March, 2012

Judgment and Measurement in Political Science

Andreas Schedler

Judgment and Measurement in Political Science

Andreas Schedler

Standard methodological advice in political science warns against the distortion of measurement decisions by judgmental elements. Judgment is subjective, common wisdom asserts, it produces opaque, biased, and unreliable data. This article, by contrast, argues that judgment is a critical intersubjective ingredient of political measurement that needs to be acknowledged and rationalized, rather than exorcised.

Following standard methodological guidelines in the social sciences, we commonly demand political measurement be free from judgmental elements. Social measurement is conventionally defined as the assignment of numbers to observations according to rules. To obtain objective, rather than subjective, data, we demand, we must renounce our judgmental faculties. We must assign our numbers on the basis of observations and rules, and nothing else but observations and rules. While the introduction of judgment produces opaque and unreliable data, we claim, a disciplined reliance on observable facts and precise rules allows us to generate transparent and reliable measures.

The reliance on rules and observations is a constitutive element of social scientific measurement. The *exclusive* reliance on rules and observations, however, is a misleading and self-deceptive ideal. It is neither a good description of what we actually do, nor a good prescription of what we ideally should and practically can do. It fails to recognize the actual role judgment plays in political measurement, as well as the potential role it may play (subject to appropriate standards). Banning judgment from measurement is neither a feasible methodological imperative nor a desirable one. It places too much faith in observation and regulation, and too little in public reasoning among scholars.

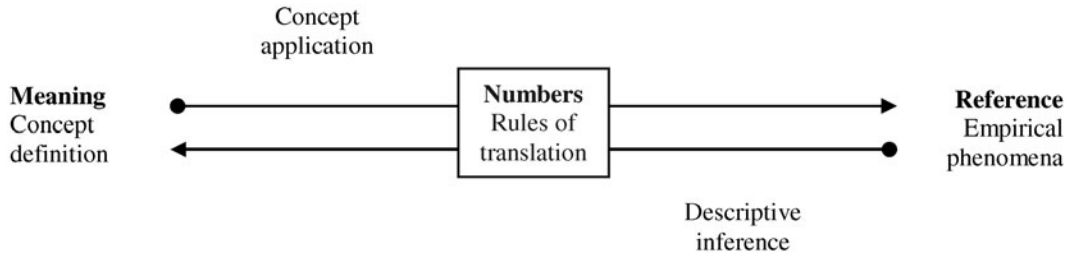
In this article, I propose a balancing act. On the one hand, I recognize the essential role rules and facts play in political measurement—yet I contend that they are seldom clear enough for us to dispense with judgmental elements. On the other hand, I recognize the essential role judgment plays, above all, in the measurement of complex concepts—yet I hold that judgment must conform to solid procedural standards to fulfil its promise of intersubjective rationality. My attempt at methodological diplomacy contributes to larger ecumenical conversations in the discipline that strive to bridge sterile methodological oppositions, especially between quantitative and qualitative methods.[1] While (diplomatically) eschewing philosophical debates,[2] my approach is anchored in an understanding of measurement as an act of *translation* between concepts and realities in the language of numbers. For measurement to be scientific, it must be grounded in shared concepts, shared realities, and shared rules of translation—all of which require judgment as well as the regulation of judgment.

While I hold my basic argument to be generally applicable to measurement processes in political science, I illustrate the twin necessity of both employing and disciplining judgment mostly with examples from the comparative study of political regimes. Perhaps even more than other subfields, contemporary research on democracy and authoritarianism has been driven by an expanding pool of cross-national data demanded by researchers, much of which, despite the discipline-wide wariness about judgmental measurement, has been more or less dependent on the judgment of knowledgeable experts.[3] Given the large units of

Andreas Schedler is professor of political science at the Centro de Investigación y Docencia Económicas (CIDE) in Mexico City and vice-chair of the IPSA Research Committee on Concepts and Methods (andreas.schedler@cide.edu). His recent publications include "The Measurer's Dilemma: Coordination Failures in Cross-National Political Data Collection," Comparative Political Studies *(February 2012).* *He thanks Alejandro Anaya, Carlos Bravo, Ronald Francisco, Jennifer Gandhi, Carolina Garriga, James Melton, Covadonga Meseguer, Alvaro Morcillo, Cas Mudde, Mauricio Rivera, and Gilles Serra, who have provided valuable comments on previous versions of the manuscript. He is also grateful to the anonymous reviewers and the Editor-in-Chief of* Perspectives on Politics *for their incisive criticism and encouragement. The usual caveats apply. All judgmental errors remain the author's.*

doi:10.1017/S1537592711004889

Figure 1
Social Measurement



analysis that regime studies focus on, their ambitions of generalizing across time and space, and the relative opacity of their objects of study, it can hardly be otherwise. This important field of political inquiry thus serves well to illustrate both the methodological necessities of relying on judgment and the methodological pitfalls of doing so in the absence of clear procedural standards.

To develop my argument on the role of judgment in political measurement first of all I delve into the logic of political measurement. I conceptualize measurement processes as acts of translation between facts and concepts in the language of numbers, and judgment as the capacity to perform such translations in an intersubjective manner. To obtain impersonal measures, common methodological wisdom instructs us to minimize judgmental elements. Instead of employing our judgmental faculties, we are to rely on the mere observation of facts and the strict observance of rules. I describe this quasi-bureaucratic ideal as a regulatory idea whose full realization is neither feasible nor desirable. In continuation, I delineate structural limits to the full "bureaucratization" of political measurement: limits to pure observation and limits to complete regulation of political data development. Both compel us to use our judgment. Finally, I turn my attention to the methodological imperative of rationalizing the use of expert judgment in political measurement. I trace the need for raising our methodological standards in five aspects of judgmental measurement: expert selection, data comparability, transparency, divergence, and accountability. I conclude by laying out disciplinary implications of exorcising, rather than recognizing, the role of judgment in political measurement.

The Nature of Measurement

What does social measurement involve? What does it require? To formulate clear and consistent methodological demands upon measurement we need, first of all, to be clear about the logical demands that measurement places upon us.

Measures are bridges. Measurement processes build numerical bridges between abstract concepts and empirical realities. They relate the meaning of concepts (their intension) with their reference (or extension) through the language of numbers. Social measurement is a linkage operation, an act of translation. Viewed from the side of meaning, it is a form of concept application: it translates concepts into numbers and attaches these to empirical phenomena. Viewed from the side of reference, it is a form of descriptive inference: it translates empirical observations into numbers and connects these to general concepts. To measure something therefore is always a relational exercise. It requires a triple operation. It requires us to clarify our ideas, to capture empirical realities, and to link the two through numbers. Figure 1 illustrates this mediated bidirectional relationship.[4]

Concepts are not observable. Scholars routinely distinguish between "latent" and "manifest" concepts, that is, between "abstract, theoretical, and unobservable concepts" on the one hand and concrete, empirical "concepts for which there are direct observations" on the other.[5] This common distinction suggests that we can observe some concepts but not others. Yet this is an illusion, for concepts are never observable. They may be more or less abstract and their empirical referents may be more or less easily observable. But concrete concepts are not more observable than abstract ones, and observable referents do not render concepts themselves observable. The concept of a dog is more concrete than the idea of a living soul, and dogs are easier to watch. Yet we observe dogs running and barking and fooling around, not the concept of dogs. The bridge we build through acts of measurement between concepts and observations may be longer or shorter, more or less solid. Yet a bridge it remains. As a matter of fact, in the social sciences, we habitually strive to engineer ambitious bridges between abstract concepts and unobservable phenomena, rather than modest ones between concrete concepts and observable phenomena.[6]

Quantification requires classification. In the social sciences, the notion of measurement commonly includes both the classification of objects, such as legislatures and political parties, and the scoring of object attributes, such as

legislative strength and the ideological position of political parties. The latter logically presupposes the former. Quantification presupposes classification. If we want to measure an attribute of some object, we can't just grasp any object, but need a suitable one.[7] For example, if we want to determine the height of things, we need to identify appropriate objects of measurement: physical objects. It would make little sense if we took out our measuring rod and strove to determine the height of free speech or the worldwide popularity of Nelson Mandela. Similarly, if we want to establish the democratic nature of political regimes, regardless of whether we conceive democracy in continuous or dichotomous terms, we first need to identify political regimes. States of political disorder, for instance, won't qualify.[8] Determining the presence of empirical phenomena (the classification of cases) and weighting their characteristics (the quantification of attributes) are distinct operations. Still, in this article, I will not distinguish systematically between them, because they pose very similar methodological challenges. This is also true for *counting,* the determination of frequencies of empirical phenomena that classify as members of the same conceptual category.

