The Power of Performance Indicators: Rankings, Ratings and Reactivity in International Relations

Judith G. Kelley, Duke University
Beth A. Simmons, Harvard University

Paper prepared for the meeting of the International Political Economy Society

© Draft. Please do not cite without authors’ permission

---

1 We are especially grateful to Sam Chase and Nadia Hajji for excellent research assistance. Kelley and Simmons are responsible for all errors.
Abstract

Global Performance Indicators (GPIs) are increasingly used to rank, rate or categorize states in a number of issue areas. Many of these new indicators are short-lived efforts to grab attention and are unlikely to matter much. But we believe there are good reasons to think that some GPIs affect important areas of state policy. Indeed, we argue that GPIs should be thought of increasingly as tools of global governance, involving rule-making and the exercise of soft power on a global scale and that their proliferation constitutes a profound social trend with implications for governance world-wide and reflects the diversity of actors and institutions attempting to influence policies across and among states. This article defines what we mean by global performance indicators, and describes their features. Using a new dataset and a series of interviews with producers of various indices, we document the proliferation of GPIs, which have been developed and promulgated by a wide range of actors, both public and private; unilaterally and multilaterally. We then elaborate possible causal mechanisms that we expect to connect externally generated GPIs with state policy, and hypothesize about the scope conditions for their effects.
In 2012, when Andrew Forrest, Australia’s richest man, decided to take up the cause of human trafficking, Bill Gates gave him some advice: use a quantifiable metric. In an interview, Forrest said: “Global modern slavery is hard to measure and Bill’s a measure kind of guy… [In] management speak, if you can’t measure it, it doesn’t exist.” 2 Forrest got fellow billionaire Virgin Group Ltd. Chairman Richard Branson on board, and by December 2012, they founded WalkFree.org, and had appealed to 25 companies to ban use of forced labor. The organization has since undertaken various campaigns, but one activity has garnered more attention than all of WalkFree’s efforts combined: in September 2013, the organization published its inaugural Global Slavery Index, which ranked states for their number of persons in “involuntary servitude” and official efforts to fight it in 160 countries. In the first month after its launch, the index was covered in over one-hundred newspaper stories around the world. In a global survey of non-governmental organizations that work on trafficking issues, 9 months after its release over 40 percent had already heard of this new index.3

Quantifiable metrics and ratings have influenced thinking at many levels of global governance. The World Bank’s Doing Business Report, which rates countries on the ease of ten sub-indicators from getting a business permit to resolving insolvency, has long been credited with bringing about reforms in countries – as many as 2000 distinct reforms since its 2003 launch (The Economist 2013). Countries sometimes openly state their plans to undertake reforms precisely to reach a certain place in the rankings (Williamson 2004). Georgia made concerted efforts to do so and managed to rise from 100th to the top 20 in two years (Schueth 2011, 52). Indeed, the perceived ability of the report to shape and encourage reforms has even led to a showdown between the Bank and some states who rate relatively poorly and would like to see the rankings weakened (The Economist 2013).

These are but two examples of actors who have come to believe that highly publicized global performance indicators (GPIs) pressure states to make policy changes that they believe will improve their relative standing in the indices. The trend to use systematic state assessments – numbers, ratings, rankings and categories – to publicize various aspects of state performance and attempt to influence states’ policies is growing. Private individuals, NGOs, think tanks, publishers, intergovernmental organizations and states have all, to some degree, participated in the performance assessment frenzy. From traditional credit ratings to such indicators as the Mother’s Index, the monitoring and rating of state performance has never been more extensive than it is today. In the last 15 years more than 8 new state-focused GPIs have been added on average each year, and today over 160 of these chart anything from gender equity to happiness.4

The importance of performance information is of course not new in international relations. Rational functionalist theories highlight the informational function of international institutions and emphasize the role of information in overcoming

---

2 Bloomberg story by Elisabeth Behrmann on 11 April 2013. Available at http://nickgrono.wordpress.com/2013/04/11/gates-helps-australias-richest-man-in-bid-to-end-slavery/. This quote is of course a derivative of the earlier claim by the British physicist, Lord Kelvin, who said "If you can not measure it, you can not improve it." See http://zapatopi.net/kelvin/quotes/
3 Survey conducted by Judith Kelley, summer 2014.
4 Earlier research has documented 178 indicators, but with a far broader definition than the one we employ in our research (Bandura 2008).
coordination and collective action problems among states (Keohane 1984). However, information today is often used in a more normative and intentional way than traditional liberal theories of institutions imply. GPIs are often an effort to exert influence states through “information politics:” the practice of quickly and credibly generating politically usable information and deploying is strategically so that it will have the greatest possible impact (Keck and Sikkink 1998, 18).

This strategic collection, packaging and dissemination of information about state performance is therefore an intriguing and potentially important development. We argue that GPIs should be thought of increasingly as tools of global governance (Davis et al. 2012), involving rule-making and the exercise of soft power on a global scale (Keohane 2003, Nye 2004). Some countries, such as Rwanda, have even designated bureaucrats with keeping an eye on their performance in various GPIs.5 GPI proliferation constitutes a profound social trend with implications for governance world-wide (Espeland and Sauder 2007, 2) and reflects the diversity of actors and institutions attempting to influence policies across and among states (Avant, Finnemore, and Sell 2010, Hale and Held 2011, Hale and Roger 2014)

The rise of GPIs raises some important questions: What are the effects of monitoring and ranking states in various ways? Do rankings influence states’ policies and practices? What – if anything – makes an indicator influential? Are GPIs merely the perpetuation of existing power structures? Do some GPIs mediate existing power structures, or do they represent a significant power shift to actor with the ability to collect and deploy believable information?

This article defines what we mean by global performance indicators, and describes their features. Using a new dataset and a series of interviews with producers of various indices, we document the proliferation of GPIs, which have been developed and promulgated by a wide range of actors, both public and private; unilaterally and multilaterally. Many of these new indicators are short-lived efforts to grab attention and are unlikely to matter much. But there are good reasons to think that some GPIs affect important areas of state policy. Accordingly, we elaborate possible causal mechanisms that we expect to connect externally generated GPIs with state policy, and hypothesize about the scope conditions for their effects in hopes of laying the ground for future testing of these ideas.

I. Global Performance Indicators: definition and trends

We define a global performance indicator (GPI) as a public, comparative and cross-national indicator that governmental, intergovernmental and/or private actors use regularly to attract attention to the relative performance of countries in a given policy area. GPIs are generally understood to “provide information on matters of wider significance than what is actually measured or make perceptible a trend or phenomenon that is not immediately detectable” and hence go well beyond mere data (Hammond 1995, 1). While such indicators exist to rank a broad range of entities – from cities to

hospitals, from universities to firms – we are interested specifically in those indicators that rate some phenomenon or policy at the national level. This definition does not preclude a national ranking that is based in part on ratings of sub- or non-state entities but we focus on rating schemes for which the named entity is a state. In terms of coverage, GPIs may assess a range of phenomena, from state qualities (“transparency”) to state policies (“press freedom”) to prevalent social practices within a state’s jurisdiction (“corruption”). Our definition has much in common with a prior definition by Davis et al., who, as part of a larger project on “Governance by Indicators,” defined indicators as “a named collection of rank-ordered data that purports to represent the past or projected performance of different units” (Davis et al. 2012, 72). Our definition is narrower and more specific. Because we are interested in the use of GPIs as tools of social pressures on state, we focus specifically on those that are public, comparative, regular, inclusive, and purposive. We discuss each of these elements in turn below.

Public. In order for a GPI to gain relevance internationally, it must be public and easily available. This would exclude for example any indicator that is proprietary and must be purchased, such as various political risk indicators intended to assist investors. Such information certainly has its commercial uses, but does not have the social function we believe is the essential characteristic of the proliferation of today’s GPIs. However, if the World Bank publishes data relevant to political risk on its website (rule of law indicators are a good example), it becomes a candidate for a GPI by our definition.

