On the Interpretability of Law:
Lessons from the Decoding of National Constitutions[1]

Melkinsburgtaru
(Zachary Elkins, Tom Ginsburg, Kalev Leetaru, James Melton)

March 20, 2010

I. INTRODUCTION

After almost two years of heated debate, drafting and re-drafting within committees, over 1000 roll call votes, and behind-the-scenes negotiation, the 587 delegates to the 1987-88 Brazilian Constitutional Assembly rested. The hard work was, theoretically, done. They now turned their product over to a "linguistic consultant" with whom the Assembly charged with the critical task of rendering their 245 Articles into readable Portuguese. Legal precision and carefully negotiated terms of the constitution were important, it seems, but this was to be a legal document that ordinary Brazilians would be able to read. Or so some thought. As it happened, most of the consultant's edits were rejected and the result, according to some at least, was a missed opportunity to achieve greater constitutional clarity. [2]

Clarity of language has a number of obvious virtues, especially in a document meant to usher in participatory democracy and the rule of law in a highly unequal society like Brazil, where it is estimated that one quarter of the population is illiterate. Notwithstanding disagreements about the precise definition and dimensionality of the Rule of Law, a central element of the concept is that law be clear and easily understood (Tamanaha 2004). Without clear law, citizens and legal decision-makers alike are

---

[1] Prepared for the Conference, "Measuring the Rule of Law" (School of Law, University of Texas at Austin). March 26-27, 2010

[2] *O Processo constituinte, 1987-1988*. Milton Guran (ed.).

more likely to produce inconsistent interpretations of the rules, and the law will be unable to provide predictability in social affairs. Many of the virtues of law will be thus be lost. Unclear law is indeterminate law, and indeterminate law is presumptively illegitimate (see Dworkin 1977). The stakes in understanding what features make law more or less interpretable, then, are quite high. Yet, notwithstanding the central importance of easy interpretability, we have little understanding of the factors that make the rules easier or more difficult to understand

This paper delves into this conceptual space by exploiting novel data on the interpretability of national constitutions. The data are derived from the Comparative Constitutions Project (CCP), a large-scale effort to assess the origins and consequences of constitutional choices across most independent countries since 1789. One of the by-products of this enterprise is a fairly systematic sense of the clarity, or interpretability, of a broad sample of constitutions. We analyze data from this coding exercise to develop a measure of interpretability and assess the various attributes of constitutional texts and constitutional settings that lead to more or less interpretability.

Written constitutions are a particularly fruitful milieu in which to seek to develop measures of interpretability for several reasons. Constitutions are core documents that establish the legal system and regulate ordinary lawmaking process, and so a constitution that is difficult to interpret will likely undermine the rule of law more broadly. Constitutional clarity also facilitates enforcement: without a clearly and plainly-written document, efforts to enforce the document (which, we shall argue, require widespread knowledge of, and attachment to, the constitution) will be feeble at best. Moreover, constitutions should be universally accessible documents in that they should be understood by legal professionals and laymen alike, by all members of a society no matter their language or cultural background, and perhaps most importantly, by future generations as well as contemporaries. If the constitution is highly context-dependent -- either culturally or temporally -- it unlikely to be serviceable

in highly fragmented societies and will not preserve its commitments across generations. In short, the very functions of a constitution will be gutted if it is unclear.

II. CLARITY AND THE RULE OF LAW

Discussions of the Rule of Law tend to start with the classic framework of Lon Fuller's (1964) *Morality of Law.* For Fuller, a rule had to have certain characteristics in order to be properly characterized as law. Fuller's criteria for law are: (1) consistency, which requires general rules; (2) publicity; (3) clarity; (4) non-retroactivity; (5) internal consistency in the sense of lacking contradictions; (6) potential compliance, meaning that the rules should not make demands not capable of being addressed; (7) stability over time; and (8) application as written. These features had, in Fuller's conception, an "internal morality" such that they were normatively desirable, independent of the substantive concept of law. The utter lack of any of these criteria would mean that the system could not be properly be characterized as meeting the rule of law.

This is a general framework that can be used to evaluate individual legal instruments (statutes, cases, rules), areas of law (e.g. administrative or corporate law regimes), or legal systems as a whole. At a conceptual level, then, measuring the rule of law implies measuring each of these attributes. We leave it to others to consider dimensionality of the concept and how the various attributes combine (if at all) to provide any overarching sense of the rule of law. Presumably, the overall extent to which the rule of law is found in any legal milieu might be thought of as some function of the individual measures of the component criteria. To be sure, one might argue that some of the criteria are more important than others and that some attributes are dependent upon others; at the very least, any overall measurement scheme will inevitably deal with the weighting of items and their functional form. Certainly, the practical challenges of measuring any one dimension will vary across rules and environments, and issues of comparability might constrain particular measurement choices.

Our focus is on clarity, Fuller's third criterion. As we shall see, however, clarity has direct implications for his first, sixth, and eighth criteria: (1) consistency, (6) compliance, and (8) application as written. But consider clarity, first. Why is it important that the law be clear? Fuller (1964:63) stated the problem thusly: "…it is obvious that incoherent and obscure legislation can make legality unobtainable by anyone, or at least unattainable without an unauthorized revision which itself impairs legality." This "obvious" problem is a pragmatic concern: in the extreme, one cannot expect any form of communication (including that about law) to be at all meaningful in a Tower of Babel. A lack of clarity has other downstream effects on the rule of law as well. We discuss three in particular: predictability and enforceability.

Predictability

Predictability is a central goal of the rule of law. As the Organisation for Economic Cooperation and Development (OECD 2005:2) put it, the rule of law concept "first and foremost seeks to emphasize the necessity of establishing a rule-based society in the interest of legal certainty and predictability." Legal theorists have long wrestled with the concept of predictability under the rubric of legal indeterminacy (Leiter 2007). As Dorf (2008) puts it: "For over a generation, the debates in jurisprudence and constitutional theory have struggled to reconcile the fact of considerable legal indeterminacy with, respectively, law generally and constitutional legitimacy in particular." The basic problem is that law is supposed to constrain decisionmakers. But if we only know what the law is after legal decisionmakers like courts and administrative agencies have deliberated about unclear meaning, then law cannot be said to provide an *ex ante* guide to behavior.