Measurement requires rules. In the world of physical objects and standardized units of measurement, measurement is traditionally understood as "the practice of attempting to identify the magnitude of a quantitative attribute by estimating the ratio between that magnitude and an appropriate unit."[9] In the social sciences, we lack standardized units of measurement for most purposes of quantification. Commonly, we cannot even tell how such units might be conceived in the first place. How much is an ounce of power, a gallon of legitimacy, a pound of deliberation? Lacking standardized units and instruments of measurement, we have to craft formal rules of measurement. It is only under the guidance of explicit rules that we can translate concepts and observations into numbers "in meaningful ways"[10]—in ways that others can understand, criticize, and replicate. In the social sciences, to measure is to legislate. Rules are constitutive for our conceptions of measurement, classically defined by psychologist Stanley S. Stevens as "the assignment of numerals to objects or events *according to rules.*"[11]

Measurement requires more than observation. If social measurement is the meaningful assignment of numbers to observations according to rules, the methodological admonition to base measurement decisions on "observations, rather than judgments"[12] is *incomplete.* Measurement requires the ability to observe, which is a sensorial faculty, but also the ability to comprehend meaning and follow rules, which are symbolic competences.[13]

If measurement is a process of translation between ideas, facts and numbers, to what extent can we cleanse its constitutive components—the comprehension of concepts, the apprehension of realities, and a command of the language of numbers—from judgmental elements? And to what extent should we do so? The methodological ambition of eradicating judgment commonly rests upon a conceptual misunderstanding: the equation of judgment with unbounded subjectivity. Revaluating judgment, first of all, requires severing this definitional link.

The Nature of Judgment

Judgment has a bad name, in particular among quantitative methodologists who tend to describe (and disqualify) it as "subjective"[14] and to issue urgent calls for "bringing objectivity back in."[15] Equating judgmental with subjective measures rests upon a twin confusion, though. It confuses the nature of both subjective and judgmental data.

Subjectivity. The common notion of subjective measures conflates two possibilities: subjective objects and subjective procedures of measurement. In terms of measurement *objects,* subjective or self-reported measures are meant to provide a glimpse into the inner world of subjectivity. They strive to capture beliefs, values, and desires of individuals. Subjective measures in this sense generally are not deemed methodologically problematic per se. As a matter of fact, whole fields of empirical research, such as the study of public opinion, rely on subjective measures. Their reliability simply depends on the controlled nature (context-independence) of the accompanying procedures and the representative nature of the underlying sample. For decades now, survey researchers have been refining their sampling techniques, survey design, and interviewing procedures to assure the impersonal collection of subjective data.

In terms of measurement *procedures,* I understand the notion of subjective measures to describe substandard measurement practices that fail to explicate their concepts, empirical referents, or rules of numerical translation, and thus generate idiosyncratic measures of uncertain meaning. Rather than representing recognizable phenomena in a recognizable manner, they "somehow" reflect the private knowledge and personal criteria of individual scholars—although we cannot know how, as long as underlying concepts, facts, and rules of translation remain unclear. When measurement procedures are subjective, different measurers produce different measures, without providing criteria of how to reconcile their differences. For instance, some cross-national expert-based estimations of the rule of law show low and even negative (!) correlations among each other. Such incongruent patterns of measurement are likely to reflect unacknowledged subjective biases as well as divergent definitions.[16]

Judgment. Expert judgments are often dismissed as "subjective" data. Yet, in contrast to subjective measures, they

Table 1
The Regulation of Measurement

	Concept meaning		Numerical translation		Empirical reference
I. Classification of cases					
Rule design	1. Case definitions	→	2. Classification rules	→	3. Observation rules
					↓
Rule application	3. Case identification	←	2. Classification decisions	←	1. Empirical observations
II. Quantification of attributes					
Rule design	1. Attribute definitions	→	2. Coding rules	→	3. Observation rules
					↓
Rule application	3. Attribute scores	←	2. Coding decisions	←	1. Empirical observations

are not supposed to be subjective, but *intersubjective*: grounded in public facts and public reasons, defensible in the face of critique. Judgments are not mere opinions, personal beliefs, or private epistemic preferences, arbitrary and unbound by claims of truth. Judgment is a child of reason. It is the "critical faculty" of forming an independent opinion in the face of conflicting arguments. A person of "good judgment" is a "competent critic." Exercising "sound judgment" requires "discernment, discretion, wisdom, understanding, good sense."[17] Immanuel Kant defined judgment as "the faculty of thinking the particular as contained under the universal."[18] In this wide sense, judgmental faculties include the capacity to understand and discern everything we need to understand and discern in social measurement: concepts, numbers, rules, and social realities.

The Ideal of Bureaucratic Measurement

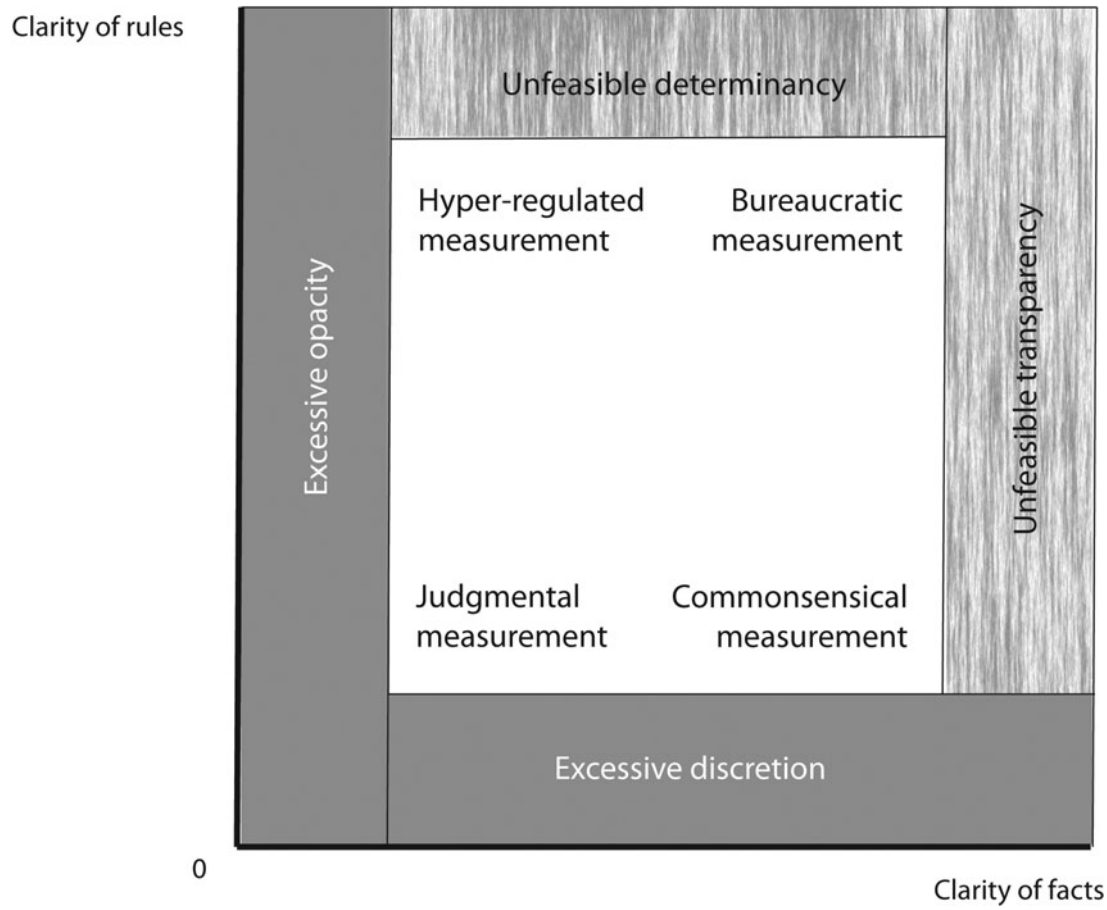
Modern science is a social enterprise of impersonal argumentation. The validity of scientific propositions about the world, be it the physical or the social world, do not depend on the personal identity of the speaker. Claims of privileged private access to knowledge do not constitute valid bases of scientific argument. Only public reasons, open to the interplay of justification and criticism, do. Just like causal arguments, descriptive arguments are supposed to be impersonal. In physical and social measurement, the numbers we assign to objects are intended to reflect properties of these objects. They are not to be determined by the private beliefs of the observer, or by the vicissitudes of the measurement process. Accurate measurement establishes a relation of correspondence between the symbolic realm of arithmetic numbers and the empirical realm of objective attributes. It requires validity, the control of systematic measurement error, as well as reliability, the control of random measurement error. Both demand that the numbers we assign to empirical phenomena do not flow out of chance or divine inspiration or personal intuition. They demand public justification and transparency at all three stages of the measurement process:

1. *Concepts must be transparent:* Concepts are the substantive anchors of the measurement process. If our concepts are unclear, contradictory, or shifting, our measures will lack clear, consistent, and stable meaning.
2. *Facts must be transparent:* Empirical realities are the factual pillars of measurement. If our empirical referents are unclear, contradictory, or shifting, our measures will lack clear, consistent, and stable empirical content.
3. *Translations must be transparent:* The translation of concepts and observations into numerical expressions must follow explicit, consistent, and stable practices. Otherwise the mediating role numbers play between concepts and facts will become unclear, inconsistent, and volatile.