Inclusive. A GPI must for our purposes include multiple states, with the aim (in principle) of full inclusiveness within a region or worldwide. Many practical reasons make it difficult for raters to monitor large numbers of states, and few of the indices we have found to date are comprehensive. However we exclude monitoring and rating systems that are intentionally limited to a handful of states.

Regular. We consider only GPIs that are issued regularly on a predictable schedule, for example annually. One-shot or occasional rankings may have some short-term impact, but they are likely to signal a lack of interest or commitment by the rater, and thereby reduce social pressure and any incentive by the rated to adjust their policies in anticipation of future rounds of ratings.

Comparative. We consider only comparative indicators as candidates for GPIs in this research. Comparisons can be made either through numerical assignments or the use of clear labels that render normative judgments, such as blacklists or watch lists (discussed in more detail below). To qualify as a GPI, it must be easy to sort and compare states using the proffered rating, ranking or categorization system. We exclude monitoring systems, such as Amnesty International’s annual reports, that result in

---

6 This excludes for example various city ratings, such as the Bike Friendly Index. See [http://copenhagenize.eu/index/](http://copenhagenize.eu/index/).

7 For example, on this basis we exclude Maplecroft's Human Rights Risk Atlas, which is “designed to help business, investors and international organisations assess, compare and mitigate human rights risk across all countries.” The data themselves are pay-walled and not publicly available. See [http://maplecroft.com/themes/hr/](http://maplecroft.com/themes/hr/).

8 While we do not include occasional monitoring and ranking systems within this study, we have collected a significant number of GPIs that have apparently “died;” that is, they were apparently in use for a few years, but have been discontinued for some reason. These are noted as “defunct” in Figure 2.
narratives but do not rate, rank or categorize states. This is critical because as we argue below, rankings are an especially potent lever of social pressure; they foster explicit comparisons that, once promulgated, are difficult to dislodge from public discourse (Andreas and Greenhill 2010). Such systems tend to simplify a complex reality, grab attention, and appear objective. Indeed, as we discuss below, many raters present a detailed methodology in an effort to appear as though hard data and legitimate criteria underlie what is ultimately, in most cases, a judgment call. Whether they are presented as rankings from first to last, ratings from excellent to poor, categories such as cooperative or non-cooperative or even express explicit social disapproval, such as watch lists and blacklists, the essential feature of performance indicators is the itchy feeling they create of being socially watched, judged and especially compared with peers.

**Purposive.** Most difficult to determine in practice, GPIs must be produced with an intention to influence policy or practices in the area at which they are aimed. The intended policy effect could be aimed specifically at public actors (specific officials, bureaucracies, legislators) or it could be aimed to influence policy indirectly via the actions of private actors (investors, NGOs, or voters). While apparently neutral “information” alone can also influence states without its producers intentionally aiming to do so, we are interested in systems that potentially represent the intentional exercise of social power through comparative information, broadly understood. Evidence of such intentionality includes costly efforts to gather new data; repackaging existing data and labeling it in ways designed to attract attention, the provision of accompanying recommendations, and even launching specific marketing efforts to increase a GPI’s visibility by alerting the media, instigating an internet or social media campaign to draw attention, and the like. We exclude the thousands of indicators whose primary purpose appears to be to passively support research, such as those that appear in the World Bank’s *World Development Indicators*, unless there is evidence that some actor takes active ownership and packages, publicizes, touts and even brandishes the GPI in a strategic way. In short, we are less interested in the unintended consequences of quantification (Davis et al.) and instead focus on information that is deployed strategically by purposive actors to influence outcomes.

Given this definition, we can now explore the empirical terrain. While indicators are hardly new – sovereign credit ratings first appeared in 1916 – the vast majority in use today were created after 1990 (Löwenheim 2008b). Charting their rise is challenging. Some GPIs may have been launched, say, in the 1960s, but have been discontinued for whatever reason and simply do not show up on any radar screen today. There are also

---

9 We also exclude comparative information that could clearly and easily be converted into a numerical index, but the source declined to do so. See for example the data displayed at [http://en.wikipedia.org/wiki/LGBT_rights_by_country_or_territory](http://en.wikipedia.org/wiki/LGBT_rights_by_country_or_territory).

10 On this basis we exclude for example the indexes listed at Aneki.com ([http://www.aneki.com/](http://www.aneki.com/)). There is no apparent policy purpose behind this list of rankings of everything from countries with the “largest class sizes” to “world’s most dangerous roads.”

11 Many GPIs that meet our criteria are in fact collections of “raw” data, repackaged and labeled often to draw attention to the creator’s purpose. For example, we do not consider a measure such as GDP growth or literacy rates to be a global performance indicator in our sense. However, when it is repackaged as a subcomponent of an indicator such as the *Basic Capabilities Index* – which combines such common data as under-five child mortality and adult literacy rates, among others – it contributes to a global performance indicator as used here.
some interesting silences – areas one might have thought an indicator might have been likely, but none are to be found (tobacco control measures, for example, have an index across US States, but not globally). However, our research to date has turned up over 160 measures that meet our five criteria and appear to still be actively updated as of 2011. Figure 1 presents a cumulative count (from their start date) for every active, eligible GPI we can locate. Defunct indicators are included for their active periods as well. We realize that more recent years may be bloated by indexes that may yet fly-by-night. Nonetheless the figures reinforce the general impression of other researchers that the use of reasonably durable GPIs began to grow significantly in the second half of the 1990s (Löwenheim 2008b), and even more so in the early 2000s. The nearly 40 active GPIs we could locate that had their start in the early 2000s are by now nearly a decade old. Even if we disregard the two most recent half-decades, reasonably durable GPIs seem to have taken off significantly by the late-1990s and appear to have accelerated since.

Economic GPIs are especially common. Figure 2 shows that, among the GPIs we found that meet our 5-point criteria and are still “active,” a significant plurality attempt to rate or rank economic activity. Social issues and development are also common areas


13 We located at least 39 indicators that had only a very short existence and/or do not appear to be actively updated as of this writing, but otherwise would meet our criteria for inclusion. See Figure 1.
subject to external GPIs. Rarer by far are GPIs relating to security, conflict or military issues. (Note that the total count of GPIs is larger in this Figure than in Figure 1 above because we have double counted cases that straddle issue areas, such as health and development.)

II. Background

Why such a proliferation? One might suspect that it can be explained at least in part by the growth of international laws and institutions, such as the United Nations and the World Bank, that collect data and create reporting and monitoring obligations. Information reporting is also a standard provision of many international treaties and its collection and dissemination is a common function of international institutions (Chayes and Chayes 1995b, 183-184, Dai 2007). However, increased legalization and institutionalization is a partial explanation at best for indicator proliferation. Many of the reporting regimes integral to international legal agreements are voluntary, ad hoc, confidential and result in narrative reports, findings or views (Chayes and Chayes 1995a,
Creamer and Simmons forthcoming), but rarely if ever entail comprehensive efforts to rate and rank states parties. Moreover, as Figure 3 shows, international bodies account for less than a quarter of all recent indicators; the rest are created by NGOs, individual states and private entities acting *unilaterally* and *extra-legally*. Thus, international law and institutions alone cannot account for this striking empirical reality. Indeed, IGOs monitoring and reporting systems – fraught with opacity, diplomatic discretion, and partial reporting compliance (Chayes and Chayes 1995b, 155-156) – can be thought of as the status quo that some of these new measures are meant to bolster or even to circumvent.

