Rating Governmental Excellence

Cary Coglianese
University of Pennsylvania Law School

Discussion Paper for the
Penn Program on Regulation's
International Expert Dialogue on
"Defining and Measuring Regulatory Excellence"

March 19-20, 2015

University of Pennsylvania Law School

Rating Governmental Excellence

Cary Coglianese
University of Pennsylvania Law School

Rating and measurement systems abound in contemporary life. *Michelin Guides* rate restaurants and hotels. *Consumer Reports* offers ratings for new washing machines, microwave ovens, and a host of other consumer products. Movie reviewers summarize their assessments using symbols that range from stars to thumbs up to the ripeness of tomatoes. *U.S. News and World Report* ranks colleges and universities (and the Obama Administration wants to start its own system for rating higher educational institutions too). Accreditation standards define the attributes of quality that hospitals, schools, and other institutions must meet. Regulators in some states have created a "hygiene" rating for restaurants. A host of systems for rating corporations exist to guide investors, from the Institutional Shareholder Services' Corporate Governance Quotient to the Dow Jones Sustainability Indices. A variety of popular magazines regularly rank the "best cities" in which to live, whether for unmarried individuals, retired persons, outdoor enthusiasts, and so forth.

Most of these rating systems exist to help guide choices, especially by consumers or investors. These systems articulate a set of criteria or attributes of quality, and then in some fashion aggregate the various attributes to achieve an overall rating or score. For example, *Consumer Reports* generates an overall rating for cell phones

> based mainly on Ease of use, Messaging, Web browsing, Display quality, Voice quality, Phoning, Battery life, Camera Image and Video quality, and Portability. Music, camera, and other features and capabilities are also considered. The [overall] score is out of a total of 100 points.

In general, measurements of quality depend upon both the identification of *attributes* to score as well as a method of *weighting and summing* these attributes to achieve an overall score or ranking. Presumably the most popular rankings or rating systems succeed because the attributes measured – and the weighting of them – generally mirrors the preferences of most users of these systems. After all, what consumer today does not want a cell phone that is easy to use and that offers sharp screen images along with crisp, clear sound?

Although rating systems abound to guide consumer choices, and many are very useful, they may also have their limits. For one thing, they will not be helpful if the attributes and weighting used by the raters does not match the preferences of an individual decision-maker. *Consumer Reports* may prioritize "ease of use" in a smart phone, for example, but a savvy, young computer engineer and a senior citizen are likely to care about that attribute differently. Parents

of small children probably find more useful than do other adults those movie rating systems that measure violence and sexual content. The Centers for Medicare and Medicaid Services cautions about over-reliance on its rating system for nursing homes:

> No rating system can address all of the important considerations that go into a decision about which nursing home may be best for a particular person. Examples include the extent to which specialty care is provided (such as specialized rehabilitation or dementia care) or how easy it will be for family members to visit the nursing home resident. As such visits can improve both the resident's quality of life and quality of care, it may often be better to select a nursing home that is very close, compared to a higher rated nursing home that would be far away.

And of course, even if a rating system captures the "right" attributes, it still has to measure them accurately, which is not always guaranteed. For example, a restaurant's hygiene scores are typically based on the results of a single visit by a health inspector; they also don't guarantee that countertops will be wiped down cleanly on the day that you dine there.

It is also possible for rating systems to miss the "forest" by focusing on the "trees." Studies of corporate governance rating systems, for example, have found that the rankings they provide do not necessarily correlate well with firms' actual financial performance, which is presumably what investors care about most. In the wake of the 2008 financial crisis, credit rating agencies have been subjected to intense criticisms for favorable ratings given to Lehman Brothers and other firms heavily invested in risky mortgage-backed securities. Ultimately, the sum of the parts does not necessarily lead to an accurate "whole" assessment of quality.

With these various considerations in mind, what are we to make of the use of performance measurement systems in the context of governmental entities, in particular regulatory authorities? Rating systems do abound, after all, in the governmental sphere. Management consultants have applied a range of assessment tools, such as the Balanced Scorecard or Six Sigma, to governmental organizations. The financial news site, *24/7 Wall St.*, issues an annual survey of the "best and worst run states in America" (with the best-run state in 2013 apparently being North Dakota). The federal government has formally institutionalized its own performance measurement systems; examples include the annual program performance reporting called for under the Government Performance and Results Act (GPRA) and the six-year experience with the Program Assessment Rating Tool (PART) used during the Bush II Administration.

Like their private-sector counterparts, these governmental rating systems can have value for decision-makers, but they may also present similar limitations to those present with consumer or investor ratings: i.e., they might not rely on the "right" attributes; errors might arise in measuring the attributes; the weights given to different attributes by the rater might

differ from the weights others think they should have; and the sum of the attributes might not lead to the resulting "whole" that the decision-makers care about most.

In addition to whatever general limitations exist with rating systems, the application of rating systems to government might well pose some distinctive issues.

First, who is (or should) be the target user of rating systems for governmental entities or programs? Perhaps rating systems in government should be intended first and foremost to inform the voting public. Perhaps, but voting decisions are seldom based on ratings of governmental performance; instead, voting usually is based on factors such as party ideology, candidates' characteristics and personalities, and even an overall "sense" about governmental performance based on conditions in the world (e.g., "peace and prosperity").