In social research, the most common means to ensure impersonal measurement has been *regulation*: the formal definition of concepts, the introduction of explicit rules of observation, and the formal definition of rules of translation that govern the assignment of numbers to facts and ideas. To produce more and more precise measures of less and less visible phenomena (great and small alike), natural scientists have been building larger and larger machines. Leading research institutions in both nanophysics and astrophysics have been building gigantic measurement machines, such as particle accelerators and "extremely large" astronomical telescopes. In the social sciences, we don't build machines. We knit rules. What natural scientists try to accomplish by constructing sophisticated machines, we strive to accomplish by constructing (more or less) sophisticated rules: impersonal measures—valid, reliable, and precise—of the empirical phenomena we are interested in.[19]

In social measurement, as elsewhere, formal regulation is a two-step process. It involves a "legislative" phase of rule design and an "administrative" phase of rule application. Table 1 illustrates this process. Although the table distinguishes between case classification and attribute scoring, the underlying logic is identical: By subjecting it to formal

Figure 2
Types of Measurement



rules, measurement is expected to work in a rational, predictable, impersonal manner. Formal rules are designed to eliminate discretionary decision-making at each step of the measurement process, from concept definition (meaning) to observation (reference) and the assignment of numbers (translation). Formal regulation is meant to transform decisions at all three stages of measurement into mechanical applications of formal rules, rather than discretionary practices.

The methodological ideal of impersonal, rule-based, non-judgmental measurement can be described as "bureaucratic" or "legalistic" insofar as it is based on a similar ideal conception of "mechanical" decision-making as we find it in modern administrative and judicial systems: Decision-makers do not properly "decide," but draw practical conclusions from available facts and applicable rules. Their role is supposed to be inferential, not judgmental. Just as the ideal bureaucracy involves a perfect separation between politics and administration, the ideal measurement involves a perfect separation between judgment and observation. According to *the idealized conception of bureaucracy,* politicians define the rules, while lower-level public officials only apply them in a mechanical, impartial, and impersonal manner. On this conception politics constitutes the realm of discretionary decision making, administration the realm of non-discretionary rule application. According to *the idealized conception of social measurement,* empirical researchers define the rules, while research assistants (measurers) only apply them in a mechanical, impartial, and impersonal manner. On this conception theory building and concept formation constitute the realm of discretionary decision making, measurement the realm of non-discretionary rule application.

The Boundaries of Measurement

The notion that the judgmental legislation of measurement rules and their bureaucratic application can be held apart in a neat manner rests upon demanding presuppositions about the nature of concepts, the nature of reality, and the nature of rules. When applying formal rules of measurement, we can only dispense with our judgmental faculties to the extent that concepts are commonsensical,

facts obvious, and rules of translation determinate. Yet what happens when these ideal conditions of bureaucratic decision-making fail to apply?

Figure 2 maps variations in the clarity of rules and the clarity of facts as two orthogonal dimensions of measurement.[20] The former ranges from clear and determinative rules to absent, defective, or controversial rules (designated by point zero). The latter ranges from obvious and commonsensical facts to absent, opaque or controversial facts (again designated by point zero). The figure illustrates two central claims that form the argumentative axes of my subsequent discussion:

Unfeasible demands: Just like the ideal of non-discretionary bureaucratic and judicial decision-making, the ideal of non-judgmental measurement is unreachable. It is a regulatory idea, not a feasible practice. Rules are never fully determinate and realities never fully transparent. This holds even for simple measurement assignments, like establishing the presence of national legislatures, which may become complicated when states are weak, fragmented, and contested, so that it is unclear who makes the laws or whether anybody makes anything resembling laws. The textured regions of Figure 2 show the extremes of unviable methodological demands. The textured region along its Eastern border shows the structural limits to factual clarity, the textured region along its Northern border the structural limits to regulatory clarity.

Minimal standards: Even if we cannot aspire to maximal clarity of facts or rules, we do need to aspire to minimal clarity of facts and rules. Even if we cannot reach the methodological ceiling, we still need firm methodological ground to stand upon. If we are to trust the quality of their judgmental data, even renowned firms, like Political Risk Services, that offer cross-national estimations of abstractions like bureaucratic quality and the rule of law, need to provide minimal information on their measurement rules and sources, rather than referring us to informal conversations on their weblog to learn "what is on [their] experts' minds."[21] The grey regions of Figure 2 show the prohibited areas of methodological relaxation. The Western region shows excessive factual opacity, the Southern one excessive discretion.

The boundaries of feasible clarity and necessary clarity delimit a property space that accommodates four basic modes of measurement that are all methodologically viable as well as methodologically licit, but differ according to the quality of rules and facts they employ. (a) When both rules and facts are sufficiently clear, measurement can approximate the "bureaucratic" model of mechanical decision-making. (b) When neither rules nor facts are sufficiently clear, measurement needs to rely on "judgmental" modes of decision-making. (c) When empirical realities are nebulous, researchers may strive to compensate for the opacity of facts by "hyper-regulating" the measurement process through myriads of minute rules (as I will illustrate below in a hypothetical example concerning electoral integrity). (d) When facts seem obvious, researchers may renounce the guidance of formal rules and take measurement decisions in a "commonsensical" fashion.

Common sense is what we can trust to understand implicitly without need of "making it explicit" (Robert Brandom). For instance, when Alberto Alesina and his co-authors (2003) introduce their cross-national data on ethnic fractionalization without explaining their conception of ethnicity, they appeal to commonsensical understandings of ethnic divisions. For coding rules to be fully bureaucratic, i.e., to allow implementation in a mechanical fashion that does not require further reflection, we need to rely on common sense as well. Otherwise, we would be drawn into infinite regresses of formal operational definitions. For example, if we follow the Cross-National Time-Series Data Archive and define "riots" as any "violent demonstration or clash of more than 100 citizens involving the use of physical force,"[22] establishing the outbreak of riots is an unproblematic "bureaucratic" exercise (only) as long as we can rely on commonsensical notions of violence, demonstrations, and citizenry. In political measurement as elsewhere, common sense is the ultimate anchor of bureaucratic regulation.

The remainder of this article will focus on delimiting the boundaries of political measurement, while giving only passing attention to the four types of measurement that lie within them. The next two sections delineate structural limits to "pure observation" and "mechanical rule application." These limitations involve the necessity of accepting certain degrees of judgment in our measurement practices, in the observation of facts as well as in the observance of rules. The methodological demand that we should base our measurement decisions on facts and rules, and nothing but facts and rules, is self-deceptive. Judgment is unavoidable. However, the opposite notion that we can base our measurement decisions on judgment, and nothing but judgment, is self-defeating, too. If we embrace the need to develop judgmental measures in the study of politics, we need to embrace the corresponding need of "domesticating" or "rationalizing" judgment through the definition of minimal standards. We need to turn judgmental measurement from a private practice into an intersubjective procedure. The final section offers a tentative catalogue of regulatory standards for judgmental measures.

The Limits of Observation

In empirical terms, for political measurement to operate in a non-judgmental fashion, we need (1) transparent empirical phenomena whose observation do not depend on our judgmental faculties and (2) complete public records on those phenomena. In the study of politics, both conditions are problematic. Table 2 provides an overview.