![Figure 3: GPI Creators, by Type](source: Authors’ database of GIs. N=165)

Rather, the emergence of governance indicators resulted initially from a conflux of factors. One might think of the evolution of indicators as two-phased. The first phase began in the 20th century, with the 1916 creation of sovereign credit rankings, but it really did not flourish until the 1970s and 1990s. This period included the creation of several ratings by Freedom House such as *Freedom in the World* and *Freedom of the Press*. In the 1990s, the *Human Development Index* and the World Bank *Governance Indicators* blazed a trail for many more to come. These indicators were created partly in response to a growing demand for policy data, which in turn was driven by increased international investment, geopolitical changes leading to new interests in improving governance and development, and the realization that many past development policy reforms were failing (Arndt 2008, Arndt et al. 2006). As part of the increasing push towards evidence-based policy making and accountability, policy makers and investors wanted better information for making decisions about aid allocations, investments, and foreign policy decisions. For example, Daniel Kaufman, then responsible for the World Bank’s *Governance Indicators*, noted in a 2004 Financial Times article that Transparency International’s *Corruption Perception Index* had “played a very important role in identifying corruption...
as an important focus for advocacy work” (Williamson 2004). Thus, the main purpose of most of these indicators was initially to provide better information to inform policy.

A considerable body of scholarship arose around these phase-one indicators. This work falls into three major groups. Many people deployed these indicators descriptively to detect changes and trends. Others used these indicators analytically, to test for causal relationships, for example, between growth and democracy; for example, the Freedom House indicators of states that are “free” and “not free” became commonplace in the policy and academic literature. The third research wave, which is huge, critically analyzed the indices themselves, questioning the methodology, construct validity, reliability, or the fairness of a given measure, sometimes even performing analyses revealing biases (Høyland, Moene, and Willumsen 2012, Ginsburg et al. 2005, Christiane 2006, van de Walle 2006). Indeed, some researchers turned their critiques into a “ranking of the rankings” (Hood, Dixon, and Beeston 2008). The assumption for the most part, however, was that indicators were useful ways to summarize some kind of objective reality that could inform better policymaking.

By the late 1990s, the number and quality of governance indicators began to change. The number of indicators, rankings and indexes exploded and many in this new wave were not only meant to provide useful policy information for decision makers; they were deployed specifically as “technologies of power” (Hansen 2011); that is, they were intended to influence the behavior of the rated. Today, they are created by private actors, NGOs, IGOs and states who want to publicize their judgments and to influence state policies by packaging information and selling it in a catchy and easy-to-digest format.

The comparative nature of these newer GPIs is by all appearances intentionally used to pressure and shame countries. This can be inferred from the way many GPIs are introduced to the public. When Transparency International (TI) released its Corruption Perceptions Index in 1995, it issued a press statement with the headline, “Indonesia Worst in World Poll of International Corruption,” and said its purpose was “to raise awareness of international corruption and to create a coalition of interests from both the public and the private sectors to combat it” (Transparency International 1995). TI is frank about its intentions: the goal of the index is clearly to influence policy (Wang and Rosenau 2001). When the OECD launched the Program for International Student Assessment (PISA) ratings in 2000, it was seen as an “[e]xercise of international public authority through national policy assessment” (Bogdandy and Goldmann 2008). Based on this and the success of other OECD indices, Ángel Gurría, the OECD’s secretary-general, said in 2013 that the organization would start scoring countries on climate change, because this would “help nudge countries to stop burning fossil fuels” (Clark 2013). Most recently, the Kofi Annan Global Commission on Elections, Democracy and Security encouraged a new global civil society organization to start “grading” elections in countries every year to “allow citizens to see how their country’s elections fare against international standards and, over time, to track whether electoral integrity in their country is getting worse or better,” specifically in the hopes that such grades would create “domestic pressure on national governments” (Global Commission on Elections 2012).

The explosion in this type of purposeful GPIs resulted from a conflux of factors. Actors began to realize that “indicization” of information promoted not only transparency and accountability (Mathiason 2004), it could also constitute a form of pressure to conform. States in particular began to realize that promulgating GPIs was a potentially
effective way to influence other states’ affairs even as normative prohibitions against forceful intervention gained strength and as economic interdependence made it costlier to use material levers such as sanctions to influence other states (Kim 2012). In short, GPIs in some cases were a flexible soft power resource that could be used to encourage some behaviors and policies and discourage others.

At about the same time, the cost of exerting pressure via information declined. While not costless, it has never been easier to collect and distribute apparently credible information from highly decentralized sources on a global scale than it is today. Technological developments have facilitated data gathering and distribution. Monitoring human trafficking for example relies on technologies from global positioning to web-based reporting platforms deployed to detect trafficking, to aggregate country profiles, and to disseminate assessments of government efforts to counter trafficking (Latonero 2012). Many types of data can now be crowd-sourced or automatically collected because it is digitally recorded and stored in formats susceptible to extraction and compilation. It is becoming increasingly easy to survey experts in real time and analyze and compile the results quickly.

Meanwhile, social media and growing internet penetration facilitate cheap global dissemination. Reports and news can be shared in a matter of seconds and ever more powerful search engines have made it easier to connect interested parties with the relevant information. Figure 4 suggests that many of the GPIs in this study reverberate broadly around the internet and are referred to on thousands of web pages. The figure graphs the internet “popularity” of GPIs resulting from a Bing search for each of the index names we have located. The vertical axis indicates the number of times Bing found the GPI named on a webpage (as of July 2014). Most of the GPIs we could locate were cited between a thousand and a million times on the internet, suggesting wide attention, but significant variance across GPIs. Twenty of the GPIs fitting our criteria were found on fewer than 1000 pages (not shown in Figure 4), but the grand prize went to Standard and Poors Sovereigns Rating List with over 1.7 billion citations on the World Wide Web (also not shown).

Research on this new type of GPI has developed more recently. One important strand of work criticizes the use of indicators, arguing for example that they oversimplify reality, strip concepts of their context and history, conceal their underlying theoretical origins, and offer a false sense of precision and certainty (Merry 2011, S84). Many of these points echo earlier criticisms of indicators’ construct validity. Some critics also note that indicators affect normative discourse: by calling itself an index of the “rule of law,” for example, an index is implicitly laying claim to the definition of that concept. Even if the label is not an accurate reflection of the existing consensus of a given concept, it can still contribute to how the concept comes to be understood (Davis, Kingsbury, and Merry 2012). By constructing “standards and scripts for action” it is argued that performance indicators determine “what constitutes legitimate social practice” (Hansen 2011).

This observation has led to a good deal of discussion about how indicators may impact power and authority relationships (Buthe 2012, Kelley and Simmons 2014, Davis, Kingsbury, and Merry 2012, Halliday 2012, 215). Arguably, the cumulative effect of
widespread quantification is to reinforce global power structures (Löwenheim 2008b). After all, rich states and wealthy people tend to do the rating, while the poor are often the rated. This is clearly illustrated in Figure 5. Most of the GPIs we have located are produced by actors anchored in the developed world. Indeed, among the active indicators that meet our criteria, more than half – 52 per cent – were developed and are promulgated by organizations headquartered in the United States. That said, as figure 3 showed, the variation in the types of actors that promulgate these indices could contribute to a dispersion of power in these countries.

Quantification may also serve to reapportion power by enfranchising technical experts who gain a new-found influence over political debate (Merry 2011). However, much more ground-breaking work on where and when indicators influence more immediate outcomes, such as aid flows and state policies, needs to be done before larger claims about global power relationships can be sustained.

Most importantly for what this paper attempts to do, however, is that relatively little research directly addresses the effects of global performance indicators on the behavior of the rated. There are a few case studies on the Extractive Industries Transparency Initiative (EITI), which is a set of standards “to improve openness and accountable management of revenues from natural resources.”14 There is hardly a consensus on the effects of the EITI’s compliance ratings: on the one hand, some studies find the EITI has stimulated useful debate over policies relating to the extractive industry; on the other, some studies raise questions about the ability of the EITI to really improve

---

14 See http://eiti.org/eiti.
governance in this sector (Short 2014, Smith, Shepherd, and Dorward 2012). A study about the “PISA effect” in Europe (Grek 2009) found that low educational rankings have shocked Germany, stimulating significant educational reforms. The United States’ Special 301 Report’s “Watch List” and “Priority Watch List” designates countries with lax intellectual property rights protection, and have been found in one study to have had some impact on IPR policy in China (Tian 2008). Blacklisting is among the “discursive, power-based mechanisms” of the Financial Action Task Force (FATF) found to be at the root of adopting anti-money laundering measures around the world (Sharman 2008). The United States Trafficking in Persons Report and its “Tier” ranking system has been linked to the criminalization of human trafficking policies around the world Kelley and Simmons (2014).

