Empirical Comparative Law

Empirical Comparative Law

Holger Spamann[*]

Abstract: I review the empirical comparative law literature with an emphasis on quantitative work. After situating the field and surveying its main applications to date, I turn to methodological issues. I discuss at length the obstacles to causal inference from comparative data, and caution against inappropriate use of instrumental variables and other techniques. Even if comparative data cannot identify any single causal theory, however, they are extremely important in narrowing down the set of plausible theories. I report progress in measurement design, and suggest improvements in data analysis and interpretation using techniques from other fields, particularly growth econometrics.

JEL codes C18, K00, P50

(forthcoming, 11 Annual Review of Law and Social Science (2015))

[*] Assistant professor, Harvard Law School. hspamann@law.harvard.edu. For very helpful comments, I am indebted to Jonathan Beauchamp, Martin Gelter, Tom Ginsburg, Yehonatan Givati, Dan Ho, Katerina Linos, Mila Versteeg, members of the Harvard Law School Empirical Legal Studies Group, and especially Ralf Michaels. Valerio Romano and Harin Song compiled very helpful lists of relevant articles. I gratefully acknowledge financial support from Harvard Law School's Summer Research Program.

1. Introduction

In this article, I review the literature that uses cross-country legal data to test causal theories in an explicit hypothesis-testing framework. This literature, which I call empirical comparative law (ECL), has grown tremendously in the past two decades. The appeal of ECL is that cross-country variation is large. In fact, cross-country variation is often the only variation available, as many characteristics, especially legal characteristics, are fixed within countries. It is difficult, however, to harness this variation for convincing tests of causal effects or even to establish robust associations. Randomized experiments are unavailable, "natural experiments" are rare at best, and even standard observational studies face considerable challenges: units (countries) are highly heterogeneous, samples are small (there is only one sample[1] of at most 200 countries), and data are sparse (i.e., unavailable for many countries or variables). Moreover, replication on independent samples is generally impossible.[2] As a consequence, comparative data require particularly careful analysis and interpretation, and even then can rarely if ever isolate any particular causal effect. Nevertheless, they can considerably reduce the number of plausible effects---or so I here argue.

The vast majority of published ECL is quantitative. This does not imply that qualitative comparative evidence is not important---far from it. Comparative evidence is most powerful when a single country provides a counterexample to what could otherwise seem a necessary relationship. Partially because of this, however, most nontrivial yet credible hypotheses are probabilistic and hence lend themselves to quantitative tests (Spamann 2009). I make only occasional mention of qualitative methods in this survey.

Section 2 situates ECL at the crossroads of empirical legal studies, comparative law, and sister empirical disciplines such as comparative politics. This section is not necessary for understanding the rest of this article, but it may be helpful for readers wondering what, if anything, is new in ECL. Section 3 surveys the main applications of ECL to date. Because ECL is a method rather than a body of knowledge, this survey's goal is not to be exhaustive but merely to illustrate the method's use. Sections 4 and 5 form the substantive core of this article. Section 4 reviews the obstacles to causal inference from comparative data. I take a very skeptical view but argue that ECL remains at least an important filter for causal theories and complements other empirical tests. Section 5 discusses other methodological issues of particular relevance for ECL in data collection (measurement), analysis, and interpretation. Section 6 concludes this review.

[1] Throughout the article, I refer to the collection of countries on earth as a sample rather than the population. In other words, I take the relevant population to be an abstractly defined set of possible countries rather than the set of presently or formerly existing countries on earth. The reason is that ECL aims or should aim to test theories that can predict what will happen in a country that was not yet in the sample, for example, because it is new or because it implements a reform.

[2] A well executed study will use all the data points (countries) to which the hypothesis under investigation applies. This will leave no other countries for replication. (It is nevertheless appropriate to speak of "sample," see previous note.)

2. Relationship To Other Literatures

ECL is closely related to, and partially overlaps with, at least three literatures: empirical legal studies, comparative politics, and comparative law. The dividing lines are not sharp and may have as much to do with the people involved as with the questions and the methods.[3]

Substantively, ECL is a subfield of empirical legal studies. The distinguishing feature of ECL is the use of cross-country data. This is, however, not a fundamental distinction. For example, the comparison of national constitutions is not fundamentally different from the comparison of state constitutions (e.g., Dixon & Holden 2012), and both can shed light on the same questions. That being said, cross-country data often offer more variation then subnational data. Often, a particular question is decided at the national level, such that the only available variation is cross country. The greater variability has both advantages and disadvantages. For example, an investigation of dictatorship effects would benefit from comparative data that contains dictatorships, whereas an investigation of details of democratic design against the backdrop of US-style mass media may be better served by a comparison of states. As a practical matter, cross-country data tend to be much more difficult to obtain, at least in a consistent format.