Table 2
Unclear facts and the role of judgment in political measurement

Elements of Measurement	Methodological challenges	Analytic dimensions	Role of judgment	Regulatory remedies
Empirical phenomena	Unobservable realities	• Symbolic realities (such as institutions, actions, written and spoken texts)	Comprehension	None
		• Causal relations and counterfactuals (as they are constitutive to concepts)	Inference (assumptions and arguments)	None
		• Subjective realities (such as intentions, perceptions, and moral commitments)	Comprehension, reconstruction from speech and behavior	Selection of observable proxies (or secondary sources)
Empirical information	Unobserved realities	• Incomplete or uneven information	Filling in gaps on the basis of local knowledge. Ad hoc amendment of rules of ignorance	Ad hoc amendment of rules of ignorance
		• Inconsistent or contradictory information	Adjudication on the basis of local knowledge. Ad hoc amendment of rules of rules of adjudication	Ad hoc amendment of rules of adjudication

Unobservable Realities

Observation is the cornerstone of methodological positivism. What we see is what we believe in.[23] Privileging the eye over other human organs may be a plausible methodological choice—as long as there is anything *to see.* The common language of measurement theory suggests that political measurement consists in putting numbers on visible, tangible objects, just like physicists did back in the old days, before they started inventing complex machineries to observe invisible things, phenomena too small or too distant to be registered by the human eye.[24] Political methodology textbooks use to tell us that our measures carry systematic information on "observable implications" of our theories and derive from "the observation of facts."[25] Yet this is a methodological illusion. In the study of politics, most of the empirical phenomena we are interested in are not accessible to direct observation—not to "pure" observation, anyway.[26]

Meaning. Social sciences are exercises in "double hermeneutics" (Anthony Giddens): we try to make sense of others (and ourselves) who are trying to make sense of others (and themselves). We do not study inanimate, objectively given realities, but symbolic, socially constructed and intersubjective realities. To grasp them, we need to understand them. Simple and pure observation won't do. Martians can't practice social sciences—at least, they can't on Earth. They need more than to register the outward movements of men and objects in order to understand the actions and institutions they are watching. For instance, by merely witnessing conglomerations of people sitting in ostentatious buildings, raising their hands from time to time or pushing buttons, they cannot comprehend the practice of law making. Whenever we talk of "observation" of social or political realities, it is never "pure" observation we are referring to, but meaningful observation, guided by our pre-existing social knowledge and conceptual tools. In this sense, we are all interpretivists. We all interpret pre-interpreted realities.

This may be an uncontroversial claim.[27] We all know that facts do not speak for themselves. Yet, since we sometimes speak as if they do, it seems pertinent to recall the intrinsically interpretive nature of our observations.[28]

Causes and counterfactuals. Notoriously, we cannot observe causation or the counterfactual states that causal reasoning invokes (such as the absence of effects in the absence of necessary causes). Our disciplinary reflections on "causal inference" accordingly revolve around problems of explanation under conditions of limited observation.[29] Yet, the non-observable nature of causal relations and counterfactual worlds creates methodological challenges not only at the level of explanation, but also at the level of description, and thus of measurement. Causes and counterfactual conditions are often integrative parts of the phenomena we try to describe and explain. They are built into the very concepts we try to measure.

For example, the concept of vote buying assumes that buyers and sellers of votes establish effective relationships of commercial exchange.[30] If we wish to measure "vote buying," it is not enough to estimate either the amount of client-list investments that parties and candidates realize, or the magnitude of electoral support they receive. We need to establish the causal relationship between the two—which is a demanding enterprise that involves, among other things, counterfactual reasoning about voter choices in the absence of vote-buying efforts. Any effort to measure concepts that rest upon causal assumptions (be they overt or hidden) cannot live on observation alone.

Subjectivity. The bounded world of subjective beliefs, values, and emotions is shut off from external inspection. There are indirect ways for gaining access to the realm of subjectivity. We can ask people about their thoughts, desires, and feelings, and take seriously what they tell us. Or we can watch how they behave and try to decipher the outward symptoms of their inner states. Yet, notoriously, regardless of how we try to comprehend what goes on inside the minds and hearts of others, we cannot observe it. We can only infer it from what we hear and see.

Scholars who investigate the realm of subjectivity, as in the study of public opinion, are familiar with the difficulties of making visible phenomena that are essentially invisible. The non-observable nature of research objects is less obvious in those cases in which ostensibly observable phenomena contain inbuilt elements of subjectivity. For instance, the notion of political violence refers to observable acts as well as to subjective motivations. Its political motives (however defined) distinguish political violence from other forms of violence, such as domestic violence or criminal self-enrichment through the use of force. What we can *see* (although in most instances, luckily, not in the first person) is the exercise of violence. What we need to *infer* (often based on knowledgeable judgments by others) are the motives that drive the exercise of violence.[31]

Unobserved Realities

Some empirical phenomena we cannot observe in principle, others we cannot observe in practice. More often than not, the high informational demands of bureaucratic observation cannot be met in reality. In particular, in the comparative study of politics, empirical information is often incomplete or inconsistent.

Incomplete information. Despite the dizzying expansion of cross-national political datasets in the past years, we lack basic information on innumerable questions in comparative political inquiry. Entire spheres of politics and categories of data are off our screens. For instance, we suffer from chronic and systematic information shortages with respect to: (1) political phenomena like crime and corruption that are hidden from public view due to their illicit nature;[32] (2) political phenomena like contentious actions or subnational processes that are observed by domestic agents but hard for international scholars to collect from the outside; and (3) official data on state institutions and decisions, such as judiciaries and judicial findings, that are generated by national public agencies but not pooled at the international level.[33]

In the complete absence of information, direct or indirect, on political phenomena, researchers must abandon hope of measuring them. But usually we do have at least some bits and pieces of information about at least some cases we are interested in. In such situations of incomplete information, we have to bridge informational gaps either by relying on contextual knowledge (expert judgment), or by devising general rules that deal with incomplete information (rules of ignorance). The creation of ad hoc rules in response to emergent measurement problems is a matter of judgment. Yet, once created, those rules can be used to resolve analogous problems in the future (as well as, if necessary, to recode data collected in the past).

Inconsistent information. Data from different sources are likely to diverge due to observational error. Moreover, in the study of politics, they are likely to diverge due to political bias. The providers of information are often parties to political struggles in which information itself represents an essential resource. For example, social movements and government agencies notoriously tend to diverge in their estimates of attendance at anti-government demonstrations. In such situations of inconsistent information, once again researchers can arbitrate between diverging accounts by relying on contextual knowledge (expert judgment), or they can devise general rules that deal with informational inconsistencies (rules of arbitration).

The Limits of Regulation

In regulatory terms, the requirements of impersonal measurement are similar to the requirements of impersonal justice. Rules, not people, are to determine the relation between facts and decisions. For political measurement to operate in a non-judgmental fashion, we thus need (1) complete, consistent, and determinative rules that eliminate discretion in the assignment of numbers to the phenomena we observe, as well as (2) low levels of conceptual complexity (abstraction, dimensionality, and aggregation). Table 3 provides an overview.

Fuzzy Boundaries

In a complex and changing world, we should always be prepared to encounter cases that do not quite fit our conceptual boxes and operational guidelines. We create our concepts and trace their boundaries, but we do not create and control the realities we try to measure. It's like the settlement of legal disputes: universal law can never foresee the infinite variation in particular cases that may arise in the future. Thus the need for judicial decision-making. Our universal rules of measurement try to grasp heterogeneous and evolving realities. They may fit well the standard situations we had in mind in devising them, but we are likely to encounter problems of application in the "gray zone"[34] of non-standard cases where our generalizing assumptions about the structure of the world do not hold. Even the ostensibly simple task of establishing the presence or absence of a phenomenon may turn problematic. For instance, when counting opposition parties in non-democratic regimes, we may encounter parties that belong

Table 3
Unclear rules and the role of judgment in political measurement

Elements of Measurement	Methodological challenges	Analytic dimensions	Role of judgment	Regulatory remedies
Rule application	Unexpected realities	• Unforeseen, hard cases, borderline cases, that undermine mechanical applications of coding rules.	Application of rules in the light of their spirit. Ad hoc amendment of boundary rules	Ad hoc amendment of boundary rules.
Concept structure	Conceptual complexity	• Abstract (rather than concrete) concepts: multiple levels between root concept and indicators	Synthesis and integration of information	Conceptual jumping
		• Composite (rather than simple) concepts: multiple sub/dimensions	Synthesis and integration of information	Litmus tests
		• Aggregate (rather than singular) referents: spatial, temporal, or social aggregation	Synthesis and integration of information	Rules of representation

to the opposition in name, yet are created, sponsored, directed, subverted, or manipulated by state agencies.[35] Again, we can make sense of such "hard" or borderline cases either by relying on contextual knowledge (expert judgment) or by devising general rules that resolve problems of operational delimitation (boundary rules).