While other studies that actually look into the effects of ranking systems do exist, they are few and far between. Meanwhile, there are thousands of articles that use or discuss ranking systems, such as the Ease of Doing Business Rankings, which are peppered with claims that these rankings have spurred reforms (Williamson 2004, The Economist 2013). It is safe to say we know little about whether such claims are true and how broadly they apply across different types of indicators created by different types of actors across many issues. That is, we lack a theoretical framework for thinking about GPIs, the mechanisms though which they operate, and most notably what explains variation in their possible effects. The next section moves in this direction.

III. Why GPIs and why might they matter?

The proliferation of indicators suggests that the agents who fund their creation and promulgation expect some kind of effect. One motive, as discussed above, could simply be to inform better policy assessments or academic research. Another possibility is that GPIs are little more than public relations products to draw attention to the creators (Davis, Kingsbury, and Merry 2012, 82). Especially when developed by NGOs or private actors, an index is a catchy lure to bring attention to an organization and raise its profile within a given issue area. The organization Freedom House likely owes its prominent position among human rights and democracy organizations to the early promulgation of its “Freedom in the World” indicator. But one very likely motive for the creation of GPIs is to influence policy. Of course, there are many ways to influence states’ policies: lobbying, shaming and vocal criticism, and material incentives are all possibilities as well. This raises the question: why GPIs?

We argue that, under some circumstances, GPIs are an effective way to focus social attention on a state’s policies, qualities, or practices. First, they imply systematic observation. Second, they imply comparative and criteria-driven judgment. Third, ranking systems can be persuasive, and have the potential to impact public discourse.

*Systematic Observation: Monitoring*

Most performance indicators originate in some form of monitoring. Monitoring involves observing and checking the progress or quality of a policy, practice or condition over an extended period of time. It implies systematic review that is repeated, often even routinized. Research shows that monitoring has behavioral consequences: in experimental
settings subjects behave differently when they know they are being watched. Referred to as the “Hawthorne effect,” individuals may re-arrange their priorities to meet external expectations when they are aware of being observed (Adair 1984). Sociologists use the concept of reactivity – the tendency for people to change their behavior in response to being evaluated – to explain the effect, for example, of US News and World Report rankings on university priorities (Espeland and Sauder 2007). Reactivity has its social uses: though a nuisance to some research designs, it can potentially be harnessed in some cases to change behavior in desired ways (Espeland and Sauder 2007, 7).

Monitoring has long been theorized as a potent form of social control (Foucault 1995, 201-202). Its power lies in its latent potential to shame those who are revealed to “underperform.” When it is regularized and ongoing, targets may internalize the regime and potentially self-regulate (Sauder and Espeland 2009). One reason may be that monitoring signals the social importance of specific tasks or values to the monitor and other actors (Larson and Callahan 1990). Some researchers stress that monitoring is especially effective in impersonal settings where its “disciplining” function outweighs its tendency to undercut personal trust relationships (Frey 1993). When a monitoring regime is applied generally to like units rather than on an ad hoc basis it may gain acceptance if not legitimacy by undercutting claims that the monitors have singled out specific targets “unfairly” (Löwenheim 2008a). As described further below, there are good reasons to expect monitoring to influence both individual policymakers and organizational routines.

The concept of reactivity certainly has its counterpart in international politics. Oran Young, for example, stressed the prospect of being exposed as a norm violator over the risk of actual material sanction as the primary reason states comply with international norms (Young 1992, 176-177, see also Keck and Sikkink 1998a, Hawkins 2004). The theory seems to have some empirical basis. For example, Kelley and Simmons find that the act of being monitored in the United States Trafficking in Persons Report alone had a significant effect on states’ human trafficking policies (Kelley and Simmons 2014).

Comparative and Criteria-driven Judgment

Once established, monitoring systems constitute governing spaces over which monitors can wield considerable influence. A primary consequence of setting up a ranking or rating “system” is the de facto creation of a set of standards or expectations that define desirable qualities and acceptable behaviors. Criteria definition – making it clear what the monitor is seeking – has the potential to focus attention on what the evaluator is looking for, and, moreover, to set expectations regarding the value of those criteria. The creators of the Heritage Foundation’s Index of Economic Freedom, for example, created the index to frame an economic ideology that aligned with the emerging Washington Consensus. By designing the index the Foundation effectively also framed a debate. The index has garnered considerable attention and has been taken into consideration by other actors such as the World Bank and US policy makers. To monitor for specific qualities, policies or practices through criteria-based monitoring can therefore be thought of as a bid to define value and set standards. Actors know this, which may in part explain the proliferation of rankings systems within some issue areas.

that can be thought of as competitors. The rise of GPIs may in fact signal a struggle over valuation of appropriate state qualities and policies themselves. The rise of indicators of sustainable development to replace conventional measures of economic welfare is but one example of how creators of GPIs are working hard to shift values and judgments in preferred directions (Miller 2005).

Comparative judgments are the raw material for the production of systems of social control. Indeed, GPI creators strategically choose forms that emphasize their comparative nature. Indeed, as Figure 6 shows, nearly half of all GPIs are pure ranking systems, which are not only comparative, but also explicitly competitive. Another 29 percent use rankings combined with another form of assessment, so that nearly 80 percent of all indices use ranking. Such explicit comparisons create contexts in which judgments are formed and identities are established and reinforced (for a review see Ashmore, Deaux, and McLaughlin-Volpe 2004). The concept of “commensuration” in sociology, or “the comparison of different entities according to a common metric” denotes how objects or entities that are hard to compare can be simplified and judged according to a common metric. Sociologists characterize commensuration as a general social process by which people make sense of the world and judge others (Espeland and Stevens 1998a).

Ranking systems are among the most common of “judgment devices” by which value is attached to entities, objects or policies for which it is difficult or impossible to assign a market generated value (Karpik 2010). Such devices are heuristics that help to establish value when the market does not clearly do so. Ranking systems that involve quantification or categorization are especially useful in this regard (Hansen 2011, 508, Buthe 2012). When monitoring produces comparable numerical indicators or watch lists or blacklists, it can enhance traditional shaming by drawing a bright line under socially acceptable behavior. Comparable numerical indicators also stimulate competition (Marginson and Van der Wende 2007, Hazelkorn 2009, Ehrenberg 2003) and may indeed be specifically designed to do so (Schueth 2011).

![Figure 6: Index Form](Source: Authors’ database)
Quantification is not automatically broadly persuasive. Numerophobia is widespread, and some research suggests that people are often much more emotionally moved by vivid examples than by quantitative evidence. One well-known regularity in patterns of charitable giving, for example, is that people are much more likely to be generous to a single individual than large numbers of people in the same situation (Slovic 2007). While there seems to be a widespread assumption that numbers exude “authoritativeness” (March and Simon 1958) the evidence on the issue is actually quite nuanced.

