## 1 Introduction

Most research on performance measures begins with the distinction between principals and agents. It assumes that agents act opportunistically, which is why principals need measures and data to monitor task implementation. Public service provision is known for its many levels of principal–agent relationships: within and across government organizations as well as built into external contracts and grants. This is the environment from which most research on performance systems has grown. It includes work concerned with measurement questions, followed by research on data use and nonuse in decision-making, and the link between performance management and organizational improvements. Scholarship has also examined dysfunctions, contradictions, and behavioral biases related to performance management (e.g., Ammons, 2020; Gerrish, 2016; James et al., 2020; Kroll, 2015a).

Most performance systems are set up within the boundaries of one organization's or jurisdiction's responsibilities to manage subagencies, contractors, or employees. However, effective public service provision increasingly requires independent organizations (across sectors) to work together. Such collaborations are voluntary, and they occur outside of the traditionally hierarchical structure of the bureaucracy. They can involve organizations from different levels of government, nonprofit organizations as well as civic and community groups, particularly if the issues they try to address are complex and cannot simply be fixed via government intervention or regulation (Bryson et al., 2015; Emerson & Nabatchi, 2015a; Isett et al., 2011).

Examples of collaborations include initiatives dealing with traffic congestion in urban areas, mitigating alcohol consumption in the amateur sports community, or addressing environmental watershed issues (Douglas & Ansell, 2021; Imperial, 2004; Page et al., 2015). In other instances, collaborative governance regimes help coordinate efforts among otherwise independent government agencies or departments. One such case is the collaboration between the U.S. Border Patrol and the Forest Service to better operate a segment of the Mexico–U.S. border (Emerson & Nabatchi, 2015b). Another example includes collaborations among agencies within the federal government using cross-agency priority goals (Choi & Moynihan, 2019).

The collaborations mentioned earlier vary greatly regarding their policy area and context, but they share a commonality. The researchers studying these cases make suggestions on how to adapt performance regimes to the specific needs of collaborative settings. I will review the specific ideas proposed in these pieces, and in other relevant work, in more detail later. Taken together, however, they all point to the following broader questions: How do we conceptualize

"performance" in collaboratives that consist of diverse actors with different interests and objectives? How can we use *shared* performance measures to manage for results in the absence of hierarchical and principal–agent relationships? What can we do to facilitate the collective use of shared measures and data in collaborative decision-making?

## 1.1 Shared Measures

When referring to shared measures, I use them as an umbrella concept that incorporates shared goals, indicators, and data. While this is an expansive use of the term *measure*, it is also intuitive and understood by government practitioners. Managing with performance measures implies that such measures are linked to goals and that data will be collected based on these measures. In that sense, "shared measures" is short for an entire set of shared performance management practices.

Shared measures have at least one similarity with more conventional performance metrics: they are the result of a systematic and routine management effort. That is, performance feedback is generated based on quantitative indicators that are supposed to capture the achievement of a predefined objective (Ammons, 2020; Hatry, 2006). Hence, shared measures – and performance metrics more generally – are different from all types of non-routine feedback (Kroll, 2013). The latter includes performance information that is ad hoc, episodic, verbal, and often not actively pursued but passively received via several sources (Mintzberg, 1975; Olsen, 2017; Tantardini, 2019).

What makes shared measures different from other routine performance practices is their use in the context of collaborative governance regimes. Such regimes have been characterized as "processes or structures … that engage people constructively across the boundaries of public agencies, levels of government, and/or the public, private and civic spheres in order to carry out a public purpose" (Emerson et al., 2012, p. 2). In this context then, I employ the following definition: *Shared measures are goals, indicators, and data that were agreed upon by a collaborative, span across organizational boundaries, and capture quantifiable changes in output or outcome performance for which the collaborative is jointly responsible*. As such, it is not the metric that makes a shared measure different from a conventional performance indicator, but the collaborative processes through which the metric is selected, defined, and used. Of course, my definition is that of an ideal type, acknowledging that, even in collaborative settings, some performance practices may be "more shared" than in others.