A closely related aspect of predictability concerns the consistent application of the law by officials. Without clear law, one cannot know how to behave and so the relationship between enforcers and subjects can potentially become arbitrary. When law is vague, it can be interpreted by the adjudicator in a discretionary fashion. In the words of F. A. Hayek (1944: 81), "One could write a history

of the decline of the rule of law, the disappearance of the *Rechstaat,* in terms of the introduction of these vague formulae into legislation and jurisdiction, and of the increasing arbitrariness and uncertainty of, and consequent disrespect for, the law and the judicature." Law, by its nature, requires the theoretical possibility of compliance, and unclear law can lead to its arbitrary application, either intentionally or unintentionally. Consistent application over time will be undermined.

Beyond its effect on the predictability of application by legal decision-makers, clear law helps subjects of the law organize their own affairs. This in turn requires inter-subjective understanding of the requirements of the law among subjects. If citizens have different interpretations of the rules, the law does not help them to coordinate their behavior and thus undermines social cooperation. To illustrate, a law requiring that everyone must drive on the right is functionally equivalent to a law requiring everyone to drive on the opposite side of the road as the British do. But the former requires fewer conceptual steps, and less knowledge about the state of the world. Citizens operating under the latter rule will have to expect others to have specific knowledge about British driving habits, and any citizen who is mistaken may cause an accident on the road.

Vague law thus implicates several aspects of Fuller's indicia of the rule of law besides clarity *per se*. If a law does not provide clear instructions to decision-makers *ex ante*, it is difficult to meet the criteria of being consistently applied as written. If interpreted differently by different subjects, the law loses its generality. These problems of multiple interpretation are likely to be exacerbated in heterogenous societies, where meanings may change across geography and culture, to say nothing of differences in interpretation across generations. This lack of consistent application and generality will have implications for Fuller's sixth criterion, compliance.

Enforceability

In the now classic line of work on self-enforcement, scholars such as Russell Hardin (1987), Peter Ordeshook (1992), Barry Weingast (1997), and Rui de Figueiredo (2005) have argued that constitutions

are sustained when they provide a focal point for private enforcement efforts. Only when the subjects of a constitution can credibly threaten to enforce it will constitutional order (and the rule of law) be sustained. Such "self-enforcement" is critical because, in most cases, there is no external agent who will enforce the rules of the constitution. Self-enforcement occurs when subjects of the constitution are willing to take costly action, and they will only do so when they believe that others will join them. This inter-subjective expectation is facilitated when the parties have common knowledge of the rules.

To generate self-enforcement, it is important that the constitution be clear. The constitutional text can provide a focal point (Schelling 1980) to help parties coordinate their enforcement efforts. A text that is clear is more likely to satisfy the criteria of common knowledge that helps to establish an effective focal point. To illustrate, when New Yorkers were asked where and when to meet someone for an unscheduled appointment, they chose Grand Central Station at noon not simply because they thought it a convenient place, but because they had expectations about what others would think was a convenient place. Similarly, when trying to evaluate whether a government has overstepped its role, it helps to have clear textual statements of the relevant constitutional rules so that all subjects can agree on the definition and predicate of a violation.

Note that an unclear text will fail to generate self-enforcement *even if every agent has the same interpretation of the constitution*. This is because the lack of clarity impedes the common knowledge necessary for an effective focal point. If the government violates my right to free speech, but the formulation of the right is so vague that I am unsure that others will share my definition, I will discount the probability that others will join me in the enforcement effort. This means that self-enforcement is less likely. The implication here is that *constitutionalism* requires clarity: effective limits on government will only be possible if they are clear.

III. ANALYTIC STRATEGY

Leveraging the Comparative Constitutions Project

Over the last five years, we have devoted much of our time to reading and interpreting a large set of historical constitutions. This experience, and the systematic manner we have undertaken the reading, allows for an assessment of constitutional clarity and interpretability. Constitutions, admittedly, constitute a very specific kind of law and we do not assert that our insights apply generally across all domains of law. However, of all types of law, clarity is arguably especially important with respect to constitutions. Unlike the fine print in a credit card contract, constitutions are intended to be plainly and clearly written. Moreover, because constitutional contracts are not enforceable by an external guarantor, the quality of self-enforcement (and therefore clarity) is especially relevant to constitutions (Weingast 1997; Hardin 1989; Ordeshook 1992).

Constitutions might fail to be clear for a number of different reasons. Most obviously, the language itself can be complicated and difficult to understand. But even if each of the individual provisions is relatively clear, a constitution might still be difficult to interpret because of the inter-relationship of its various parts. It might contain extensive cross-references, referring to definitions found elsewhere in the text. It might contain contradictions. Or it might describe complicated institutional schemes, which will require more skill to understand, even if the language used is itself clear.

Our *Comparative Constitutions Project* (online at https://comparativeconstitutionsproject.org) records some 668 characteristics of the written constitutions of independent states (including micro-states) since 1789. By our accounting, the universe of cases numbers 867 new constitutional systems, which have been "amended" 2,186 times (for more on constitutional systems and the distinction between new and amended see Elkins, Ginsburg, Melton 2009.) Figure 1 provides some sense of the universe of states and constitutions, as well as our sample as of February 2010. We are adding cases

weekly, but the current sample includes full information on 420 of the 867 constitutional systems, including nearly all constitutional systems currently in force and just under 50% of all constitutional systems since 1789. We say "full information," since our coding process involves several stages. Constitutions are coded (at least) twice by separate coders after which the codings are reconciled by a third person, who reviews all answers but focuses primarily on discrepancies in the answers from the original coders. For the purposes of this paper, our sample includes all constitutional systems that have been coded twice *and* reconciled.[3]

The theoretical motivations of the CCP project more broadly revolve around a set of historical questions about the origins and consequences of constitutional choices. However, the process of reading and interpreting constitutions in a systematic fashion yields a great deal of information about the ease of interpreting constitutions, more generally. A short description of our coding process will demonstrate some of analytical possibilities. Given the scope of our project, we have employed a set of graduate students (both in law and political science) and highly competent undergraduates to assist in the data collection.[4] Three graduate students in political science have worked with the project since the summer of 2005, when coding began, and some five or so others worked on the project for over three years over this period. In the hierarchy of the project, all research assistants begin as coders and some (those in the experienced group just mentioned) become "reconcilers" (see below) at a certain point. In total, 96 individuals have worked with us as coders, and of this group, 14 have become reconcilers.