Whether numbers are persuasive – that is, whether they are useful in changing peoples’ attitudes – has long been a subject of research about marketing and consumer behavior. This research does not support the idea that numbers are always automatically authoritative and persuasive; indeed, some researchers argue that the consensus is to the contrary (Kazoleas 1993). Rather, quantification is apparently persuasive only when it comes from a highly respected source (Yalch and Elmore-Yalch 1984); otherwise, reactions can range from suspicion to indifference. Sometimes people simply tune out when confronted with too much quantitative information. Moreover, much depends on what is being quantified and what other forms of information are available (Kadous, Koonce, and Towry 2005). Some studies find that quantification can significantly reinforce messages that depend primarily on ordinary language examples, and can change opinions to some extent when they contradict examples (Boster et al. 2000). One study found that negative statistical reviews of products were perceived as more credible than negative narrative reviews, but that quantitative evidence did not affect attitudes about the product any more than qualitative evidence (Hong and Park 2012). Interestingly, in marketing studies, however, some of the least numerate consumers seem to be most persuaded by quantitative evidence (Ju and Park 2013).

We do not expect to resolve these debates over the general effect of quantification on attitudes and human decision-making. We do note, however, that at least one widely cited meta-study firmly concluded that people generally find statistical evidence more persuasive than narrative evidence alone, although the study was limited to the United States and it did not test the proposition that their combination might be the most persuasive (Allen and Preiss 1997). There is little doubt that vividness of the message matters, however it appears that vivid statistical evidence impacts attitudes longer for a longer period of time than do vivid narratives (Baesler and Burgoon 1994). It is important to keep in mind that comparative indexes are often designed specifically by their creator-owners to attract attention, to be readily understood, and to influence public discourse. Many creators of indices have testified to this fact, stressing that they chose to create an index to attract attention, and noting that the media is particularly fond of indices.16 Numerical indicators of the kind we are discussing readily serve as “psychological rules of thumb.”17 They are promulgated precisely because they reduce complexity (Sinclair 2005, 52). A column of numbers can be scanned in seconds (and easily sorted in a

16 This theme was evident in a serious of 10 interviews conducted by the authors in Washington DC, August 12-14, 2014.
17 Indexes with greater “topical relevance” – that match the name of a document or search – are more likely to be found and noted, potentially making them more influential. See for example (Rieh 2002).
spreadsheet), while reading the underlying reports on which they are based (which may or may not be translated into the local language) could take weeks (Espeland and Stevens 1998b, 316, Löwenheim 2008b, 257-258). Numbers facilitate comparisons among units and over time. They can establish trends and can also be averaged, thereby helping to establish “norms” or “standards” against which it becomes straightforward to compare different units (Weisband 2000). Rankings and categories form focal points that define not only what it is we ought to pay attention to (Espeland and Sauder 2007, 16) but also who exactly does – and does not – make the grade. For these reasons, we think there are good reasons to expect actors to respond differently to GPIs than to words alone (Hansen and Mühlen-Schulte 2012, 457, Robson 1992).

Evidence suggests that a broad range of actors and entities respond to ranking systems. Applicants and universities respond to the famous U.S News and World Reports rating system (Monks and Ehrenberg 1999, Meredith 2004, Luca and Smith 2013); patients respond to hospital rankings (Pope 2009); investors even respond to firms’ environmental rankings (Murguia and Lence 2014, Aaron, McMillan, and Cline 2012). But do states respond to any of the GPIs that purport to rank everything from their infrastructure to their rule of law capacities? We hypothesize there may be some conditions under which they do.

IV. Mechanisms of Influence

The discussion above suggests that indicators may be especially persuasive forms of social pressure. But we need a theory about how they might influence state actors to take decisions implied by the GPI. We propose three general mechanisms that are likely to link these systems with state policy making (Figure 5): domestic politics, elite politics, and transnational pressure.

---

18 Murguia and Lence (2014) found that getting one position closer to the top of Newsweek’s “Global 100 Green Rankings” increases the value of an average firm in the list by eleven million dollars.
First, performance indicators can influence policymakers to the extent that they influence domestic politics. Higher rankings in a domestically salient policy area such as human rights or environmental protection can help to attract or retain domestic political support (Dai 2007). Salient, negative rankings potentially mobilize domestic political actors (NGOs, economic actors) who in turn press decision makers for behavioral or legislative change (Simmons 2009). In a study of the Extractive Industries Transparency Initiative (EITI), David-Barrett and Okamura argue that “EITI works because of the nature of the implementation process and the way in which it improves the capacity of civil society to hold governments to account, whether governments like it or not” (Dávid-Barrett and Okamura 2013, 3). This mechanism does not necessarily depend on external material power, although some groups may mobilize to protect an economic stake that could be threatened by an external sanction or by the fear of economic repercussions for business or investments. Mobilization can strengthen vocal domestic political coalitions who are inspired or incensed enough by the rating to demand official attention to the matter. Such demands can in some cases raise the costs of not responding for politicians.

Even the anticipation of publicity and negative domestic reactions could prompt preemptive policy review by government officials. Indeed, policy makers may be affected by publication information even if the public pays little attention, simply because policy makers anticipate and fear public backlash (Cook et al. 1983). This is also consistent with the idea that elites in general pay much more attention to media than the
mass public, because they believe that the media does, in fact, affect public opinion (Mutz 1998).

Direct Elite Response

Second, performance indicators can work directly through elite channels. GPIs sometimes target policies for which specific government officials are directly responsible, thereby affecting the personal status of an individual (e.g., government minister) or that of a collectivity such as a department or bureaucracy (Kelley 2013). Poor rankings are, simply put, embarrassing for an organization and its leaders. As discussed earlier, GPIs enhance traditional shaming through their comparative and heuristic attention getting nature and make allegations of poor behavior appear more objective. Because of this, when organizations’ managers get rated lower than they thing they deserve – referred to as “perceived identity-reputation discrepancy” (Martins 2005) – this can stimulate a search for ways to move up the scale by introducing policy changes before the next “grading period.” As with domestic politics, this mechanism can work independently of the material power of the rater; what is critical is the subjective regard of the rated for the rater and the need or desire to maintain a good professional reputation. Officials may initiate policy change to deflect criticism that could damage their personal or professional reputations.¹⁹

Sometimes monitoring and ranking can even influence ongoing bureaucratic operations and capacities. Monitoring may elicit compliance activity and stimulate information gathering. Many of the organizations that create indices have said that they dialogue with policy makers in the countries they are assessing both to get information, but also to give them opportunities to react to the information the index presents.²⁰ Publish What You Fund, the producer of the Aid Transparency Index, for example, engages in a six month back and forth before the index comes out and meet with the heads of donor agencies in different countries to discuss the index and the information in it.²¹ Such activity by external monitors may prompt bureaucrats to comb through records, assign employees data collection tasks, and forge connections with private actors who may have useful information. Some researchers have argued that the “collection, processing and dissemination of information” itself shapes the cognitive framework of policy-making (Bogdandy and Goldmann 2008, 242). More strategically, bureaucrats are adept at learning what it takes to improve their state’s ratings by consulting the bank of “approved” policy advice that monitoring summaries sometimes contain (Cialdini 2012) or by directly seeking the advice of the rating organization, a practice many index

¹⁹ This argument builds on a well-developed literature in international relations that argues state elites are susceptible to social pressures (Checkel 2001, Johnston 2001b). Shaming, or overtly singling out governments, and sometimes even individual leaders, for public opprobrium, is a tactic used by states, intergovernmental organizations (Joachim, Reinalda, and Verbeek 2008, Lebovic and Voeten 2006) and many non-governmental actors (Risse and Sikkink 1999, Hafner-Burton and Tsutsui 2005, Hendrix and Wong 2012, Ron, Ramos, and Rodgers 2005, Murdie and Davis 2012).

²⁰ This theme was evident in a serious of 10 interviews conducted by the authors in Washington DC, August 12-14, 2014.

producers have experienced. Teasing out whether monitoring primarily affects bureaucratic operating procedures or involves individual cognitive remapping (or both) is difficult, but both mechanisms suggest monitoring and GPIs are likely to influence both individuals and organizational routines.