*Shared measures are agreed upon by a collaborative*. In the absence of a hierarchy, goals and measures are not set by a principal and then worked toward by an agent. Instead, the members of the collaborative need to identify and agree upon goals, measures, and data sources as the result of a group effort. *Shared measures span across organizational boundaries*. The rationale behind forming collaborations is commonly that a single organization is unable to solve a complex problem without other groups or agencies cooperating. If the production of outputs or the improvement of outcomes requires multiple actors to work together, then the key performance measures need to reflect this collective effort. *Shared measures capture changes in performance for which the collaborative is jointly responsible*. Members of collaboratives are the ones defining goals and measure as well as implementing actions to achieve these goals. Often, there is no external overseeing body whose purpose is to monitor the collaboration members and hold them accountable. Hence, the collaborative as a whole needs to take on responsibility for performance improvements (or a lack thereof) and self-manage.

There are performance practices that involve collaboration but do not necessarily fall under my definition. For instance, collaborative performance summits may not necessarily use shared measures (Douglas & Ansell, 2021). While summit attendees represent a diverse set of organizations that work toward the same goal, each organization may use its own separate performance system in lieu of shared measures. This is particularly true if the summit members' systems are actor-centric as opposed to network-centric (Douglas & Ansell, 2021, p. 953). In such cases, performance practices would not be jointly designed, and they would lack shared responsibilities. So, while performance summits certainly have collaborative elements, such as boundary-spanning lesson learning, they do not necessarily make use of shared measures.

PerformanceStat systems, such as CompStat or CitiStat, most likely do not use shared measures. While *Stat* meetings bring together different actors from various departments and sometimes even organizations, they miss the "sharing part" that is supposed to occur among partners who work together on equal terms. In fact, most Stat approaches are very hierarchical with a clear principal–agent structure, where top-level leaders use measures and data to hold middle managers accountable for changes in their unit's performance (Behn, 2014; Pasha et al., 2021). While PerformanceStat approaches can vary in terms of their aggressiveness, they rarely mirror the idea of a horizontally structured and voluntarily formed collaborative in which partners have equal say regarding the design and use of shared measures. In its more aggressive configuration, Stat puts managers in the "hot seat," where they are quizzed or even "cross-examined" about the performance deficits of their individual units.

Learning forums may use shared measures, but they often do not. First and foremost, learning forums tend to coordinate activities within one agency rather than across several organizations or groups (James et al., 2020; Moynihan, 2008). They facilitate a dialogue among actors from different hierarchical levels, but whether such an approach is truly collaborative may vary across organizations. Another feature that is often missing is that learning forums focus on the use and sensemaking of existing systems rather than jointly developing shared performance practices. As such, learning forums can serve as a vehicle for the use of shared measures; however, in practice, they rarely do.

In this subsection, I have provided a definition of the term shared measures, which I hoped to be broad and inclusive. I also reviewed a few popular concepts and approaches, some of which lend themselves to the application of shared measures. However, not every performance system that features one or more collaborative element(s) fits under the shared measures label. This distinction is important because "shared measures theory" may only apply to cases where collaboration members are truly able to share in the development and use of performance practices.

## 1.2 Approach of This Element

### *1.2.1 Argument*

Collaborations should make use of performance practices. And they do, as the previous examples illustrate. As such, the general logic behind the performance management principle will likely also apply to collaborations. Defining goals and using measures and data to track achievements may help groups to close performance gaps and yield improvements. At the same time, adoptions of performance approaches should be informed by the experiences made within the traditional agency setting. This includes awareness of misleading assumptions about rationality and objectivity as well as the unintended consequences of overly rigid systems that may, in fact, incentivize dysfunctional responses and behaviors (Heinrich & Marschke, 2010; Kroll, 2015a; Moynihan, 2009).

However, one must be cautious with simplistic transfers of "systems" from the traditional agency to the collaboration context. This is mostly because important administrative routines featured in public organizations are largely absent in collaborative arrangements. Think, for example, of performance contracts or bargains, reporting requirements, and traditional top-down monitoring that are present in most traditional performance systems but almost impossible to recreate (and probably unwanted) in horizontally configured collaborations. Hence, some of the content of performance management could be adapted to collaborative settings (definitions of performance, requirements

for high-quality measures, etc.), whereas related processes, such as *how to* develop measures or facilitate the use of data, need to be comprehensively revised.