Coding constitutions involves a certain back-and-forth between the project's survey questions (the online "survey instrument") and the text of actual constitutions. Coders go through an extensive training process and detailed instructions regarding known issues of interpretation are included both in

---

[3] In addition to these systems, we have 121 that have been coded twice but not reconciled and 102 that have been coded once. Hence, we have *some* data on 643 out of 867 constitutional systems (including nearly all systems since World War II). Since the unit of analysis below is the coder-reconciler dyad, we exclude the 223 codings that have not been reconciled our present sample.

[4] All personnel are listed (thanked!) on the project website (comparativeconstitutionsproject.org).

the "survey instrument" and a manual for coders. Of course, one doesn't know exactly where the shoe pinches until it is on the foot. Accordingly, we have developed a rather comprehensive process by which to adjudicate ambiguous cases, and communicate our decisions on these cases. The project's online portal includes a message board where coders post questions about interpretive problems. A typical coder will post three or four queries per constitution, although as we shall see, this rate varies by coder and by constitutional text. The principal investigators monitor the board and issue "rulings" on these cases. In turn, these rulings serve as the controlling instructions for future coders who face comparable issues. (The rulings can be searched and retrieved by topic).

The system bears some resemblance to a kind of legal system in miniature. Our primary regulators are the reconcilers, with the principal investigators serving as a court of final review for all decisions. The rulings form a system of precedent, recorded and easily accessible to all actors in the system. For the most part, the rulings are treated as settled law, although on several occasions a principal investigator has overturned a decision, an action which has then sometimes precipitated the retroactive coding of affected cases. As we discuss below, this ongoing process of review has decided consequences for the degree of shared understandings, as one would expect.

Interpretability and Reliability

We begin with the assumption that there is some determinate answer to each of our survey questions. That is, for example, to the question "Does the constitution provide for the right to free speech?", one should be able to produce inter-subjective agreement about the answer with respect to a given constitution. This does not mean that the answer should be a definitive "yes" or "no" in every case. Certainly, the answer could be an intermediate one, such as "yes under certain stipulated conditions." However, we assume that excluding errors of interpretation, multiple readers will reach the same conclusion about the answer. Of course any written text – from Solon's constitution to Shakespeare's *Othello* -- will communicate nuanced differences in meaning across readers. In the realm

of literature, our assumption would clearly be untenable. There is no right answer to whether, in killing Desdemona, Othello should be considered "the greatest poet of them all" (Bradley) or is simply "egotistical" (Leavis), to cite two prominent literary critics. However, if a constitution is to serve as a general contract underlying all political activity, we expect its terms to be mutually intelligible. We can therefore treat any inconsistency in interpretation across readers as measurement error.

As we would expect, coders are able to assess the meaning of constitutions with varying degrees of error, some of which is associated with their own characteristics, some with those of the constitution or the constitutional setting, and some with aspects of the coding process. In part, our goal in this essay is to disaggregate the error in interpretation into its various parts and assess the proportion of variation associated with attributes of the constitution and the constitutional setting. Assuming that we can isolate and estimate this component -- call it the "interpretability" or "clarity" of a given constitution -- we can then say something about how this quality varies across constitutions.

To state this more formally, we are interested in estimating the interpretability (Int) of i constitutions in j countries and we assume interpretability is a function of attributes of the constitution and the country in which it is written:

$$Int = \gamma_1 X_1 + \gamma_2 X_2 + u \qquad \text{(equation 1)}$$

where X_1 is an p x i matrix of p constitutional attributes, X_2 is an r x j matrix of r country attributes, u is a vector of error terms with rank i, and γ_1 and γ_2 are vectors of coefficients of rank p and r, respectively. However, interpretability is a latent variable, so we cannot estimate equation 1 directly. We can estimate the reliability of our measures, that is degree of consistency acrosss repeated measurement. We then assume that the reliabilities (Rel) of k codings of i constitutions are a function, in part, of interpretability:

$$Rel = \beta_1 Int + \beta_2 Z + \varepsilon \qquad \text{(equation 2)}$$

where Z is an q x k matrix of q coder attributes, ε is a vector of error terms with rank k, and β_1 and β_2 are vectors of coefficients of rank 1 and q, respectively. Substituting equation 1 for Int in equation 2 provides the reduced-form equation:

$$Rel = \beta_1(\gamma_1 X_1 + \gamma_2 X_2 + u) + \beta_2 Z + \varepsilon = \alpha_1 X_1 + \alpha_2 X_2 + \beta_2 Z + \beta_1 u + \varepsilon \quad \text{(equation 3)}$$

where $\alpha_1 = \beta_1 \times \gamma_1$ and $\alpha_2 = \beta_2 \times \gamma_2$. Thus, equation 3 allows us to estimate the effects of X_1 and X_2 on interpretability and even predict the interpretability of each constitution. A valid measure of coders' reliabilities, identification of all constitutional and country attributes that affect interpretability, and identification of all coders' attributes that might affect reliability are critical for equation 3 to provide unbiased estimates. We discuss each of these topics in the following sections.