Transnational Pressure

Third, indicators may impact policy by activating transnational pressure. Most notably, indicators may influence market expectations. For example, after Georgia jumped over 80 places in the *Ease of Doing Business* rankings by the World Bank, foreign direct investment tripled (Schueth 2011, 52). Even if the rater does not have direct control of material resources, indicators can influence policymakers in the target state if they contain market or other relevant information to which private economic agents respond. For example, Walt Disney blacklists countries based on their ranking in the World governance indicators. Credit rating agencies control minimal material resources of their own, but their ratings can touch off a tsunami in capital or exchange rate markets. Indeed, states may be concerned that ratings are linked; for example, credit rating agencies may be influenced by other indicators such as Transparency International’s *Corruption Perception Index* (Mellios and Paget-Blanc 2006). An indicator produced by one entity may also inspire third parties to apply additional pressure on a particular target. For example, the US uses many indicators in its assessment of whether countries qualify for Millennium Challenge Corporation (MCC) funds. Likewise, some indices are used as input into the World Bank’s *Ease of Doing Business* rating, causing countries be concerned about these input indices. A rater therefore need not have significant material power for the indicator to change the incentives of policymakers in a target country, although their assessments would need to have enough credibility to be taken seriously by the market or other actors.

Through each of these mechanisms, the performance indicator may have a “multiplier effect,” raising the target’s perceived risk that undesired behavior might have political, reputational or material consequences. These consequences likely vary by issue area. Money laundering black lists may work through transnational market pressures, but human trafficking ratings may work through the mobilization of domestic NGOs. A particular ministry or minister may pay special attention to indicators in his or her bailiwick; the World Wildlife Fund’s *Ecological Footprint Index* might particularly embarrass Ministers of the Environment, for example. Even when fairly tightly coupled with the material power of the rater, the added value of social pressure via indicators resides in their ability to signal community displeasure to the target and to stimulate a policy response. It is also important to recognize that positive ratings may stimulate

---

22 This theme was evident in a series of 10 interviews conducted by the authors in Washington DC, August 12-14, 2014.
24 This study does not document the effect of TI ratings on credit ratings, but uses the former as a predictor of the latter.
25 The MCC aids countries on the basis of 17 indicators generated by third parties, from IGOs (e.g., UNESCO) to NGOs (e.g., Freedom House) to universities (e.g., Columbia/Yale), and aids only those who score above the median. See [http://www.mcc.gov//pages/selection/indicators](http://www.mcc.gov//pages/selection/indicators).
“status maintenance,” or efforts to maintain good ratings. Either way, indicators can be used by powerful actors, to paraphrase Kofi Annan, to amplify the effect of their own moral, institutional, and material resources (Annan 1998, 129). Prima facie evidence of their influence can be found in their contestation (Hansen and Mühlen-Schulte 2012, 458). The fact that ratings are cited, discussed, and sometimes excoriated indicates their power to draw attention and to set the terms of the policy debate.

Distinguishing domestic, elite and transnational processes help us to organize our analyses of the influence of GPIs, but in reality there is considerable overlap and spillover across these mechanisms. The Country Indicators for the Foreign Policy Project, for example, has established a “watch list” for the possible eruption of violence across states, on the assumption that “The value of a watch list is that documenting such information and putting a country on a public list can itself bring about positive changes in behavior both by external and internal actors.” GPIs that excite broad publics, such as human rights or rule of law ratings, are likely to work at least in part through public reactions. GPIs that align with ministry responsibilities may work primarily through elite processes. GPIs that markets actors or aid donors care about are likely to work through transnational pressures. All of these mechanisms may also be influenced by the propagator nature of the GPI itself, propositions we consider in the following section.

V. When do GPIs Matter?

Some GPIs are widely recognized as influential while others linger in obscurity. When might we expect GPIs to influence state policies through any of the mechanisms discussed above? This likely depends on a wide range of factors. Some are implied by the discussion of mechanisms and likely depend on the characteristics of the target country; for example, domestic pressures are much more likely to be accommodated by democratic regimes than authoritarian ones. In this section, however, we theorize aspects of the GPI owner/propagator and the indicator itself (including the issue area) that may affect GPIs’ capacity to exert social pressure on states.

Attributes of the owner

GPIs do not exert influence in a vacuum; they leverage power. Part of that power flows from the status of the “creator” or “owner” who implements the regime and exercises control over it. GPIs are tools of influence exercised by the former vis-à-vis the latter. Because of this, the characteristics of the GPI creator matters. Creators may have material resources they can wield to “encourage” the rated to comply with desired policies. It usually takes material resources to produce and promulgate believable GPIs in the first place, especially if these involve collecting and creating new information. Moreover, at least a few GPIs were created specifically to legally justify material sanctions (the United States Trafficking in Persons tiers) or explicitly to set criteria for development assistance (the 15 indicators used to dole out aid through the MCC). Such implicit or explicit links to material resources likely enhance the effectiveness of the GPI.

26 See for example the Country Indicators for Foreign Policy Project, p. 11 at: http://www4.carleton.ca/cifp/docs/archive/watchlistreport1.pdf.
That said, status in international relations is not isomorphic with material power (Wohlforth 2009). Material leverage can be direct and explicit (the threat of sanctions) but more often it will be indirect and implicit. Rated states may be concerned about a range of material consequences of their behavior, from decreased resource allocation to reduced trading opportunities. As noted above, even when the rater does not possess significant material resources, GPIs can indirectly leverage the resources of third parties if they inform investment or resource allocation choices of others (Brooks, Cunha, and Mosley 2013).

More generally, the status or authority of a GPI creator boosts the GPI’s perceived legitimacy and increases attention. At the most basic level, people find information more credible when it comes from what they believe is a “trustworthy,” “credible,” or “authoritative” source (Wilson 1983). Trust of the source is then used to filter all kinds of information including (or, especially) information posted on websites (Rieh 2002). More profoundly, authoritativeness is deeply intertwined with an actor’s legitimacy, which bolsters their ability to set the policy agenda and to initiate shaming. Nye speaks of this type of influence as “soft power.” An effective GPI can over time create a positive feedback loop that bestows further authority on the creator or permits the creator to leverage the material resources of other actors. Among state actors, international organizations are often thought of as respected actors (Galtung 1973, Bearce and Bondanella 2007, Johnston 2001a). Hall and Biersteker (2002, 4) note that many new non-state actors are also perceived by states as authoritative, for purposes of ‘certifying’ norm fulfillment, for example.

An actor’s network position can also contribute to its ability to wield an influential GPI. First, actors central to their network are often best positioned to set agendas as discussed above. As Carpenter has shown, “An issue’s salience within an advocacy network will depend on its adoption by organizations most central to that network” (Carpenter 2011). Second, network position defines social relationships central to shaming and reputation building. “Social impact theory” emphasizes the importance to the target of the actor or group of actors engaging in pressure, the nature and extent of the target’s exposure to the group, and, to some extent, the size of the group attempting to enforce conformity (Latané 1981). Third, network position can also impact information flows, facilitating both data collection and GPI dissemination. For example, the United States Department of State, which creates the Trafficking in Person Report and grades country performance, uses its embassies and NGOS around the world to tap into many sources of information. As Diane Stone notes, “A [policy-relevant knowledge] network amplifies and disseminates ideas, research and information to an extent that could not be achieved by individuals or institutions alone” (Stone 2002). Actors who are located where they have access to decision makers are better able to mine information and assert authority (Hall and Biersteker 2002, 14). Overall, well-connected actors have the potential to use GPIs in ways that exploit their network positions (Kahler 2009, Hafner-Burton, Kahler, and Montgomery 2009).

