstein, J., & Stewart, D. (2013). How should MDG success and failure be judged: Faster progress or achieving the targets?. *World Development, 41*, 19–30.
George, A., & Bennett, A. (2005). *Case studies and theory development in the social studies*. Cambridge: MIT Press.
Gerth, H. H., & Wright Mills, C. (Eds.) (1970). *From max weber: Essays in sociology*. London: Routledge and Kegan Pau.
Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New York: Aldine de Gruyter.
Glouberman, S. & Zimmerman, B., (2002). Complicated and complex systems: What would successful reform of Medicare look like? Discussion paper no. 8. Toronto: Commission on the Future of Health Care in Canada.
Goldberg, J., & Bryant, M. (2012). Country ownership and capacity building: The next buzzwords in health systems strengthening or a truly new approach to development?. *BMC Public Health, 12*, 531–540.
Guijt, I. (2007). *Assessing and learning for social change: A discussion paper*. Brighton, UK: Institute of Development Studies.
Guijt, I. (2008). *Seeking Surprise. Rethinking Monitoring for Collective Learning in Rural Resource Management*. PhD thesis, Wageningen University.
Hague, R., Harrop, M., & Breslin, S. (1998). *Comparative government and politics: An introduction* (fourth ed.). Basingstoke: Macmillan.
Haily, J. (2000). Learning for growth: Organizational learning in south Asian NGOs. In D. Lewis, & T. Wallace (Eds.), *New roles and relevance: Development NGOs and the challenge of change* (pp. 52–72). Bloomfield, CT: Kumarian Press Inc.
High Level Forum on Aid Effectiveness, (2008). *Accra Agenda for Action*. Accra, Ghana: Author.
Hodson, R., Martin, A., Lopez, S., & Roscigno, V. (2012). Rules don't apply: Kafka's insights on bureaucracy. *Organization, 20*, 256–278.
Holvoet, N., Gildemyn, M., & Inberg, L. (2012). Taking stock of monitoring and evaluation arrangements in the context of poverty reduction strategy papers: Evidence from 20 aid-dependent countries in Sub-Saharan Africa. *Development Policy Review, 30*(6), 749–772.
Hood, C. (2012). Public management by numbers as a performance-enhancing drug: Two hypotheses. *Public Administration Review, 72*(S1), 585–592.
Hyden, G. (2008). After the Paris Declaration: Taking on the issue of power development. *Development Policy Review, 26*, 259–274.
IICA (Inter-American Institute for Cooperation on Agriculture), (2009). *The Outlook for Agriculture and Rural Development in the Americas*. San Jose, Costa Rica: Author.
Imas, L., & Rist, R. (2009). *Road to results. Designing and conducting effective development evaluations*. Washington, DC: The World Bank.
IOEC, (2014). The International Organization for Cooperation in Evaluation. Retrieved August 29, 2014: <http://www.ioce.net/en/index.php>.
Khaleghian, P., & Das Gupta, M. (2005). Public management and the essential public health functions. *World Development, 33*(7), 1083–1099.
Kremer, M., & Glennerster, R. (2012). Chapter 4, Improving health in developing countries: Evidence from randomized evaluations. In M. Paul, T. McGuire, & P. Barros (Eds.). *Handbook of health economics* (Vol. 2). Walthma, MA: North Holland Publications.
Laforce, J., & Silva, E. (2013). *Evaluacion del Programa Desnutricion Cero*. La Paz, Bolivia: Canadian International Development Agency.
Leroy, J. L., Habicht, J. P., Pelto, G., & Bertozzi, S. M. (2007). Current priorities in health research funding and lack of impact on the number of child deaths per year. *American Journal of Public Health, 97*, 219–223.
Lindblom, C. (1959). The science of muddling through. *Public Administration Review, 19*, 79–88.
Lindenberg, M., & Bryant, C. (2001). Accountability, evaluation and organizational learning. In M. Lindenberg, & C. Bryant (Eds.), *Going global, transforming relief and development NGOs* (pp. 209–242). Bloomfield, CT: Kumarian Press Inc.
Lipsky, M. (1980). *Street level bureaucracy*. NY: Russell Sage Foundation.
Locke, E., & Latham, G. (2006). New directions in goal-setting theory. *Current Directions in Psychological Science, 15*(5), 265–268.
Mackay, K. (2006). *Institutionalization of monitoring and evaluation systems to improve public sector management*. Washington, DC: World Bank.
Mahler, J. (1997). Influences of organizational culture on learning in public agencies. *Journal of Public Administration Research and Theory, 7*(4), 519–540.
May, E., Shand, D., Mackay, K., Rojas, F., & Saavedra, J., (2006). *Towards the Institutionalization of Monitoring and Evaluation Systems in Latin America and the Caribbean: Proceedings of a World Bank/Inter-American Development Bank Conference*. Washington, DC: World Bank.
Mayer, J. P. (1956). *Max weber and german politics*. London: Faber and Faber.
Mazmanian, D. A., & Sabatier, P. A. (1983). *Implementation and public policy*. Glenville, Ill: Scott, Foresman.
Mills, A. (2012). Health policy and systems research: Defining the terrain; identifying the methods. *Health Policy and Planning, 27*, 1–7.
MOH Ministry of Health, (2010). *Plan Etrategico Institucional 2010-2014*. La Paz, Bolivia: Author.