I argue that the collective use of performance information in collaborations is different from organization-centered use, which I categorize as being either institutionalized or discretionary. Both of these uses are located within the public sector hierarchy and their focus is on managing individuals or organizations. Data use is institutionalized if it is regulated by formal requirements or informal norms. The part of use that cannot be regulated – which is cognitive and shaped by individual judgment – I label as "discretionary." Collective use is different, in that it occurs within multi-organizational networks and its reference point is the group level. It refers to data use that is negotiated among equal partners, and performance information here likely fits my ideal-type definition of shared measures.

While collective use is different from the other two modes, some of what we have learned about institutionalized and discretionary use can be adapted to the "collaboration case." For example, lessons from interactive dialogue theory may travel well across contexts. Yet, explaining collective use means shifting away from agency theory, which emphasizes self-interest-driven incentivization, and better-involving stewardship approaches, which are built around the ideas of aligning goals and values. To understand the use of performance data, and shared measures in particular, in collaborations more comprehensively, I argue that it is necessary to turn to relational theoretical perspectives and approaches.

While relational theory has been applied to model relationships among organizations, it has been largely disregarded in the performance information use literature, which is mostly focused on formal system requirements, organizational features, or the individual data user. In this Element, I examine six relational perspectives and, hence, take a closer look at the role of group composition, egalitarian structures, social relationships, distributed leadership, group culture, and value congruence. Overall, studying the collective use of shared measures requires taking performance management research into a new, mostly unexplored direction since existing theories will need to be revised and expanded, given the unique characteristics of the collaboration context.

### 1.2.2 Research Design

To examine the argument outlined earlier, the Element employs what others have called an instrumental or explanatory cases study (Stake, 1995; Yin, 2017). The case will be used to understand and explain causal relationships,

specifically those between the collective use of shared measures and its potential antecedents. The case-study approach has been employed as it allows an in-depth inquiry into a complex phenomenon and is particularly useful if the boundaries between the phenomenon and its context are blurry (Yin, 2017). I use the case study to illustrate the extent to which the explanations of shared measures use vary from the use of performance metrics in traditional, hierarchical settings as documented in the extant literature. Furthermore, I am concerned with identifying more and less influential explanatory variables. Generalizations will be analytic rather than following the paradigm of sample-to-population inferences.

My case study is on opioid-response collaborations in North Carolina (2019–2020). Across the United States, the opioid epidemic poses a major governance and public health challenge. Between 1999 and 2019, about 500,000 people died from an overdose involving prescription and illicit opioids (CDC, 2021a). Recently, the COVID-19 pandemic seemed to have supercharged a death spike related to drug and opioid overdoses (Katz & Sanger-Katz, 2021). What started out with the overprescription of painkillers turned into a multifaceted "wicked problem" (Lee, 2018) that involves abuse, addiction, mental disorders, criminalization, and illicit drugs as well as social and economic hardship. In such a context, the case study is concerned with how communities in the state of North Carolina responded to the epidemic. A context factor that these communities share is that their efforts are organized via collaborations that involve local government agencies, nonprofit actors, and civil society.

The main reason for selecting this case is that it offers a de facto natural quasi-experiment that directly maps onto this Element's main research question. The School of Government (SOG) at the University of North Carolina (UNC) at Chapel Hill set up a program to help communities within the state address the opioid epidemic locally (Nelson, 2021; SOG, 2021). As a part of that program, ten community collaboratives received training regarding the management of such groups, including training on the use of shared measures. This then created the rare opportunity to observe how the use of shared performance practices played out across ten newly formed collaboratives that all received the same training, thereby allowing the examination of the impact of different group dynamics and configurations on such practices.[1]

[1] I served as one of the instructors providing SOG's training "treatment." Again, this treatment was given to all ten groups, and the interesting question was to study whether – and why – groups responded differently to the input. My research, despite my involvement in delivering the treatment, does not qualify as action research since the study participants did not take on the role of researchers, and the research interest was not tailored to answering a practical question unique to the case (Zhang et al., 2015).