GPI creators also vary in their reputational resources, which can boost their authority. Actors may gain reputational resources based on their track record and expertise (Hall and Biersteker 2002). Well-respected raters such as Consumer Reports have a tremendous impact on consumer behavior, even when they contradict previously
held beliefs or occasionally provide erroneous information (Simonsohn 2011). There is a long tradition of research on the influence of epistemic communities on states’ policies (Haas 1989, Haas 1992), and we would expect many of the findings of that literature to be relevant to discussions of GPIs. For example, were Amnesty International to produce an index of human rights (which it does not), this would command some authority because of its longstanding record as a non-political research-based organization with extensive knowledge about human rights violations. Similarly, a World Competitiveness Index created by the Swiss will carry different weight than a similar index were it to be developed by Russians. This is not because the Swiss have tremendous coercive power; rather they have much higher credibility for the purposes of such a rating. Actors who themselves are thought to uphold certain norms are more likely to have normative power (Sikkink 1993), which Manners defines as the ability to shape conceptions of ‘normal’ (Manners 2002, 236). While people tend to discount data and websites they link with various interest groups, they tend to trust what they consider to be objective sources of information (Flanagin and Metzger 2007). In sum, we hypothesize that:

**H1:** GPIs are likely to attract more attention and be more effective when their creators possess greater authority, either because, a) their creators have recognized expertise, experience or normative power in an area, or b) their creators have central network positions.

**H2:** GPIs are likely to attract more attention and be more effective when their creators directly control material resources and can leverage these over the rated.

**Attributes of the indicators**

Are all GPIs created equal? What is it about the nature of the indicator itself – or the issue area it purports to address – that might vary and might influence policy? While we believe that the nature of social relationships between rated and rater is likely more important than variance in the GPIs per se, there may be good reasons to think through the rating processes and formats that could potentially influence state outcomes. Here we hypothesize about six factors: the potential to shame, the potential to leverage third party resources, the engagement of the rated, the actionability, the credibility of the methodology and the ecology of the issue areas.

Because shaming is a primary way for GPIs to affect policy makers, all else equal, GPIs more able to engender shame ought to be more effective. Poor performance in areas with greater moral appeal may be more shameful, or have greater ability to produce moral outrage and mobilize domestic publics. Margaret Keck and Kathryn Sikkink have argued that issues involving bodily harm or issue involving equality have higher moral

---

27 When Consumer Reports provided wrong information that was uncorrelated with existing beliefs people acted on it. For every position lost/gained in the safety ranking CR provided, the average car seat price dropped/increased 3%. This illustrates that people respond to what they consider to information provided by “expert” sources even when it is wrong and even when it contradicts previously available information. See Simonsohn (2011).

28 A Swiss NGO, the International Institute for Management Development, has created and disseminated such an indicator. See http://www.imd.org/news/World-Competitiveness-2013.cfm
appeal and generate more shame (1998b, 26-28, see also Hawkins 2002). GPIs that use bright lines that place some actor’s practices or policies firmly in an “unacceptable” category, like watch lists or black lists, may also be particularly effective at producing shame. Kelley and Simmons (2014) have found for example that when the United States placed states on the human trafficking “watch list” this has a significant impact on their likelihood of criminalizing human trafficking – a policy at the top of the U.S. priority list.

**H3:** Indicators are more likely to draw attention and be effective when they produce more shame, either by a) addressing issues that inherently considered more morally shameful or b) by drawing bright lines under unacceptable behaviors by using watch lists or blacklists.

As noted in the discussion of transnational pressure, leverage over material resources likely magnifies the impact of indicators. Some issue areas have greater ability to influence transnational third party resources that bear on trade, investment or aid flows. These include governments’ budgets, transparency, rule of law, economic policies and the like; issues that influence the market environment.

**H4:** Indicators are more likely to draw attention and be effective when they concern issues that can leverage third party resources.

Even if they do not explicitly or implicitly shame or leverage resources, the process of creating and promulgating GPIs can be an opportunity to socialize states to existing or emerging global standards. This is especially likely if raters engage the rated. Involving bureaucracies in information gathering and interpretation provides opportunities for talking about, measuring, documenting and justifying behavior – all processes critical to socialization (Risse 2000, 6). The interaction also provides the creators of the index an opportunity to suggest solutions and approaches to the issue at hand. Other times GPIs can stimulate discussion between the rated about the policy steps they have taken to improve their rankings, thus creating, in the words of Egypt’s foreign minister in 2008 “a forum for exchanging knowledge.”29 While much literature has focused on the socializing role of international organizations (Voeten 2014, Johnston 2001b), it is at least plausible that actively participating in the latest GPI wave – as opposed to passively allowing that wave to wash over – could contribute to the socialization process. So what about a GPI encourages such interaction?

GPIs based on original data have the potential to stimulate more interaction. Some indicators merely repackage other data at relatively low cost to the producer. Producing original data is far more engaging. If the policy makers in a country are requested to fill out surveys, or provide information on a particular topic regularly, this topic will be primed in their minds, and could have the potential to spawn policy action (Weaver 1991, Scheufele 2000). If governments have to gather information on a particular issue, they are

29 Dr. Mahmoud Mohieldin, quoted in World Bank and International Finance Corporation (IFC). 2004. Doing business in 2004: Understanding regulation. Washington, D.C.: World Bank, p iv: “What I like about Doing Business […] is that it creates a forum for exchanging knowledge. It’s no exaggeration when I say I checked the top 10 in every indicator and we just asked them, “What did you do?” If there is any advantage to starting late in anything, it’s that you can learn from others.”
also more likely to institutionalize these processes into bureaucratic structures, with the result that some policy makers and bureaucrats gain jurisdiction and thus ownership. Original GPIs are also more likely to create new information, creating new pressures on state policies.

“Voluntary” participation may matter as well. Even if they are not keen on being rated, states may consider it in their interest to cooperate, especially if not doing so implies serious shortcomings. Research on election monitoring for example has revealed social pressure to “cooperate” since not doing so appears as a self-declaration of cheating (Kelley 2012, Hyde 2011). Some GPIs such as the Aid Transparency Index are explicitly founded on cooperation among states that agree to be monitored according to some specified standards.

Finally, engaging the rated is akin to providing a “grading rubric.” When one knows what it is they are expected to do, they tend to do it. Educators are keenly aware of the power of a grading rubric to influence the quality of student work (Howell 2013). Perhaps states supplied with a “rubric” will be more engaged in policy development.

H5: GPIs are more likely to attract attention and be effective if they solicit engagement either by a) creating original data that fosters interaction with the rated, or by b) soliciting voluntary cooperation by states.

Indicators also vary greatly in their actionability which might also impact their effectiveness. The World Bank notes that actionability “implies greater clarity regarding the steps governments can take to improve their scores on an indicator, i.e. if the government successfully undertakes reforms in certain areas, relevant indicator(s) will respond in a favorable direction” (World Bank year unknown). Some indicators focus on a relatively narrow issue areas that policy makers can address through concrete steps, while other are widely encompassing and offer little short term, or in some cases even long term, control. The topics of some indices are inherently more governable (Kooiman 2008); the government can adopt policies to address the problem if it has fairly tight control over an issue area.