As stated above, the observation of social realities depends on our interpretive faculties, our capacity to understand symbolic realities. The elementary dependence of observation on interpretation cannot be bridged, circumvented, or mitigated by any amount of bureaucratic regulation. Whether we like it or not, whether we recognize it or not, we have to live with it. There is no escape either from the fact that we cannot observe causal relations or counterfactual worlds. Establishing causes and counterfactuals is a matter of argumentation (causal inference), not observation. By contrast, there are bureaucratic remedies for all the other challenges to bureaucratic rule application in political measurement that I have discussed so far, at least in principle.

We cannot directly observe subjective realities, but we can often devise indirect indicators, observable (visible and readable) symptoms that reveal underlying subjective realities. We often have to deal with incomplete or inconsistent information and cannot foresee all possibilities of informational gaps and contradictions. Yet we can devise rules of ignorance and rules of adjudication that anticipate some of these informational problems; and we can amend these rules if we encounter novel problems in the process of measurement, so that we have general rules to guide us next time when we stumble over similar problems. The same applies to hard cases that inhabit the gray zone between categories or measurement scores: If we encounter borderline cases that are difficult to make sense of on the basis of existing coding rules, we can amend these rules and thus provide formal and explicit guidance for similar measurement decisions in the future.

These regulatory bridging devices allow us to measure what we cannot observe, and to create supplementary rules of measurement when the application of existing rules runs into difficulties. They work well as long as levels of conceptual and empirical complexity remain moderately low. Yet, they become unfeasible in the face of abstract, multi-dimensional, and/or aggregate concepts.

Complex Concepts

In scientific measurement, just as in real life, we can try to eliminate discretion and surprise by weaving dense webs of authoritative regulation. As long as the phenomena we are subjecting to formal and explicit rules are relatively simple, stable, and clearly bounded, regulation may indeed work to create predictability and constrain human agency. But in politics, at higher levels of complexity, the legislative pretense that one can foresee everything and regulate everything is certain to create bureaucratic nightmares. (Recall the tragic comedy of communist command economies.) Similarly, when the concepts we wish to measure reach a certain degree of complexity, the methodological pretense that we can devise a full catalogue of coding rules that establishes clear and precise links between all possible elements of empirical evidence (including all possible gaps and inconsistencies of evidence) is likely to produce a bureaucratic nightmare, too. In the face of complex realities, the notion of complete, consistent, and determinate law represents an idealized fiction—in the realm of legal regulation as much as in the realm of methodological regulation.

Conceptual complexity. In comparative political measurement, simple concepts can be quantified in bureaucratic fashion. They can be measured through the mechanical application of formal rules that relate observations to numbers without giving rise to doubt or ambiguity. Tellingly, it is not easy to come up with clear-cut examples. Still, the holding of elections, the abolition of national legislatures, and the declaration of international war (all categorical events) may count as plausible exemplifications. By contrast, the measurement of complex

concepts—such as state capacity, the rule of law, electoral integrity, civil society, and many others—is generally not susceptible to full bureaucratization (except at the price of radical reductions of complexity). Their observation and measurement imposes informational demands that cannot be processed though formal regulation. They can only be met by expert judgment, grounded in analytical and synthetic competence as well as in local knowledge. Conceptual complexity entails three structural properties: abstraction, composition, and aggregation.

1. *Abstraction:* Complex concepts are abstract, rather than concrete. Situated at high levels of generality, they oblige researchers to travel a long way on the road from conceptualization to operationalization. To get from the general concept to concrete indicators they have to laboriously descend the ladder of abstraction by multiple steps.
2. *Multi-dimensionality:* Complex concepts involve multiple dimensions and subdimensions. The challenge of anchoring abstract ideas in concrete realities multiplies by the number of dimensions a concept accommodates.
3. *Aggregation:* Complex concepts refer to aggregate, rather than singular, phenomena. Aggregation may be spatial (across territories), temporal (across time), or social (across groups of actors). In the comparative study of politics, concepts routinely refer to properties of national political systems. These macro-level properties often do not capture single events at the center stage of national politics. Rather they represent aggregate results of countless events that take place in a decentralized fashion, on countless locations far away from the capital city. Measurement thus involves the challenge of collecting and aggregating streams of information that tend to be overwhelmingly rich and hopelessly incomplete at the same time.

These three structural features of conceptual complexity may vary independently of each other, but they are frequently associated. Consider the integrity of elections. Within the liberal consensus, the existence of "free and fair" elections represents a constitutive dimension of modern representative democracy. Political elections are notoriously complex processes, however, and so is the concept of electoral integrity. Integrity is an aggregate, abstract, and multidimensional concept: Rather than distinct decisions by single actors, integrity describes patterns of interaction among state actors and citizens in a territorial state. Rather than concrete properties of specific actions (such as the expediency of candidate registration by electoral authorities), integrity denotes abstract properties of administrative and judicial systems (such as the impartiality of electoral dispute settlement). And rather than a single dimension (such as the formal independence of the national election commission), integrity comprises multiple dimensions (such as bureaucratic integrity and the absence of violence).

Hyperregulation. Translating complex concepts into the language of numbers usually requires the knowledge and judgmental faculties of experts. If we try to subject their measurement to bureaucratic standards we invite regulatory pathologies. The idea of electoral integrity exemplifies the dependence of complex concepts on expert judgment. The expanding archive of case reports produced by the community of election observers over the past two decades, as well as the expanding scholarly literature on electoral manipulation, have borne testimony to the difficulties of ascertaining the democratic quality of elections—particularly when they are neither clearly democratic nor clearly dictatorial.

As election experts Jørgen Elklit and Andrew Reynolds affirm, the "systematic study of election quality" requires the systematic study of eleven dimensions of electoral processes: legal framework, electoral management, constituency and polling district demarcation, voter education, voter registration, ballot design, party and candidate nomination and registration, campaign regulation, polling, vote counting and vote tabulation, electoral dispute resolution, and post-election procedures. According to the authors, to ascertain the democratic quality of these dimensions we have to address a total of fifty "performance indicators" (an average of just under five indicators per dimension). Many of these indicators, such as the questions on levels of violence and voter intimidation, contain further subdimensions and demand collecting information on the entire voting process, with thousands of locations and millions of actors.[36]

The complexity of the measurement task at hand is staggering. If we were to try to devise coding rules that would allow election observers to assign numbers to all fifty performance indicators in a mechanical fashion, without exercising discretionary decision-making power, we would need at least a dozen rules for each indicator, most likely many more than that. If it were only ten rules per indicator, we would need five hundred coding rules (!) to assess the democratic quality of an election (and then, of course, more rules to determine their aggregation). This is a regulatory and administrative nightmare, a recipe for mental as well as methodological insanity (a state defined by the northwestern corner of "hyper-regulation" in Figure 2, above). No wonder that Elklit and Reynolds reach the conclusion that most of their indicators have to be ascertained through "expert panel assessments."[37]

1. *Bureaucratic shortcuts:* Are there any methodological shortcuts that allow us to measure complex concepts bureaucratically, without recurring to the judgmental faculties of experts? There are some, indeed, although they come at the cost of radical simplification

2. *Conceptual jumping:* We may bridge the complexities of abstraction through "conceptual jumping," that is, by drawing direct linkages between abstract concepts and concrete indicators, while ignoring all intermediate levels. Although the most abstract or "basic level" of conceptualization is usually "too abstract and complex to be directly converted" into specific indicators,[38] conceptual jumping has been the "dominant" approach in statistical research where it is habitual to see scholars connect abstract concepts and quantitative indicators (misnamed "proxies") without mediating steps.[39] Consider common practices in quantitative comparative politics, like using per capita GDP as a proxy for societal modernization, public revenue as a proxy for state capacity, or linguistic fractionalization as a proxy for ethnic conflict.
3. *Litmus tests:* We can circumvent the complexities of multidimensional concepts if (and only if) we can come up with reliable "litmus tests," that is, if we can identify specific symptoms whose presence firmly indicates the presence of the general concept we wish to measure. If we can observe such symptoms, we need not observe the underlying condition that produces them. Here, in the compelling logic of symptoms, resides the ingenuity of Adam Przeworski and his collaborators' designation of alternation in executive power as a key indicator of the democratic quality of elections.[40] Authoritarian elections preclude opposition victories at the polls (substantive certainty); democratic elections make them possible (substantive uncertainty).
4. *Samples:* We may reduce the complexities of aggregation by devising "rules of representation," that is, rules and procedures that allow us to select a subset of observations we can plausibly take as "representative" of the whole. Random sampling, the selection of worst observations, and the selection of end-of-year observations are examples of such strategies.