The study applies an embedded single-case study design (Yin, 2017). The first unit of analysis is the "North Carolina case" or, more specifically, statewide patterns in responding to the epidemic that hold across community collaboratives. At this level, I draw most of my inferences in an attempt to identify generalizable relationships; important context factors specific to the state and the SOG program are located here as well. I organize findings at this macro level by variables of interest (rather than sites) and – where helpful – pool data across groups. The second unit of analysis consists of the ten community collaboratives. Most of my data collection instruments employ the community level as their point of reference. Furthermore, at this level, I will contrast findings across two divergent community groups in more detail.

As conducting a case study allows the use of mixed methods (Yin, 2006), in order to strengthen the validity of the findings, I triangulate (a) methods, (b) data sources, and (c) the times of data collection. Regarding the methods, I use surveys, interviews, focus groups, and document reviews. With respect to data sources, my qualitative instruments collect data from "key players" (core group members and trainers[2]), whereas the quantitative surveys cast a much wider net. Table 1 shows which instruments were used at what point in time for the purpose of data collection.

All interviews were semi-structured and followed an open-ended question route (each took about forty-five minutes). In total, I conducted fifteen interviews (including those with trainers). The number of interviews was sufficient to gain rich descriptions of the perceptions of a selected few, but not large enough for more extensive text analysis. Rather, I wrote summaries of the

**Table 1** Data collection instruments

| Date | Instrument | N |
|---|---|---|
| 2019/03 | Main Survey (Part 1) | ~145 |
| 2019/09 | One-Page Survey (Forum 3) | 48 |
| 2019/11 | Review of Ten Performance Plans | – |
| 2019/11 | Focus Groups/Interviews with Ten Groups | 10* |
| 2020/02–07 | Interviews with Trainers and Follow-Ups | 5 |
| 2020/08 | Main Survey (Part 2) | ~120 |

**Note:** * Each of these focus groups/interviews was conducted with 1–4 individuals.

[2] Faculty and staff who worked on an interdisciplinary team that helped support and provide training and resources to the groups.

recorded material and contrasted major similarities and differences across groups. The purpose of the interviews was to collect background information about the groups and their work products, while being aware that the perspective of most interviewees was that of the "in-group member." For the document analysis, I reviewed each group's performance plans, including vision documents, action plans, and performance indicator sheets.

The surveys allowed me to better quantify some of the findings and reach out to a larger sample of group members. The main surveys were administered at the beginning and end of the project, and they targeted the full population of group members. Providing reliable response rates is difficult because the surveys were sent to all individuals who had given their contact information to one of the coordinators of a collaborative. However, some of these individuals only attended one meeting, left the group after a few meetings, or they were on the contact list by mistake. In such cases, said individuals should be considered as falsely included in the sampling frame rather than being nonrespondents. An additional survey was distributed during one of the forums where about fifty representatives of all groups came together to meet with the SOG team and trainers.

The remainder of the Element is organized to position my arguments within the extant literature (Section 2), propose a set of causal mechanisms (Section 3), conduct an empirical case study (Section 4), and draw out conclusions for theory and practice (Section 5).

## 2 Institutionalized, Discretionary, and Collective Data Use

This section categorizes the existing performance management literature into the streams of institutionalized and discretionary data use. This is done to differentiate these two perspectives, which were both established in the context of the public-sector hierarchy, from collective use that I associate with collaborations. To that end, this section provides conceptualizations of all three frames in which data use occurs. It synthesizes theories and empirical work to illustrate similarities and differences across these frames and identify the nature of, and mechanisms behind, collective use. However, I begin with a brief review of what the collaboration literature says about performance measures.