Even if rating systems do not primarily help voters, they could be (and are) used by government officials themselves in managing and overseeing programs and personnel. This is precisely the use contemplated by GPRA's performance management scheme. It is also the use contemplated by many educational testing requirements that American states have adopted, namely, to use test scores to decide which teachers or principals to promote or fire. Yet, when performance measures are used to evaluate employees and provide internal incentives, they may also crowd out intrinsic motivations and lead to problems captured under the banner of "teaching to the test." Shelley Metzenbaum, who headed up the Office and Management and Budget's responsibilities for implementing GPRA in the Obama Administration, has cautioned about overreliance on government rating scores for management decisions in government:

> [P]erhaps the biggest problem is that [directly linking incentives to performance measures] mistakenly suggests that the true objective of performance management is hitting a target rather than improving performance and increasing public-value return on investment. Many of us working in and with government are trying hard to reset this mistaken mind-set, treating target attainment as the purpose rather than a means to an end. It is my hope that researchers, in choosing areas and methods of study, will redirect their inquiries to the real purpose of performance management: continually finding and applying government practices that work better (Metzenbaum, "Performance Management: The Real Research Challenge," *Public Administration Review*, 2013)

Of course, the potential for misuse of performance measurement systems exists in any setting where ratings are used to measure the performance of individuals, teams, or organizations – whether in the private or public sector. But, if rating systems in the public sector are primarily intended to be used for managerial decisions (as opposed, say, to informing consumer or investor choices), concerns about misuse or misaligned incentives may well take on heightened importance when used to rate governmental performance.

A second potential concern about rating systems in the governmental sphere relates to the relative importance given to the "parts" versus the "whole." The specific attributes, or parts, of an electronic product like a cell phone do matter to people, so it makes a lot of sense to rate such products based on these attributes (e.g., display quality, battery life, etc.). With respect to governmental programs or agencies, do the specific parts matter as much? Or is it the outcome of a program or agency (the "whole") that matters most? Is the air getting cleaner? Is the economy prospering? Are highways safe? To be sure, citizens do and should care about certain attributes or parts of a governmental entity, such as its fidelity to democratic principles, its transparency, and so forth. Indeed, what we know from social psychologists about procedural justice suggests that, in addition to substantive outcomes, people care about the nature of their interactions with government; they care about process and *how* they are treated. Nevertheless, on many attributes that might be used to measure governmental quality, perhaps few will care very much about the specific attributes of the program or agency, such as its organizational practices and its processes. Perhaps as long as government "works," it matters little to many people whether governmental entities organize their routines in specific ways, what kind of human resources and IT systems they deploy, whether they use specific policy tools (e.g., performance standards versus design standards), or whether they rely on adversarial versus cooperative enforcement strategies. One might well imagine that if Rome is literally burning (or if it is prospering), few people will ultimately care if its governmental entities check all the boxes in a rating system of regulatory quality.

Finally, the application of rating systems to the governmental sphere may be complicated by the fact that government's performance – especially the performance of government regulators – is ultimately dependent on the performance of others, namely those they regulate. Unlike the rating of a manufacturer's cell phone, which can be based on the phone that the company produced and is in the tester's own hands, a regulator's performance is literally in the hands of someone else (the regulated entity). This not only creates some difficulties in accurately measuring a regulator's performance (and especially comparing across different regulators), but the multi-layered nature of that performance may well hold two other important implications. First, a regulator could well rate very highly on any number of attributes (e.g., it is highly transparent about its rules; it treats its employees well and trains them to meet high professional standards; etc.), and yet, for whatever reason, the industry it regulates might still experience a disaster that the regulator was supposed to prevent. In other words, since responsibility for risk control in the regulatory sphere is by necessity shared between the regulator and the regulated, a failure by the latter will inevitably be viewed as a failure on the part of the former, notwithstanding even high performance of the former, at least in terms of attributes that make up a rating system.

Second, the converse is also true. That is, *good* outcomes on the part of the regulated community might not really correspond, causally, to how a regulator scores on a rating system. It may look like a regulator is doing well, both because it is "hitting its marks" and because outcomes look good, but the regulator's performance on scored attributes really may have little

to do with the good outcomes observed in the world. For example, in the United States, the Environmental Protection Agency might very well be rated as a well-run, analytically sophisticated regulatory agency, and yet the overall improvement in environmental quality in the United States over the last several decades may have come about largely because of a shift in the U.S. economy away from manufacturing to services – something unrelated to the work EPA does. Does an excellent rating therefore mean the same thing if the outcomes in the world happen to be disconnected from what is being rated?

The inherent nature of a regulator's challenge – that is, of trying to control outcomes caused by the behavior of others – may well mean that thinking about what makes an excellent *regulator* is a lot like thinking about what makes an excellent *parent*. The measure of success for both of them is irreducibly out of their hands. Probably we all know – or can at least imagine – parents who are by all accounts quite excellent (e.g., caring, nurturing, wise, etc.), and yet at least one of their children turned out to be rather self-centered, rude, needy, or indolent as an adult. On the other hand, examples abound of highly successful, self-actualizing individuals who nevertheless had parents who were, if not abusive, at least neglectful and decidedly subpar. If a child's successful maturation is only at most partly affected by parenting quality, what implications would this have for a rating system of parental excellence?

The relationship between a regulator's attributes and its performance raises similar questions. Indeed, this relationship between attributes and performance would seem to capture a central, if not the central, issue in applying a rating system to a governmental organization such as regulatory agency. To be reliable, such a system needs to capture what matters much if not most of the time – even if it can never capture everything that matters all of the time. It may well be true, for example, that the offspring of *some* nurturing and attentive parents turn out to be miscreants, but presumably most do not. Similarly, without a doubt more individuals do struggle when they have grown up in inattentive and decidedly non-nurturing environments, even if the occasional Horatio Alger story can be told. Ultimately, the challenge in applying a rating system of governmental organizations may well be the challenge faced of all rating systems: capturing what matters most – either because the rated attributes are intrinsically valued or because those attributes are what, generally speaking, will be more likely to result in desired, intrinsically valued outcomes. What type of rating system to use in the governmental sphere may simply be the one that will best achieve the goal, aptly articulated by Metzenbaum, of "finding and applying government practices that work better."