MPD (Ministry of Planning and Development), (2007). *Decreto Supremo 29272: Plan Nacional de Desarrollo: Bolivia Digna, Soberana, Productiva y Democrática Para Vivir Bien, Lineamientos Estratégicos 2006–2011*. La Paz: Gaceta Oficial de Bolivia.
Morales, N., Pando, E., & Johannsen, J. (2010). *Comprendiendo el Programa Desnutricion Cero en Bolivia: Un Analysis de Redes y Actores*. La Paz, Bolivia: Inter American Development Bank.
Naidoo, I. (2013). Growth and integration in the evaluation profession: Some perspectives for consideration. *American Journal of Evaluation, 34*, 572–576.
Natsios, A. (2010). *The clash of the counter-bureaucracy and development*. Washington, DC: Center for Global Development.
OED (Operations Evaluation Department) (2004). *Evaluation capacity development: OED self-evaluation*. Washington, DC: World Bank.
OECD/DAC (Organization for Economic Coordination and Development/Development Assistance Committee), (2005) *Paris Declaration on Aid Effectiveness*. Paris: Author.
Ordoñez, L., Schweitzer, M., Galinsky, A., & Bazerman, M. (2009). *Goals gone wild: The systematic side effects of over-prescribing goal setting. Working Paper 09-083*. Cambridge, MA: Harvard Business School.

Osterloh, M., Bruno, F., & Homberg, F., (2007). Performance evaluation and pay for performance: Does it really motivate public officials? Working Paper for the EGPA Conference in Madrid, 2007.
Patton, M. Q. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York: The Guildford Press.
Pelletier, D., Corsi, A., Hoey, L., Faillance, S., & Houston, R. (2011a). The Program Assessment Guide: An approach for structuring contextual knowledge and experience to improve the design, delivery, and effectiveness of nutrition interventions. *Journal of Nutrition, 141*(11), 2084–2091.
Pelletier, D., Frongillo, D., Frongillo, E. A., Gervais, S. G., Menon, P., & Ngo, T. (2011b). The nutrition policy process: The role of strategic capacity in advancing national nutrition Agendas. *Food and Nutrition Bulletin, 32*(2), S59–S69.
Pelletier, D., & Pelto, G. (2013). From efficacy research to large-scale impact on undernutrition: The role of organizational cultures. *Advances in Nutrition, 4*, 687–696.
Potter, C., & Brough, R. (2004). Systemic capacity building: A hierarchy of needs. *Health Policy and Planning, 19*, 336–345.
Rittel, H., & Webber, M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*, 155–169.
Rogers, P., & Fraser, D. (2014). Chapter 9. Development evaluation. In B. Currie-Alder, R. Kanbur, D. Malone, & R. Medhora (Eds.). *International development: Ideas, experience and prospects*. Northamptonshire, UK: Oxford University Press.
Rose, R. (1991). Comparing forms of comparative analysis. *Political Studies, 39*, 446–462.
Savedoff, W., Levine, R., & Birdsall, N. (2006). *Will we ever learn? Improving lives through impact evaluation*. Washington, DC: Center for Global Development.
Shekar, M. (2008). Delivery sciences in nutrition. *The Lancet, 371*, 1751–51.
Simon, H. (1957). *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting*. New York: Wiley.
Sjöstedt, M. (2013). Aid effectiveness and the Paris Declaration: A mismatch between ownership and results-based management?. *Public Administration and Development, 33*, 143–155.
SNIS (Sistema Nacional de Informacion en Salud), (2010). Power point presentation: Propuesta de reestructuracion del sistema nacional de informacion en salud. La Paz, Bolivia: Author.
Stacey, R. D. (1996). *Strategic management and organizational dynamics*. London: Pitman Publishing.
Stake, R. (1995). *The art of case study research*. Thousand Oaks, CA: Sage Publications.
Staw, B., Sandelands, L., & Dutton, J. (1981). Threat rigidity effects in organizational behavior: A multilevel analysis. *Administrative Science Quarterly, 26*(4), 501–525.
Thomas, V. (2010). Evaluation systems, ethics and development evaluation. *American Journal of Evaluation, 31*(4), 540–548.
Tendler, J. (1997). *Good government in the tropics*. Baltimore, MD: Johns Hopkins University Press.
Walt, G., Shiffman, J., Schneider, H., Murray, S., Brugha, R., & Gilson, L. (2008). 'Doing' health policy analysis: Methodological and conceptual reflections and challenges. *Health Policy and Planning, 23*, 308–317.
Weber, M. (1946). Bureaucracy. In H. H. Gerth, & C. W. Mills (Eds.). *Essays in sociology* (pp. 196–244). New York: Oxford University Press.
Weber, M. (1978). In G. Roth, & C. Wittich (Eds.). *Economy and society* (Vol. 2). Berkeley: University of California Press.
Weick, K., & Sutcliffe, K. (2005). Organizing and the process of sensemaking. *Organization Science, 16*(4), 409–421.
Westley, F., Zimmerman, B., & Patton, M. Q. (2007). *Getting to maybe: How the World is changed*. Toronto: Vintage Canada.
Yanez-Pagans, M., & Michacado-Salas, G. (2014). Bureaucratic delay, local-level monitoring, and delivery of small infrastructure projects: Evidence from a field experiment in Bolivia. *World Development, 59*, 394–407.
Yin, R. K. (2003). *Case study research: Design and methods* (third ed.). Thousand Oaks, CA: Sage.