### 2.1 Perspectives on Performance Measures in Collaboration Research

Collaboration research incorporates the topics of performance and performance management, with an emphasis on questions of accountability (Agranoff, 2007). Bryson and colleagues (2006) consider accountability to be "a

particularly complex issue for collaborations because it is not often clear whom the collaborative is accountable to and for what" (p. 51). Kettl (2006) points to the same problem – "if everyone is in charge, is anyone in charge?" (p. 17) – and McGuire (2013) identifies creating accountability as one of the core competencies for the purpose of effectively managing networks. While most people would agree that addressing accountability-related questions remains a significant challenge for collaborations, it is noteworthy that research (mostly produced outside of the collaboration realm) has shown that accountability is only one out of many purposes for which performance data and systems could be used (Kroll, 2015a; Moynihan, 2009; Van Dooren et al., 2015). In that sense then, performance data – or more specifically shared measures – could add a great deal of value for managing collaborations if used for alternative purposes, including to evaluate, control, budget, motivate, promote, celebrate, learn, and improve (Behn, 2003).

Other work is concerned with the conceptualization of what performance means for cross-organization collaborations. For instance, Emerson and Nabatchi (2015b) argue that for a collaborative regime to be considered effective, it needs to satisfy stakeholders with different interests and priorities across several potential performance dimensions. Page and colleagues (2015) use a public-value perspective to develop a framework that not only draws on widely known subdimensions of performance (efficiency, effectiveness, equity) but also incorporates novel concepts such as capacity building. Moynihan and colleagues (2011) point to the fact that defining performance and selecting shared measures can be difficult in collaborations because authority is dispersed, and principal–agent roles are blurred. Along these lines, performance practices can help better manage collaborations (Imperial, 2004), but collaborators have a variety of implementation options to choose from (Douglas & Ansell, 2021), and poor adaptations to the collaborative context can be harmful as well (Denhardt & Aristigueta, 2008).

Most research so far has given little attention to the theme that is the focus of this Element: understanding differences in the use of shared measures for group decision-making. Page (2004) labels such behavior as managing for results, and he speaks more specifically of "the capacity to use data about results strategically to assess progress and to improve policies and operations in the future" (p. 593). Yet, we find little scholarship on the empirical underpinnings of shared measures use. One exception is a study by Choi and Moynihan (2019), wherein they examine interagency collaborations in the U.S. federal government and, among other things, emphasize the adverse effects between existing agency-focused systems and collaborative performance management efforts.

This Element is about the use of performance data for decision-making or, more specifically, the collective use of shared measures for managing collaborations. One prominently featured typology that clarifies the term "data use" is known as Moynihan's (2009) four Ps: purposeful, passive, political, and perverse use. In a nutshell, purposeful use is the type that is normatively desired and often associated with improvements, learning, and better-informed management. Passive use is about complying with data reporting requirements. Political use refers to employing data as ammunition in negotiations. Finally, perverse use captures dysfunctional behaviors such as data manipulation and gaming. Most research on these types of use has been conducted within the traditional public-sector hierarchy (for an overview, see Kroll, 2015a; Moynihan & Pandey, 2010).

To conceptualize data use in a collaborative setting, I take a step back from purposes of use and instead theorize "use" based on different organizational structures in which it is embedded. I begin with a hierarchical perspective that distinguishes between institutionalized and discretionary use. I then shed light on the collaborative environment that features collective use. The purpose behind this categorization is to articulate the idea of collective use by contrasting it against the two more conventional modes. While my typology of the three modes of use is novel, I will show that a large amount of the existing empirical literature fits under the umbrella that these three ideal types provide. In fact, I will synthesize and connect theories and empirics related to institutionalized and discretionary use to better substantiate our understanding of collective use.

## 2.2 Modes of Data Use: Institutionalized, Discretionary, and Collective

Conceptualizations such as the four Ps define performance information use by the varying purposes they serve. My starting point here is different, in that I distinguish modes of use based on the organizational frame in which they occur. Such a frame can be explicit or implicit. In an explicit frame, use is institutionalized via formal or informal norms, while in an implicit frame, use tends to be discretionary. The explicit frame captures the part of use that is organized through the establishment of structures, rules, and values that configure the relationships among actors or agencies. In contrast, the implicit frame describes the part of use that is more internal and rather difficult to structure and regulate. The explicit-implicit distinction makes few normative assumptions. Essentially, all purposes of use that the four Ps feature could occur under either the explicit or implicit frame.