Circumnavigating the complexity of concepts through bureaucratizing remedies (conceptual jumping, litmus tests, and sampling strategies) involves radical reductions in conceptual and operational complexity. If we are not prepared to accept the courageous simplifications (and ensuing losses of validity) they often impose, if we wish to measure complex concepts at higher levels of complexity, we must rely on the local knowledge and judgment of experts. The challenge is to reach measurement judgments in valid and reliable ways, without giving way to subjective arbitrariness.

The Rationalization of Judgment

Reliability constitutes the core value and main achievement of bureaucratic measurement procedures. Although solid bureaucratic measures ideally would be both valid and reliable, effective bureaucratic regulation only guarantees their reliability, not their validity. It ensures that different researchers who measure the same phenomenon on the basis of shared conceptual choices, empirical evidence, and measurement procedures are likely to reach similar results. The central value of judgmental measurement, by contrast, is validity—the validity of informed and reasoned public argument. Expert judgments do not strive to be replicable, but convincing. Their validity remains uncertain, however, as long as the validity of the arguments behind the judgments remains uncertain. Judgmental measures need to put a clear distance between themselves and subjective measures.

Measurement is of little use if the numbers experts assign to cases correspond to patterns of private reasoning, rather than patterns of empirical reality. In the face of complex concepts and complex realities, the general case for judgmental data development is strong. Still, to gain confidence in the quality of experts as well as in the quality of their measurement decisions we need to "discipline" or "rationalize" the process of judgmental data construction. We must not suffocate judgmental elements through finely knitted webs of bureaucratic regulation. But we do need to define demanding common standards and operating procedures in five crucial areas: expert selection, measurement comparability, transparency, convergence, and accountability.

1. *Expert selection.* By contrast to bureaucratic measures, expert judgments are not supposed to be impersonal. Coders of factual observations are fungible, experts are not. While the identity of the former must not matter for the results of factual measurement, the identity of the latter is constitutive for the construction of judgmental data. The quality of expert judgments therefore depends, first of all, on the quality of experts. The identification and selection of genuine experts is key to the production of genuine expertise.

Expert polls as well as other forms of judgmental measurement often fail on this account. For instance, while the notion of representative random sampling makes limited sense in expert communities (either you know or you don't), expert polls face recurrent critiques against their use of "very small samples of individuals."[41] In response, following the misguided notion that "more is better," they sometimes strive to maximize their number of data entries (through snowballing and self-appointment), which makes them prone to include individuals with highly varying degrees of knowledge on the specific theme under investigation. In cross-national research, expert surveys often recruit individuals who are simply identified as general "country experts," even when survey questions demand precise knowledge on rather specific substantive issues *within* countries. In particular in small and poor countries with fragile social-scientific infrastructures, few people if any are likely to have the necessary information on specific policy fields or institutional arenas.

2. *Comparability.* For expert judgments to serve the purpose of cross-national comparison (or any kind of comparison, for that matter), they must be themselves comparable. That is, the numbers they assign to countries or phenomena within countries must make sense across countries (or other units of observation). If the same numbers mean different things to experts from different countries (or to different experts within one country), they are useless for comparative purposes. For example, if the think tank Freedom House, in its survey on *Nations in Transit,* asks its country experts whether "the government [is] open to meaningful citizen participation in political processes and decision-making,"[42] we need shared standards (not just shared concepts) to ensure that their judgments are meaningful in comparative terms. Observers may judge prevailing levels of "openness" to citizen participation on the basis of various criteria, such as democratic ideals, past experiences, paradigmatic cases, or regional averages. In the absence of normative anchors—of common, explicit and transparent benchmarks—their measurement decisions will lack the intersubjective quality that defines meaningful judgment.

The best way to ensure common standards is to explicate them clearly. Illustrating abstract criteria through concrete examples or introducing well-known paradigmatic cases might ease the task of constructing such collective anchors. Reducing the complexity of concepts by disaggregating them helps, too. Conceptual disaggregation constitutes a key feature, for instance, of the ambitious, expert-based *Varieties of Democracy* project set up to create more than 180 indicators of regime characteristics for all independent polities worldwide, from 1900 to present.[43]

3. *Transparency.* Reliability is a standard we demand from the repeated application of measurement procedures to invariant empirical phenomena. If we apply identical procedures to varying phenomena, we have no reason to expect the results to be reliably similar. Cross-national political datasets, in particular those based on expert judgements, commonly fail to disclose their information sources in a systematic and transparent manner. Frequently we learn that dataset authors rely upon a certain range of information sources, without knowing the precise information bases that motivated specific coding decisions. For example, in its recent annual reports on the state of freedom in the world, Freedom House publishes selective listings of more than 320 periodicals and over 170 organizations that it draws upon to generate its global estimates of political rights and civil liberties.[44] We learn about the rough contours of its experts' field of vision, but cannot know what exactly they have been looking at when making concrete coding decisions. In the end, we cannot relate numbers to observations in a minimally precise fashion.

Of course, one of the key assets experts bring into the measurement process lies precisely in their capacity to process and synthesize large amounts of dispersed information. In the end, it is usually impossible for them to relate the numerical conclusions they reach to the precise pieces and bits of information that have gone into them. Nor are they in a position to provide an algorithm that would trace the mental process of reasoning that led them from the assessment of empirical evidence to the assignment of scores. And yet, even if experts are unable to describe all the miniature pieces that comprise a complex mosaic of knowledge generation and analytic judgment, they should be able to document the big picture. They should be able to provide, not each and every source and mode of reasoning that have gone into their measurement decisions, but the pivotal ones. And just like historians, they should be able to describe the range of uncertainty and controversy regarding their judgmental decisions with reference to concrete documentary evidence (or the lack of such evidence).[45]

Sometimes, cross-national data collectors mobilize country experts, not because of their judgmental faculties, but because of their informational advantages. Rather than asking them to apply complex concepts to complex realities, they ask them to provide country information outsiders do not have ready access to. For instance, when the *Varieties of Democracy* project invites coders to assess whether political parties have "well-defined, consistent, and coherent ideologies," it asks for expert judgment. When it inquires into "the number of parties gaining seats in the national legislature," it seeks expert information.[46] Data developers should distinguish clearly between judgmental and informational tasks, and vary their requirements of source documentation accordingly.

4. *Divergence.* Since experts may not fully converge in their assessments, data producers must have some way of reconciling their divergences. Standard textbook advice to adjudicate among diverging coding decisions through random procedures seems reasonable in the case of bureaucratic measurement, but it makes little sense in the case of expert judgment.[47] Expert surveys, in which data producers collect judgments by *external* experts, tend to rely on *additive* procedures (the calculation of arithmetic means) that grant equal weight to all individual judgments. Examples are the Legislative Power Index assembled by Steven Fish and Matthew Kroenig[48] and the data on subnational regimes in Argentina constructed by Carlos Gervasoni.[49] If multiple independent expert-based measures exist for some variable, we may treat them in a similar manner. For instance, borrowing multi-rater models from educational testing, Daniel Pemstein, Stephen Meserve and James Melton create "unified democracy scores" by integrating almost a dozen existing regime measures through Bayesian latent-variable analysis, which yields point estimates as well as estimates of uncertainty and coder reliability.[50]

Expert studies, in which data are generated by *internal* personnel within the responsible data agency (be it a research group, university department, non-governmental

organization, government agency, or international organization), tend to rely on *deliberative* procedures (the reconciliation of discrepancies through communication), in which it is, ideally, "the forceless force of the better argument"[51] that determines final measurement decisions. Freedom House scores of political rights and civil liberties (www.freedomhouse.org) and the Bertelsmann Transformation Index (www.bertelsmann-transformation-index.de) exemplify judgmental data that arise from layers of expert deliberation.

Although deliberation seems to be the most appropriate procedure to settle disagreements among experts, we still need to devise procedural guidelines that render it compatible with our methodological demands of transparency. The quality of expert judgments is defined by the quality of their public justification. Even if deliberative processes take place within closed circles of experts, their results must obey the principle of publicity. We should not trust experts blindly; we need to trust their arguments.