Several examples clarify the point. The United Nations Public Administration Program (UNPAN) E-Government Development Index monitors and measures progress of a country’s citizens to access public information electronically, which is relatively responsive to government programs and policies.30 Similarly, it is entirely within the power of a government to decide whether to publish its budgets, as required by the Open Budget Index.31 Compare these to the Basic Capabilities Index created by Social Watch, an NGO dedicated to poverty eradication. This index purports to measure “structural dimensions that represent the indispensable starting conditions to guarantee an adequate quality of life.”32 Very broad indicators may suffer from collective action problems: no one person can easily be held responsible for conditions as broad as “the quality of life.” By contrast, an education minister may feel particularly responsible for a low score on the Trends in International Mathematics and Science Study (TIMSS).33 The politics of

31 See http://survey.internationalbudget.org/.
33 See http://nces.ed.gov/Timss/
responsibility, and thus shame, are triggered when an indicator has a specific counterpart in a state’s decision-making structure. As Keck and Sikkink have noted, “[P]roblems whose cause can be assigned to the deliberate (intentional) actions of identifiable individuals are amenable to advocacy network strategies in ways that problems whose causes are irredeemably structural are not” (Keck and Sikkink 1998b, 27). Thus, we hypothesize that:

\[ H6: \text{GPIs are more likely to attract attention and be effective if they focus on issues that are more actionable for the government and for which responsibility can more easily be assigned.} \]

Since indicators are influential to the extent that they create an impression of accuracy, it is important to note that some exude technical credibility, while others do not. Raters and rankers have an incentive to convince audiences that their indicator reflects some essential truth, and that it was developed carefully and professionally. Some performance indicators specify the methodology by which they determine their scores, while others are quite opaque. Some groups have tried to improve the “scientific” nature of their product over time. One good example of this is Freedom House, who has been in the “freedom”-rating business since Ray Gastil sat down at his desk in 1973 and rated freedom in ways that were completely opaque to the outside world (Gastil 1990). This changed over time as Freedom House’s indicators gained wide usage in policy and academic circles. By the mid-1990s, the organization used a team of experts and regional specialists to come up with their scores. While the rating scheme is still not completely clear, now Freedom House describes its sub-indexes, information sources and the expertise of the raters involved at reaching their final scores.

Professionalism in the creation of indicators implies internal consistency and transparency in methodology. Do they use existing data, or do they gather data themselves, and if the latter, how is it collected and validated? Are the sources “expert”, or are they likely quite political? Are the data self-reported, and do checks assure quality and accuracy? How are various components of composite indexes weighted, and how are the weights justified?

Many organizations discuss the methodology for their performance indicators on their websites. The group Vision of Humanity, which touts the Global Peace Index, for example, describes in clear, simple terms the components and weights of its index, and assures readers that their “22 qualitative and quantitative indicators…[]…were selected by an international expert panel and are reviewed annually.” However, there is no explicit discussion of sources or explanation for how the experts map qualitative data onto their 1-5 point scale.\[34\] Rating organizations have to decide how much to invest in the “scientific” nature of their enterprise and how much to divulge to an audience who may (or may not) bother to dig that deeply. Some organizations publish methodological explanations that explore the robustness of their measures (see for example the OECD’s Indicator of Employment Protection\[35\]) or even allow an interested user to reweight the

\[34\] See http://www.visionofhumanity.org/#/page/about-gpi.

components to view results under a range of different assumptions (see for example the *OECD Better Life Index*).  

A number of performance indicators rely on “expert surveys” for their raw data. The trend seems to be to crowd-source the experts to get answers to very subjective assessments. The World Justice Project’s *Rule of Law Indicator*, Transparency International’s *Bribe Payer’s Index*, and the World Bank’s *Ease of Doing Business Index* are all based on expert surveys. Some even offer their individual level data, should researchers (or skeptical governments) wish to re-do their analysis. Some GPIs also use external reviewers to boost their credibility. For example, the *Aid Transparency Index* publishes the names of such reviewers in the report and the creators note, “The credibility of these experts is important to the credibility of our methodology, which is hugely important to the credibility of the index.”

In general, we would expect indicators that justify their methodology to be more credible than those that do not.

**H7:** GPIs are more likely to attract attention and be effective they are based on a publicly available methodology that appears rigorous and well justified.

Finally, it may be important to situate specific measures in their GPI ecological context. Some issue areas are crowded with GPIs; others much less so. For example, the tier ratings in the United States *Trafficking in Persons Report* get a lot of attention partly because it dominates the issue area; it has no serious competitors that measure the same or a similar policy area. These US ratings can be contrasted with the *Human Development Index*, which overlaps significantly with many other GPIs (“happiness” indicators such as the *Happy Planet Index* and the UN *World Happiness Index*; Freedom House *Freedom in the World* democracy scores; the *Mo Ibrahim Index* and the *Legatum Prosperity Index*, to name a few). A GPI can be thought of as more *focal* to the extent it captures attention in a given issue area. The more crowded the field the harder it is for any given measure to dominate. Crowded fields are likely to reduce the ability of any specific measure to focus policy discussions and public discourse. If a measure has overwhelming market share and few competing GPIs that might reflect or contribute to its authoritativeness, but it could also have to do with some production or marketing advantage of its producer. If “counter indices”– GPIs created specifically to balance the

---

40 This is the case for example with the *Ease of Doing Business Index* created by the World Bank. [http://www.doingbusiness.org/methodology](http://www.doingbusiness.org/methodology).
42 The UN publishes a less comprehensive report only every three years or so, and the *Global Slavery Index* is brand new.
perspective of the dominant index – develop within issue areas this could be a sign of the dominant GPI’s authority, as well as of political contestation.

H8: GPIs are more likely to attract attention and be effective when there are fewer competing indices in the given issue area.

VI. Conclusion

The global information age invites us to think about new forms of influence and power. This paper has focused in particular on broadening our concepts of social power in international relations. We suggest it is useful to move beyond traditional exercises of power such as military and economic coercion, and to consider the ways in which the information environment influences states to adopt particular policy measures and behaviors.

Social pressure has long been understood as important exercise of power in international relations, but much of the focus has been on various forms of explicit uses of shaming, exclusion, and social rewards. But how do actors come to define, understand, and measure state policies in ways that gain attention and social acceptance? How is such “knowledge” propagated? And how do we come to believe that certain states “measure up” and others do not? GPIs may be part of the contemporary answer to these questions. We have begun to map the empirics of GPIs, and suggest that they may influence state policy decisions through a number of mechanisms, such as popular politics and mobilization, elite politics and bureaucratic operations, and transnational processes involving pressures from markets, aid donors and traditional transnational advocacy. The contribution of this agenda is to investigate what GPIs add to these traditional political mechanisms.

We have offered some ways forward in this regard. A proliferation of indicators implies enhanced state monitoring, which in and of itself may nudge policymakers towards “accepted” policy norms. It is important to consider whether some indicators are more influential than others, and if so, why this might be the case. We have theorized how creator and index characteristics influence the ability of an index to attract attention and possibly change state behavior and developed specific hypotheses about each of these. At this stage, though, most of these remain empirical questions. Testing these across a set of indicators will not be feasible due to lack of comparable outcome measures. However, for many of these indices, it is possible to develop measures of success and to examine their possible effectiveness through case studies of even with quantitative methods as has already been done in some cases (Kelley and Simmons 2014).

More research is surely needed, not only on the influence of indices but also about who wields such influence. There is no guarantee that the indicators that have proliferated represent a universally accepted or even a coherent agenda; indeed, many may instead represent the political agenda of the few. There is no systematic check on their accuracy. And there is no guarantee that they will be fully probed or even understood by those who consume them. That said, our data also suggests that many GPIs are reinforcing global power systems, with the great majority emanating from rich western countries, although this of course does not preclude a shift away from state power towards IGOs and private
actors given that fewer than five percent of GPIs are actually created solely by states. Some indices, such as the *Aid Transparency Index*, are pushing rich and powerful governments to be more accountable. On the other hand, some hold considerable risks that the rated try to game the system by targeting the sub-indicators in ways that fit only narrowly with the purpose of the index (Schueth 2011).

The questions about norms and power are of course of utmost importance. The agenda we have laid forth, however, is to actually theorize and study whether and how GPIs may influence state behavior. This is crucial, because knowing whether and how such GPIs matter is essential to be able to answer the prior questions about whether they actually influence norms and standards and alter global power relationships. Only once we understand more about whether and how they influence politics we will be in a position to ask, *has this been for the better?*

---

Figure 4: GPI "Popularity": Bing Hits by GPI Name

Note:
Search 20 July 2014
Excludes 20 indexes with fewer than 1000 hits.
Excludes S&P credit ratings (1.7 billion hits)
References:


Clark, Pilita. 2013. "OECD to add climate scorecard to its country surveys." Financial Times, October 9, 2013.


33


World Bank. year unknown. The World Bank's AGI initiative.