Measuring Reliability

The dependent variable in the analyses that follow is the reliability of a coder's interpretation of a set of constitutional provisions. Our measure of reliability is a version inter-coder reliability, or the probability that two independent coders will provide the same answer to the same question. A number of issues arise in the calculation of this quantity. The first, given the particular structure of our coding procedure, involves the choice of coders at which to calculate inter-coder reliability. As we describe above, we have at our disposal two (or more) independent codings of each constitution, and we could simply measure the degree of inter-coder agreement of coder pairs. This is typically the way one calculates inter-coder reliability. Alternatively, since we also have a more authoritative interpretation of the constitution (the reconciled answers), we could conceivably assess the coders' reliability against this standard. Each of these approaches has its advantages. We choose the latter approach mostly because it increases the precision of our measure of reliability and because it allows us to assess the impact on reliability associated with the reconciliation process, which is of procedural interest to us. We therefore construct a dataset of coder-reconciler dyads, with a binary measures agreement (1) or disagreement (0) across a a set of items from the CCP survey instrument.

Accordingly, the next issue that arises, then, has to do with the selection of items across which to observe agreement. As we mention, the survey instrument includes 669 questions. However, not all constitutions speak to each of these questions (recall that we have identified an inclusive set of items in order to accommodate the wide inventory of provisions that drafters have thought to constitutionalize over the years). Some of these questions are "root" questions that ascertain the presence of a constitutional provision on a particular topic and are followed by branching questions that pursue the provisions in more detail. That is to say, some constitutions will have missing observations on some of these branching questions based on a coder's response to a root question. Further, survey questions come in several varieties. Some are closed-ended questions with mutually exclusive choices ("Does the constitution provide for the right to free speech") and non-exclusive choices ("Which of the following are requirements to as a member of the lower house of the legislature?"). Some are open-ended questions with restricted choices ("What is the length of the term of office for members of the lower house [numeric responses required, with a 0 for fixed length]?") and less restrictive choices ("Describe any details (other than those already covered) of the process of amending the constitution"). Finally, questions vary remarkably by their degree of error. Some questions are highly consensual across coders with almost no disagreement, while others exhibited levels of agreement as low as 20% (see the results below for some discussion of this variation).. In terms of estimating overall reliability, it makes sense to include items irrespective of their variation in error. However, if we are interested in explaining the variation in the degree of error, it may make sense to exclude the highly consensual items lest they dilute the informational value of the other items. A solution to this problem may be to include the full set of items, but weight them by their level of difficulty, measured either by the degree of agreement or by the frequency of interpretive problems (message board posts) on the topic.

Our approach to these various issues is to (1) use only closed-ended questions from the CCP survey instrument, (2) treat answers of "non-applicable" to branched questions like any other answer,

and (3) weight questions based on their difficulty (as measured by the agreement between coder and reconciler across all cases). Or, more formally:

$$\text{Rel}_k = \left(\frac{\sum_{v=1}^{n} (w_v D_{kv})}{\sum_{v=1}^{n} (w_v)} \right) \times 100 \qquad \text{(equation 4)}$$

where D_{kv} is the correspondence between the coder and reconciler on question v of n questions and w_v is the weight assigned to question v and is equal to 1 minus the percentage of coders who agreed with the reconciler for that question. One can think of the measure as a weighted percentage of the questions that the coder agreed with the reconciler. Table 1 provides a simple illustration in the case of two coders who have answered two questions, which are then reconciled. The coders' agreement with the reconciler (1 if they agree) is represented by columns D_1 and D_2, with the resulting weights of 0 and 0.5 and reliability scores of 100 and 0, respectively. The reader should note two important features of this measure. First, the weights are determined based on the available data, so as we acquire more data, the weights could change. However, since the weights used for the analysis below are generated from data with almost 1,000 codings, updates to these weights through the addition of new data should be minimal. Second, these weights should correct for the extremely high reliability of questions that are commonly left unanswered as a result of branching within the questionnaire. At the extreme, such the influence of such highly consensual questions is reduced to zero.

IV. EXPLAINING RELIABILITY

Inter-coder error will be associated with three sets of factors, which operate at various levels of measurement. These are (1) the characteristics of the coder, (2) those of the constitution or constitutional setting, and (3) those of the process. In this section we describe a set of hypotheses within each of these categories. We specify a fairly large number of hypotheses with the intention of

assessing as precisely as possible the proportion of error associated with each component. Our particular interest is in those characteristics associated with the constitution or its context.

Characteristics of constitutions

We begin with a discussion of error attributable to the constitutional text or the constitutional setting. We can think of that sort of error as falling within two basic categories: (1) that associated with the problem of making judgments across context; and (2) that associated with the syntax and structure of the text.

Context. Consider first the issue of context and, in particular, language, culture, and era. The basis of our analytical approach here is the assumption that coders socialized and trained in one context will have more difficulty with texts produced elsewhere and/or in different eras. Roughly 95% of our coders are young (twenty-something) U.S. citizens who, while unusually knowledgeable about political institutions (given their vocation), are not generally conversant with historical and comparative constitutional jurisprudence.

With respect to the temporal context, the question is whether contemporary readers can parse constitutional text written generations earlier. This is a question of obvious importance to constitutionalism more generally. Indeed the idea that constitutional commitments would constrain future generations is central to the very basis of higher law and, of course, is a source of the counter-majoritarian dilemma. It is hard to think how the "dead would govern the living" (Jefferson) if the living cannot understand the dead. The hypothesis that we test here is that the age of the constitution, measured from the year it enters in force, decreases inter-coder reliability.

The issue of interpretation across language is equally important. As we know, the words and phrasing chosen by constitutional drafters is often scrutinized, interpreted, and re-interpreted carefully. While constitutions in multiethnic states will be disseminated in multiple languages, distortions in translation can alter meanings as well as cloud them. If the important rule-of-law criterion of generality-

of-law is to hold, then the meaning of law should retain across translations. The empirical scope of this challenge is significant: depending how one counts, roughly half of contemporary states include a sizable minority group whose members speak a different language than do those in the majority. The constitutions of some countries have even been written in languages wholly foreign to the majority. The Norwegian constitution, for example, was drafted in 1814 in Danish, since a standard written form of Norwegian had not yet materialized. Even the original Norwegian versions of the text, first transcribed in the 1900s, were written in an archaic form of the language by today's standards.