5. *Expert accountability.* In addition to improving ex ante mechanisms of expert selection, we also need to improve ex post mechanisms of expert accountability. When commissioning country "ratings from knowledgeable observers," current APSA president Henry Brady once suggested, "we might want to think about whether we should scale the raters as well as the countries that are rated."[52] To guard the guardians, judge the judges, rate the raters, we may invent codes of ethics and disciplinary mechanisms. The key to accountable expert measurement, however, is publicity. Rather than treating experts the same way as we treat survey subjects, whom we grant full anonymity, experts need to assume public responsibility for their measurement decisions. Establishing their personal responsibility involves two things. First, we need to know who codes what in which manner. If the identity of coders counts, it must be revealed. Second, coders must be prepared to defend their decisions and engage in processes of public debate and revision.[53]

Conclusion

The modern project of forging Weber's "iron cages of bureaucracy," of imposing formal rules to constrain human agency and assimilate it to the smooth operation of machines, regularly runs into obstacles. Some of them are epistemic and arise from the messiness of our concepts, the messiness of reality, or the messiness of our rules. Understanding and relating concepts, rules, and realities often requires elements of judgment. In all known realms of bureaucratic regulation, the notion of self-applying rules has revealed itself as a regulative idea that we may approximate, yet never fully realize. We may complain when public officials exercise judgment, and accuse them of usurping political functions ("bureaucratic discretion"). We may complain when judges exercise judgment, and accuse them of usurping legislative functions ("judicial activism"). And we may complain when hunters and gatherers of political data exercise judgment, and accuse them of subjectivity ("interpretive activism"). But we have to recognize that formal rules can constrain, but never eliminate the need for, human decision-making.

In this article, I strove to define and describe the empirical and conceptual conditions that require judgment (informed and reasoned decision-making) in the collection of political data: unobservable realities, unobserved realities, unexpected realities, and conceptual complexities. In the comparative study of political regimes these conditions are pervasive. If we were to renounce our judgmental faculties in the measurement of regime properties and regime dynamics, we would have to renounce the measurement of most of the most interesting regime properties and regime dynamics. If we truly had expelled judgment from data development, quantitative research on political regimes could not have blossomed as it has over the past decades. Yet also, if measurement experts had been able to exercise judgment on the basis of accepted methodological standards, the quantitative foundations of regime studies would be less controversial than they are at present.

In general terms, my conclusion is simple: To the extent that we need to rely on judgmental elements in the collection of political data (be it for epistemic, theoretical, or practical reason), we should recognize that fact, rather than deny it. Once we accept the essential role of judgment in political measurement, we can develop methodological standards that guide its transparent use, rather than exercise it in opaque ways that undermine its comparative advantages: the capacity to make sense of vast amounts of dispersed and incomplete information (informed decision-making), and the public justification of measurement decisions in the light of available evidence (reasoned decision-making).

The first step, recognizing the methodological legitimacy of judgment in political measurement, may be the hardest, as it requires us to jump over our methodological shadows. Yet, unless we are prepared to do so, we impoverish our methodological thinking as well as our substantive work, while damaging the credibility of both. Firstly, we need to take our bureaucratic standards for what they are: regulatory ideas, methodological idealizations. If we pretend we are putting them into practice when we cannot, we create a methodological version of false consciousness. Failing to practice what we preach, we push the discipline towards creative denial, rather than critical self-awareness. Just like bureaucrats who are forced to operate with laws that are detached from realities, we channel our energies into inventing ever more sophisticated ways of describing our violations of the spirit of the law as being in conformity with its letter.

Secondly, if our methodology leads us to exclude or to deny essential elements of our practice, our practice itself is likely to suffer. By narrowing our methodological field

of vision we are bound to narrow our substantive camps of vision as well. If we picture the ideal political scientist as the perfect bureaucrat who works only with simple commonsensical concepts, observes only simple commonsensical realities, and applies only simple commonsensical rules of quantification, we are bound to generate substantive knowledge only in the image of our methodological standards: simple and commonsensical.

Thirdly, we undermine the credibility of the knowledge we produce if our self-knowledge is based on methodological illusions. Why should anyone trust our representations of external realities if we offer misrepresentations of our internal practices?

Notes

1 See Brady and Collier 2004.
2 See Bevir 2008.
3 See Schedler and Mudde 2010.
4 On the structure of concepts, see Sartori 2009b and Schedler 2011. On the relational nature of social measurement, see also Adcock and Collier 2001 and Carmines and Woods 2005. On the notion of descriptive inference, see King, Keohane, and Verba 1994 (chapter 2).
5 Carmines and Woods 2005, 933.
6 See also Jackman 2008.
7 Thus the famous verdict by Giovanni Sartori 2009a (18): "*concept formation* stands prior to *quantification*."
8 See Snyder 2006.
9 Michell 2005, 678.
10 Ciuk, Jacoby and Pyle 2011, 1013.
11 Stevens 1946, 677 (emphasis added).
12 Przeworski et al. 2000, 55.
13 I am echoing older, Aristotelian distinctions between "faculties of the senses" and "the faculty of intelligence and understanding" that forms "the source of judgment" (Pettit 2008, lines 185–195). As a matter of course, measurement also demands a comprehension of numbers. Unless we understand the grammar of numbers, we cannot understand the semantics of translating concepts and observations into numerical symbols.
14 Bollen and Paxton 2000.
15 Kurtz and Schrank 2008, 8.
16 See Skaaning 2010, 456. On "judge-specific measurement errors" in democracy indices, see Bollen and Paxton 2000; in corruption indices, Hawken and Munck 2011.
17 The quotations are from the *Oxford English Dictionary Online* (www.oed.com), s.v. "judgement, judgment."
18 Kant 1943, line 220.
19 Modern states, of course, have since their inception been building measurement machines: large bureaucracies set up to survey the subjects and objects of state control, such as population, territory, taxable wealth, and political dissidence. In the social sciences, we make routine use of the administrative data these bureaucracies generate. Their logic of justification and operation lies outside the scope of the present essay, though.
20 The figure excludes the third dimension: the clarity of concepts its discussion would transcend the scope of this article.
21 See PRS homepage: (http://www.prsgroup.com/), accessed 6 August 2011.
22 Cross-National Time-Series Data Archive User's Manual (Banks 2008), Variable domestic6 (www.databanksinternational.com).
23 See also Johnson 2006.
24 Incidentally, if the natural sciences had circumscribed the realm of legitimate evidence to "observables" that we can register with our sensory organs, they would have hardly developed much beyond Aristotle.
25 King, Keohane, and Verba 1994, 24, 29.
26 The reification of realities tends to go together with the reification of concepts. Both are treated like physical objects. See Bevir 2008 (64) and Bevir and Kedar 2008.
27 Quite obviously, I am sidestepping the rather arcane methodological debates on the notion of interpretation that define imaginary cleavages in US political science. For a critical review of the (often opaque and inconsistent) philosophical foundations of political science, see Bevir 2008.
28 The big question here concerns the nature of legitimate evidence in the study of politics. Given their identical status as linguistic expressions, it is somewhat ironic, for example, that written texts are commonly treated as "observable" and thus legitimate pieces of evidence, while spoken texts are not. For an insightful treatment of the difficulties involved in coding legal text of varying clarity, see Melton, Elkins, and Ginsburg 2010.
29 See King, Keohane, and Verba 1994.
30 See Schaffer and Schedler 2007.
31 Note, though, that the contemporary debate about private versus political explanations of civil war ("greed" versus "grievance") tends to obliterate the distinction between private and political violence (see Collier and Hoeffler 2004).
32 Of course, most illicit acts do not go "unobserved" in a literal sense. Someone is watching, even if just the perpetrator and the victim. Yet private observations do not count in scientific measurement. Only recorded and publicly accessible information does.
33 Schedler 2012.
34 Goertz 2006.
35 See Wilson 2005.

36 See Elklit and Reynolds 2005, Table 1.
37 Ibid.
38 Goertz 2006, 53.
39 Ibid, 55.
40 See Przeworski et al. 2000, 27.
41 Landman and Häusermann 2003, 30.
42 Freedom House, Nations in Transit 2011, "Methodology" (p. 13).
43 See Coppedge and Gerring 2011. If we do not trust our ability to hold experts to common standards, we might infer the meaning of their judgments behind their backs, after the fact, by offering them "anchoring vignettes," ordinal lists of concrete examples they have to assess along the same scales as their abstract questions (see King et al. 2003). Such ex post techniques of control may help us understand subjective biases of expert judgment, but not overcome them.
44 See Freedom House, Freedom in the World, 2010 Edition, "Selected Sources." (www.freedomhouse.org/template.cfm?page=351&ana_page=365&year=2010), accessed 1 July 2011.
45 See also Coppedge and Gerring 2011, 256. For a systematic treatment of proximity and transparency of data sources, see Lieberman 2010.
46 See Coppedge and Gerring 2011, 256.
47 Random choice of diverging coder decisions is a standard procedure in other disciplines, like psychology and media research, yet virtually unknown in comparative political science.
48 See Fish 2006.
49 See Gervasoni 2008.
50 Pemstein, Meserve, and Melton 2010.
51 Habermas 1981, 47.
52 Brady 2004, 64.
53 See also Coppedge and Gerring 2011, 258.