In the CCP, we have completed approximately 933 codings from texts that are either translated to, or originally composed in English and 25 codings from texts in languages other than English. In the case of the 25 constitutions, we have actually completed one coding from the non-English text and the second coding and reconciliation from a text translated to English. Given this variation in translated and original sources, one can make a number of comparisons, all based on the basic expectation that translation increases error. In the analysis, we employ a simple indicator variable that identifies cases as "translated" or "not-translated" based on the whether or not, for each constitution in question, English is one of the official languages of the country at the time of drafting. Note that this variable includes as "translated" even those cases in which the constitutions were coded in a language other than English since the reconciliation was done with the use of a translated English text. (Remember that our unit of analysis is the coder-reconciler dyad). The variable therefore lumps together two different kinds of coder-reconciler dyads (those for which both coder and reconciler use the translated text and those for which only the reconciler uses the translated text). This is a distinction that we will draw in subsequent iterations of the paper.

Apart from language, we expect that differences in institutional and societal culture (that is between that of coders and that of their target constitution) will also lead to error. For U.S. coders, it is likely that constitutions from some regions – particularly Asia and Africa – will be more difficult to

interpret than will those from Europe and the Americas. We also harbor some suspicions about whether certain kinds of institutional arrangements will be more difficult to assess across. Most of our coders, for example, are more familiar with executive-legislative relationships that resemble those common in presidential systems than they are those common in parliamentary systems. Similarly, constitutions from contexts steeped in the common-law legal tradition may be more easily interpreted by our coders than will those from civil-law contexts. The implications of any of these cross-cultural, or cross-institutional effects are potentially far-reaching. It is well known that constitutional ideas migrate quite freely across states, either voluntarily or involuntarily (in the case imposed constitutions). In terms of the viability of imported ideas, one wonders whether their interpretation will be muddied in the process.

Syntax and Structure. Finally, we consider the distinct possibility that the compositional structure of the writing – regardless of its provenance – will be associated with variation in error. Some constitutions will simply be written more clearly than others. Whether deserved or not, the U.S. Constitution is often praised for its plain, accessible style. One does not have to look hard to find constitutions at the other end of the spectrum. Consider this double negative in the Zimbabwean Constitution (Art. 16.7):

> Nothing contained in or done under the authority of any law shall be held to be in contravention of subsection (1) to the extent that the law in question makes provision for the acquisition of any property or any interest or right therein in any of the following cases...

It is critical, then, to control in our analysis for some measure of linguistic complexity. Fortunately, our texts are in digital form and there exist a number of machine-mediated ways to assess complexity. For example, one set of methods, the Flesch and Kincaid indices, compute readability as a function of sentence length and word length. We are curious about these complexity measures and they deserve further elaboration, but for our purposes here we depend on them mostly as control variables.

Accordingly, we include five measures in the analysis – some of them highly inter-correlated – but put aside a discussion of their measurement qualities.[5]

We also include a measure of the length of the constitution, under the expectation that longer constitutions will be more fatiguing for the coder and, quite likely, less clearly written, and a measure of scope. We also include a measure of constitutional scope, which is roughly the density of constitutional provisions in the constitution (see Elkins, Ginsburg, and Melton 2009). Our expectation with respect to scope is that constitutions with broader content are likely to include provisions on more obscure (and confusing) rules of the game (e.g. the presence of amparo or ombudsman, the protection from *nulla poena sine lege*, the right of self determination, etc.). Although constitutional lawyers understand these concepts, they are probably less accessible to the average citizen.

Characteristics Associated with the Coder and Reconciler

Not all error in judgment can be blamed on the constitutional text or its context. Certainly, coders will vary in their abilities, experience, and interest in interpreting constitutional text. This is the problem of PICNIC, to borrow an acronym used by computer programmers: Problem In Chair, Not In Computer [Constitution].

One source of variation has to do with how *experienced* the coder is with reading constitutions and, more generally, with constitutional law. We have a direct measure of the former, since we know for any given coding, how many previous codings a coder had completed. Also, since our sample of coders draws from a set of political science graduate students, Law students, and undergraduates -- all at different points in their training and at three different academic institutions -- we are able to assess any differences associated with these at least small degrees of variation in academic experience.

[5] The measures are: (1) The IMG index; (2) percent of complex words; (3) percent of one-time words; (4) Flesch index; and (5) Kincaid index.

Apart from experience, some coders will be more *conscientious* and, perhaps, possess sharper interpretive abilities than others. Admittedly, we do not expect large difference with respect to these qualities since we selected coders based largely on these very characteristics. Nevertheless, our coders undoubtedly vary to some degree on these dimensions. We have several measures of conscientiousness. One is the number of questions, on average, that a coder posts to the message board under the theory that those who ask questions and bring cases are more engaged in the project and will exhibit lower error rates. A second is the elapsed time between the start and finish of a coding (which includes time "on" and "off" the clock), under the theory that those coders that work more steadily will be more reliable than those that interpret a document over a longer stretch of time. This expectation stems from our expectation that coding will be more reliable if a coder works in uninterrupted blocks of time rather than spreading an analysis of a constitution over a longer period. Of course, if coders vary in their rate by which they answer questions – something that should affect elapsed time and their error rate-- then the elapsed time measure will pick up two contradictory effects. (Note, of course, that we control for difficulty and length of a constitution).

To the degree that coders and reconcilers vary based on their analytical and reading skills (or any inherent knack for deciphering constitutions), we estimate a fixed-effects model (i.e., add dummy variables for each coder) in order to capture differences in the reliability of any given coder, or reconciler, over and above the other covariates in the model. The coefficients on these individual dummy variables, of course, are of procedural interest to us since they serve as a measure (although not always a perfect one) of the reliability of individual coders.

Thus far in this section we have described a set of monadic hypotheses about coders and reconcilers, but of course any measure of intercoder reliability -- like tango – is the product of at least two individuals. As such, we take into account several dyadic attributes. One such dyadic hypothesis has to do with the relative stage of the project at which the judgments were made. As we note above,

our instructions to coders regarding ambiguous interpretive cases evolve incrementally, in part, since we issue rulings on a regular basis that set precedent for future interpretation. Due to changes in doctrine, then, one would therefore expect differences between judgments made at the beginning of the coding process and those made at a later stage. We assume a relatively constant revision of doctrine and include a measure of the span of time between the completion date of the coding and the start date of the reconciliation. A more precise measure might be based on the relative accumulation of caselaw (based on the density of postings to the message board), which does not grow at a constant rate, but we do not explore that possibility here. Since the message board system came on line only a year or so into the process, we include a variable that indicates the onset of this procedure since caselaw was disseminated in a less direct way in the earlier period. In some sense, this set of variables capture mere procedural elements of the process. In another sense, they serve as intriguing indicators of the effect of established doctrine in the interpretation of constitutional texts – certainly an important substantive question.

Another set of dyadic measures can be built from the combination of the attributes of the coders, something we intend to incorporate in subsequent iterations of this paper. Specifically, one wonders whether a shared culture, training, and courses (notably, those taught by the principal investigators) in either law school or a political science graduate program will manifest itself in correlated measurement error within each group. The hypothesis here is that the codings of coders with the same training will exhibit higher rates of agreement. Note that this is unrelated to any effect on error associated with a coder's program of study, the monadic effect that we describe above. The dyadic hypothesis is that two coders from the same academic program will agree more; the monadic question is whether coders from different programs will exhibit different amounts of error.

Characteristics Associated with the Process

We also include a set of covariates in the model that are mostly of procedural interest to us, or help us control for confounds based on our measure of the dependent variable. These variables include: (1) a binary variable indicating the shift, part-way through the administration of the project, to a new survey engine; (2) a measure of the number of "non-applicable" responses per constitution, since questions will be likely produce higher agreement between coder and reconciler and may not be fully captured by our variable measuring the scope of the constitution; (3) the number codings completed for a particular constitution, since it is likely that more codings will lead reconcilers to scrutinize coders' decisions more extensively and will result in lower reliabilities.

V. EMPIRICAL ANALYSIS

Econometric Issues

The structure of our data introduces several peculiarities in the analysis. While our unit of analysis is the coder-reconciler dyad, our variables are measured at three different levels: that of the (1) constitution; (2) coder; and (3) coder-reconciler dyad. Since information from constitutions and coders will appear multiple times (but assumed to be independent of one another), we adjust the standard errors by clustering them at the level of the constitution, the level of most interest to us.

Baseline measures of Reliability

Figure 2 shows the distribution of our measure of reliability. On average, coder-reconciler agreement across the set of items in question is 82.6 percent, with a standard deviation of 5.6. This rate of error (low by our standards) provides a sense of the difficulty in interpreting constitutions. From a procedural perspective, it also validates our decision to adopt a process of multiple codings followed by reconciliation, rather than single interpretations. The distribution also suggests significant variation in agreement across coder-reconciler dyads and, probably, across constitutions.

If we calculate the mean coder-reconciler agreement per constitution, we are in a position to identify – at least in a bivariate manner – the more troublesome and least troublesome texts. Those constitutions eliciting the highest level of agreement were Haiti (1811), Thailand (1959), and Bhutan (1981), all with reliability scores above 90%. Those with lowest level of agreement were France (1958), Armenia (1995), and Guyana (1995) all with reliability scores below 72%. Remember, of course, that some of the error could be attributed to coder-specific or procedure-specific factors, something we will account for in the regression models below. Still, it is interesting at this point to inspect scores of interesting cases. The U.S. constitution, perhaps surprisingly, is only as high 88%, one of the highest to be sure, but still with more the 10% of error. The Brazilian constitution, whose story opened this essay, comes in at 81%, slightly below the sample mean. Again, after accounting for other factors below, we will be able to compute scores for these constitutions that are functions of the text and setting themselves, not the coders.

Explaining Variation in Interpretability: Implications for the Rule of Law

We described a rather inclusive set of hypotheses above, many of which probably deserve elaboration. Table 2 reports regression results from four model specifications: (1) one with only the fixed effects (dummy variables for each coder); (2) one with fixed effects plus coder-specific characteristics; (3) one with fixed effects plus constitutional characteristics; and (4) one that is fully specified. Here we focus on the effects of variables that implicate some of the challenges to sustaining the rule of law, as conceptualized above.

One question is whether constitutions are accessible to all citizens, whether legal specialists or not. That is, to return to the case of the Brazilian constitution of 1988, should we be disturbed that the "accessibility" suggestions of the linguistic consultant were ignored by Brazilian elites? As we describe above, we do not know how average citizens would respond to our questions posed by our survey instrument, however, we can say something about this idea based on the relative experience and

conscientiousness of our own coders. For example, it is clear that the reliability of coders improves with each constitution that they code (an increase of .08 per coding, based on the estimates in model 4. After 50 codings – the level reached by our veteran coders – the reliability scores increase by 4 points, an increase of almost a standard deviation of the dependent variable. For the most part, however, we did not see large differences in reliability across the characteristics of coders. Texas partisans may read something into the 3.5 point superiority in the reliability of the University of Texas coders over those from the University of Illinois and 5 point margin over those from the University of Chicago as a proof of something, but obviously the effect does not tell us much about the difference in elites and masses in interpreting constitutions. Similarly, the finding that the reliability of law students is about two points higher than that of graduate students in political science and undergraduates is mostly of procedural interest. Experience and background clearly have *some* effect on reliability, but if we want to make the sort of elite-mass claims that connect to rule of law, we may need to adopt a design that asks the man and woman on the street interpret constitutions (something decidedly not in our plans).