References

Adcock, Robert and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3): 529–546.

Alesina, Alberto, Arnaud Devleeschauwer, Sergio Kurlat, and Romain Wacziarg. 2003. "Fractionalization." *Journal of Economic Growth* 8(2): 155–194.

Banks, Arthur S. 2008. Cross-National Time-Series Data Archive, Databanks International. Jerusalem, Israel. (http://www.databanksinternational.com), accessed December 29, 2011.

Bevir, Mark. 2008. "Meta-Methodology: Clearing the Underbrush." In *The Oxford Handbook of Political Methodology,* eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Oxford: Oxford University Press.

——— and Asaf Kedar. 2008. "Concept Formation in Political Science: An Anti-Naturalist Critique of Qualitative Methodology." *Perspectives on Politics* 6(3): 503–517.

Bollen, Kenneth A. and Pamela Paxton. 2000. "Subjective Measures of Liberal Democracy." *Comparative Political Studies* 33(1): 58–86.

Brady, Henry E. 2004. "Doing Good and Doing Better: How Far Does the Quantitative Template Get Us?" In *Rethinking Social Inquiry: Diverse Tools, Shared Standards,* eds. Henry E. Brady and David Collier. Lanham: Rowman & Littlefield.

——— and David Collier, eds. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards.* Lanham, MI: Rowman & Littlefield.

Carmines, Edward G. and James A. Woods. 2005. "Validity Assessment." In *Encyclopedia of Social Measurement,* vol. 3, ed. Kimberly Kempf-Leonard. Oxford: Elsevier Academic Press.

Ciuk, David, William G. Jacoby, and Kurt Pyle. 2011. "Measurement Theory." In *The Encyclopedia of Political Science,* ed. George Thomas Kurian. Washington, D.C.: CQ Press.

Collier, Paul and Anke Hoeffler. 2004. "Greed and Grievance in Civil War." *Oxford Economic Papers* 56(4): 563–595.

Coppedge, Michael and John Gerring. 2011. "Conceptualizing and Measuring Democracy: A New Approach." *Perspectives on Politics* 9(2): 247–267.

Elklit, Jørgen and Andrew Reynolds. 2005. "A Framework for the Systematic Study of Election Quality." *Democratization* 12(2): 147–162.

Fish, M. Steven. 2006. "Creative Constitutions: How Do Parliamentary Powers Shape the Electoral Arena?" In *Electoral Authoritarianism: The Dynamics of Unfree Competition,* ed. Andreas Schedler. Boulder and London: Lynne Rienner.

Freedom House, Nations in Transit 2011. (http://www.freedomhouse.org/sites/default/files/inline_images/NIT-2011-Methodology.pdf), accessed July 1, 2011.

Gervasoni, Carlos. 2008. "Conceptualizing and Measuring Subnational Regimes: An Expert Survey Approach." Political Concepts Working Paper #23, IPSA Committee on Concepts and Methods. (http://www.concepts-methods.org/WorkingPapers/PDF/1029), accessed November 14, 2011.

Goertz, Gary. 2006. *Social Science Concepts: A User's Guide.* Princeton: Princeton University Press.

Habermas, Jürgen. 1981. *Theorie des kommunikativen Handelns: Handlungsrationalität und gesellschaftliche Rationalisierung.* Frankfurt am Main: Suhrkamp.

Hawken, Angela and Gerardo L. Munck. 2011. "Does the Evaluator Make a Difference? Measurement Validity in Corruption Research." Political Concepts Working Paper #48, IPSA Committee on Concepts

and Methods. (http://www.concepts-methods.org/WorkingPapers/PDF/1078), accessed November 14, 2011.

Jackman, Simon. 2008. "Measurement." In *The Oxford Handbook of Political Methodology,* eds. Janet M. Box-Steffensmeier, Henry E Brady, and David Collier. Oxford: Oxford University Press.

Johnson, James. 2006. "Consequences of Positivism: A Pragmatist Assessment." *Comparative Political Studies* 39(2): 224–252.

Kant, Immanuel. 1943. *The Critique of Judgment,* trans. John Miller Dow Meiklejohn. New York: Wiley. Mobilereference e-book.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton, NJ: Princeton University Press.

King, Gary, Christopher J.L. Murray, Joshua A. Salomon, and Ajay Tandon. 2003. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *American Political Science Review* 94(4): 567–583.

Kurtz, Marcus and Andrew Schrank. 2008. "Promises and Perils of Cross-National Datasets: Perceptions, Objective Indicators, and 'the Rule of Law.'" Paper presented at the American Political Science Association Annual Meeting, Boston, MA, August 28–31.

Landman, Todd and Julia Häusermann. 2003. *Map-Making and Analysis of the Main International Initiatives on Developing Indicators on Democracy and Good Governance.* Essex, UK: University of Essex Human Rights Centre and the Statistical Office of the Commission of the European Union (Eurostat).

Lieberman, Evan S. 2010. "Bridging the Qualitative-Quantitative Divide: Best Practices in the Development of Historically Oriented Replication Databases." *Annual Review of Political Science* 13: 35–59.

Melton, James, Zachary Elkins, and Tom Ginsburg. 2010. "On the Interpretability of Law: Lessons from the Decoding of National Constitutions." Political Concepts Working Paper #44, IPSA Committee on Concepts and Methods. (http://www.concepts-methods.org/WorkingPapers/PDF/1072), accessed November 14, 2011.

Michell, Joel. 2005. "Measurement Theory." *Encyclopedia of Social Measurement* vol. 2, ed. Kimberly Kempf-Leonard. Oxford: Elsevier Academic Press.

Neuendorf, Kimberly A. 2002. *The Content Analysis Guidebook.* Thousand Oaks, CA: Sage.

Pemstein, Daniel, Stephen A. Meserve, and James Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18(4): 426–449.

Pennings, Paul, Hans Keman, and Jan Kleinnijenhuis. 1999. *Doing Research in Political Science: An Introduction to Comparative Methods and Statistics.* London: Sage.

Pepinsky, Thomas B. 2007. "How to Code." Political Concepts Working Paper #18, IPSA Committee on Concepts and Methods. (http://www.concepts-methods.org/WorkingPapers/PDF/1034), accessed November 14, 2011.

Pettit, Philip. 2008. *Made with Words: Hobbes on Language, Mind, and Politics.* Princeton and Oxford: Princeton University Press. Kindle edition.

Przeworski, Adam, Michael E. Alvarez, José Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950–1990.* Cambridge, UK: Cambridge University Press.

Sala, Brian R., John T. Scott, and James F. Spriggs II. 2007. "The Cold War on Ice: Constructivism and the Politics of Olympic Figure Skating Judging." *Perspectives on Politics* 5(1): 17–29.

Sartori, Giovanni. 2009a. "Concept Misformation in Comparative Politics." In *Concepts and Method in Social Science: The Tradition of Giovanni Sartori,* eds. David Collier and John Gerring. New York and London: Routledge.

———. 2009b. "Guidelines for concept analysis." In *Concepts and Method in Social Science: The Tradition of Giovanni Sartori,* eds. David Collier and John Gerring. New York and London: Routledge.

Schaffer, Frederic Charles and Andreas Schedler. 2007. "What Is Vote Buying?" In *Elections for Sale: The Causes and Consequences of Vote Buying,* ed. Frederic Charles Schaffer. Boulder and London: Lynne Rienner.

Schedler, Andreas. 2011. "Concept Formation." In *International Encyclopædia of Political Science,* eds. Bertrand Badie, Dirk Berg-Schlosser, and Leonardo Morlino. London: Sage.

———. 2012. "The Measurer's Dilemma: Coordination Failures in Cross-National Political Data Collection." *Comparative Political Studies* 45(2): forthcoming.

——— and Cas Mudde. 2010. "Data Usage in Quantitative Comparative Politics." *Political Research Quarterly* 63(2): 417–433.

Skaaning, Svend-Erik. 2010. "Measuring the Rule of Law." *Political Research Quarterly* 63(2): 449–460.

Snyder, Richard. 2006. "Beyond Electoral Authoritarianism: The Spectrum of Nondemocratic Regimes." In *Electoral Authoritarianism: The Dynamics of Unfree Competition,* ed. Andreas Schedler. Boulder and London: Lynne Rienner.

Stevens, Stanley S. 1946. "On the Theory of Scales of Measurement." *Science* 103(2684): 677–680.

Wilson, Andrew. 2005. *Virtual Politics: Faking Democracy in the Post-Soviet World.* New Haven: Yale University Press.