Consider now the issue of time and constitutional interpretation. The results from the analysis suggest that our coders are as good at interpreting older constitutions as they are contemporary constitutions. Reliability appears to decrease by anywhere from 0.01 to 0.004, depending upon the model, with each year of age of the constitution -- a relatively small effect that is not statistically different from zero in the full model specification. This suggests that the problem of law written previous generations is not necessarily problematic from the perspective of clarity. Another finding related to time has to do with the effect of the accumulation of doctrine, as measured by the accumulation of our instructions to coders regarding interpretability problems. Here we find that the effect of the time elapsed between the coding and reconciliation of a constitution has no effect on the degree of error. Apparently, our adjudication of ambiguous cases – and the doctrinal caselaw thus created --did not shift the meaning of constitutions to any great extent. One need not read too much

into this effect, but it does seem to suggest that many of the answers to constitutional questions – at least in the context of our study – appear to reside in the text itself.

Now we turn to issues of cross-cultural interpretation. Are the judgments of our U.S. based coders distorted appreciably by constitutions translated from other languages, written in non-Western countries, or written in institutionally exotic lands? Yes (but minimally), no, and no. The reliability of interpretations of constitutions from countries with English as one of their official languages was, on average, 1.3 points higher than that of the constitutions that were read in translation. This is small effect, some of which may be due to our measurement strategy. As we note above, we intend to pursue this effect further. As for the other contextual variables, we see no noticeable effects. None of the regional dummy variables (with Western Europe as the residual category) exhibit coefficients different from zero. While we do not include a measure of presidentialism versus parliamentarism, since nearly all constitutions in Latin American have been presidential, and the Latin American dummy variable in insignificant, it seems unlikely that coders had any particular problem with coding one system or another. The same is true with respect to common law versus civil law traditions – neither of which shows up significant in the assessment of reliability.

It may be, interestingly enough, that one of the primary sources of variation in reliability has to do with how (stylistically) constitutions are written, regardless of what era, language, or region they are written. One of the largest effects that we find is that of scope. As against constitutions with the minimal amount of breadth, constitutions that include the largest number of provisions exhibit reliability scores more than nine points lower – an effect the size of almost two standard deviations in the dependent variable. It also appears possible that some of linguistic complexity measures may matter (e.g., the Flesch readability index shows a non-significant effect of a full 22 points). These measures are probably the least developed ingredient in our analysis and certainly deserve more careful attention in subsequent versions of the paper.

VI. CONCLUSION

Legal clarity is a central element of the rule of law, under virtually every definition. Unclear law is undermines predictability and compliance, leads to inconsistent application, and will fail to provide effective limits on government. While there is a good deal of agreement on this point as a matter of principle, we have to date no systematic measure that can be used to compare levels of clarity across legal texts.

Using our data from the Comparative Constitutions Project, we develop a measure of clarity and interpretability of constitutional texts. A constitution that is unclear will be difficult for our coders to understand, and presumably this would extend to members of the informed subject population, though we do not make any strong assertions of external validity at this point. Instead, our primary goal here is to contribute to ongoing methodological debates about conceptualization and measurement of the rule of law (e.g., Schrank and Kurtz 2006), focusing on one central dimension of the concept.

It seems clear that constitutions vary significantly in the degree to which readers understand their provisions. It could be that their clarity, as estimated by the kind of analysis that we present above, constitutes an aspect of rule of law that analysts would consider among other elements. In future iterations of this project, we plan to generate such scores and examine them in the context of relevant outcomes and other factors. As a start, one would be curious whether measures of clarity have any bearing on the conditions under which a constitution was written, whether it affected the endurance of these texts (Elkins, Ginsburg and Melton 2009), or whether clarity has any association with some of de facto measures of rule of law that are in use. Our larger, more general, point, is that clarity of law is a central element of the rule of law that can be profitably incorporated into more empirical and theoretical work on the topic.

Figure 1. Universe and Sample, the Comparative Constitutions Project

Table 1. Illustration of a the Reliability Calculation

	Question 1		Question 2		
	D_1	w_1	D_2	w_2	Rel_k
Coding 1	1	0	1	0.50	100
Coding 2	1	0	0	0.50	0

Figure 2. Distribution of the Reliability Measure

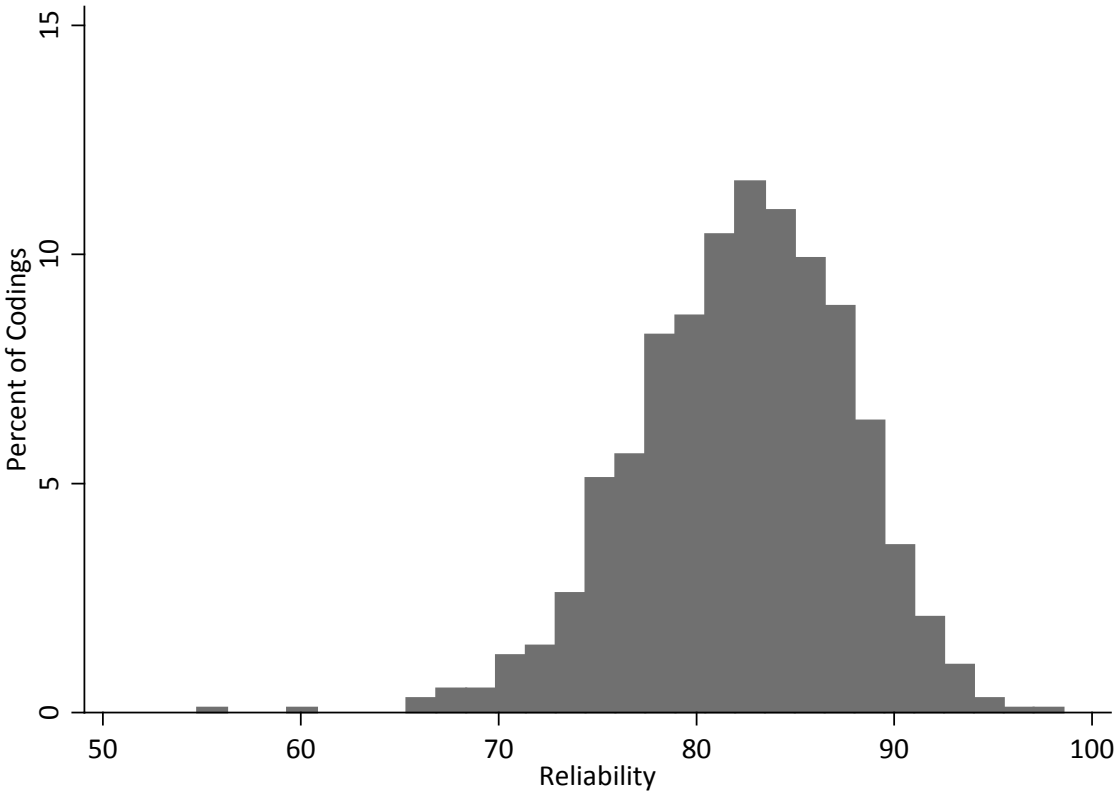

Table 2. Statistical Models of Reliability
OLS regression with fixed effects (for coders and reconcilers)

Variables	Model 1	Model 2	Model 3	Model 4
Word recognition (IMG index)			-3.778*	-1.449
			(2.122)	(1.949)
Complex Words Index			6.203	5.331
			(4.728)	(4.117)
Flesch Index			30.41	22.56
			(21.83)	(19.18)
Kincaid Index			30.41	23.16
			(21.57)	(18.87)
One Time Words			3.672*	0.741
			(1.978)	(1.989)
Length (in 1,000's of Words)			0.000355	0.0102
			(0.0286)	(0.0262)
Scope			-18.66***	-9.426***
			(1.979)	(2.212)
English Language text			0.373	-0.497
			(1.052)	(0.982)
English is Official Language			1.577***	1.343***
			(0.556)	(0.487)
Common Law			-1.054	-1.024
			(1.264)	(1.172)
Latin America			0.734	0.298
			(0.764)	(0.732)
Eastern Europe			-0.286	-0.0847
			(0.720)	(0.623)
Sub-Saharan Africa			0.318	-0.583
			(0.955)	(0.793)
North Africa/Middle East			0.145	-0.172
			(0.914)	(0.797)
South Asia			1.373	-0.0947
			(1.199)	(1.073)
East Asia			0.639	0.850
			(0.834)	(0.715)
Oceania			0.478	-0.408
			(1.381)	(1.111)
Age of Constitution at Coding			-0.00948	0.00423
			(0.00796)	(0.00754)
Codings Completed by Coder to date		0.0788***		0.0761***
		(0.0146)		(0.0158)
Total Messages posted by coder		0.0561***		0.0186**
		(0.0111)		(0.00885)
Undergraduate Student		1.588		4.123**
		(3.420)		(2.088)
Law Student		0.821		6.549***
		(3.245)		(2.054)
Graduate Student		2.899		4.362*
		(3.366)		(2.561)
University of Chicago		3.274**		-1.321
		(1.547)		(1.519)

Variables	Model 1	Model 2	Model 3	Model 4
University of Texas		3.699**		3.134*
		(1.569)		(1.900)
Reconciliations Completed to date by reconciler		-0.00236		-0.00394
		(0.00502)		(0.00518)
Total Messages posted by reconciler		-0.00134**		-0.000774
		(0.000594)		(0.000581)
Message Board in force		1.073*		1.076*
		(0.551)		(0.580)
Days between Coding and Reconciliation		-0.00104		-0.00185
		(0.00116)		(0.00134)
Number of Non Applicable Responses		0.131***		0.0793***
		(0.00854)		(0.0120)
Number of Codings Completed per Constitution		-0.411**		-0.552***
		(0.198)		(0.213)
CITES (first) survey engine in use		-0.629		-1.556
		(1.271)		(1.315)
Constant	82.56***	73.47***	88.25***	69.38***
	(1.319)	(3.743)	(27.88)	(25.29)
Adjusted R^2	0.471	0.673	0.648	0.703
Observations	959	959	887	887

Notes: Cells contain coefficient estimates with robust standard errors in parentheses; statistical significant is indicated as follows: *** p<0.01, ** p<0.05, * p<0.1; coefficients of coder and reconciler fixed-effects are omitted.

VII. REFERENCES

de Figueiredo, Rui and Barry Weingast. 2005. Self-enforcing Federalism. *Journal of Law, Economics, and Organization* 21: 103–35.

Dorf, Michael. 2008?. An Institutional Approach to Legal Indeterminacy.

Elkins, Ginsburg, Melton. 2009. *The Endurance of National Constitutions.* New York: Cambridge University Press.

Fuller, Lon. 1964. *The Morality of Law.* New Haven: Yale University Press.

Hardin, Russell. 1989. Why a Constitution? In *The Federalist Papers and the New Institutionalism*, edited by Bernard Grofman and Donald Wittman. New York:Agathon Press.

Harris II, William F. 1993. *The Interpretable Constitution*. Baltimore: Johns Hopkins University Press.

Hayek, F. A. 1944. *The Road to Serfdom*

Leiter, Brian. 2007. Naturalizing jurisprudence: essays on american legal realism and naturalism in legal philosophy.

Llewellyn, Karl. On Reading and Using the Newer.

OECD Development Assistance Committee, Issues Brief: Equal Access to Justice and the Rule of Law (2005), http://www.oecd.org/ dataoecd/26/51/35785471.pdf.

Ordeshook, Peter C. 1992. Constitutional Stability. *Constitutional Political Economy*

3(2): 137–75.

Rubenfeld, Jed. 2001. Freedom and Time: A Theory of Constitutional Self-Government. New Haven: Yale University Press.

Schelling, Thomas. 1980. *The Strategy of Conflict*.Cambridge, MA: Harvard University Press.

Tamanaha, Brian. 2004. *The Rule of Law.* New York Oxford University Press.

Thompson, E.P. ___. *Whigs and Hunters*

Weingast, Barry. 2006. 2006.Designing Constitutional Stability. In *Democratic Constitutional Design and Public Policy*, edited by Roger Congleton and Birgitta Swedborg. Cambridge: MIT Press.