Comparative politics and other comparative social sciences do use cross-country data, and many of the papers surveyed below explicitly speak and may even belong to those disciplines. I have attempted to select papers that produce and/or use more and better comparative legal data than has traditionally been the case. In particular, the involvement of lawyers in ECL has considerably improved and broadened the legal data-collection process.

ECL also has many points of contact with various branches of the heterodox field of comparative law.[4] The major difference is that comparative lawyers traditionally ask different, noncausal questions (Pistor 2010).[5] In particular, most comparative law has been devoted to understanding foreign legal systems (e.g., Lasser 2004), developing common concepts (e.g., Michaels 2006), and a comparative mapping of legal rules and institutions (e.g., Zweigert & Kötz 1998). Where comparative lawyers have tackled causal questions, as in the exploration of legal change (e.g., Glendon 1989), they have favored more exploratory, descriptive accounts over an explicit hypothesis-testing framework.[6] In spite or rather because of these differences between ECL and comparative law, the opportunities for fruitful exchange are plentiful. Comparative law offers ECL important hypotheses, concepts, and knowledge of legal rules

[3] Suchman & Mertz (2010) document a similar phenomenon in domestic empirical work. In particular, they distinguish "Empirical Legal Studies" from other empirical approaches. For brevity, I use the term "empirical legal studies" for any empirical work with legal data.

[4] On the heterodoxy of the field, see, e.g., Reimann (2002, pt. II), and compare the number of approaches surveyed in Reimann & Zimmermann (2006, pt. II).

[5] The dividing line has never been sharp. For example, Edouard Lambert (1905) speculated in his famous report for the Paris Congress of 1900 that comparative law and legal sociology are one and the same thing (p. 35), with missions such as to reveal the "natural laws" of social life (p. 32) or at least the effect of legal reforms in various countries (p. 36).

[6] Some within the field even hesitate to attach the label "empirical" to comparative law, although comparative law is by definition empirical in the sense of learning from observation. Cf. Jansen (2006, p. 313) (contrasting comparative law's attempts to capture similarity and dissimilarity of legal systems to "the empirical sciences").

and institutions and their functioning in practice. ECL in turn generates data that can inform taxonomies of legal systems and other descriptive elements of comparative law.[7] In separate work, I also argue that comparative law would benefit from applying ECL's hypothesis testing framework and empirical rigor to traditional comparative law questions such as legal families or traditions (Spamann 2013).

Comparative lawyers have often been quite critical of ECL, in particular the law and finance literature surveyed in Section 3.2 (for a summary of the criticisms, see Michaels 2009). I do not have space to address these criticisms explicitly, but my own criticisms from within the statistical-empirical framework nest most of them. For example, I interpret the charges of neglect of functionally equivalent mechanisms, arbitrary selection and weighting of index components, and neglect of law in action as problems of measurement validity (Section 5.1) or low prior probability of the tested theory (Section 5.3). Similarly, the criticism that empirical work has paid too little attention to local particularities alleges a problem of measurement or a problem of insufficient controls that I discuss at length in Sections 4.1 and 5.2.

3. Examples

In this section, I review some particularly active areas of ECL. I make no attempt to be comprehensive. This would be pointless for a "field" identified by a method rather than an object of study, and it would be hopeless given the volume of literature.[8] Rather, my goal is to illustrate the broad range of applications of ECL. For these merely illustrative purposes, I refrain from reporting statistical and economic magnitudes. I defer methodological discussions of measurement and interpretation to subsequent sections.

3.1. Constitutional Law

There is a rich comparative empirical literature on constitutions in comparative politics and political economy (e.g., Ben-Bassat & Dahan 2008; for a survey, see, e.g., Landmann & Robinson 2009). In the past decade, however, this research has received a major boost through the Comparative Constitutions Project (CCP) directed by political scientists Zachary Elkins and James Melton and lawyer Tom Ginsburg (comparativeconstitutionsproject.org). The CCP is a publicly available cross-national historical data set of all written constitutions from 1789 through the present. It codes 668 constitutional characteristics and tracks their development over time, including adoption, amendments, and suspensions.

[7] For example, Siems (2015) performs a cluster analysis of the Doing Business investor protection data and finds no confirmation for standard taxonomies. Such work still remains too rare, even though the situation has improved since Reimann (2002, p. 686) wrote "[C]omparative law has still not acquired a solid empirical basis. We have ridiculously little statistical data about the legal systems we study and compare. Without such data, most of our conclusions rest on personal intuition, anecdotal information, or plain speculation, rather than on systematic observation of hard facts." See Siems (2014) for a textbook-length attempt to integrate "numerical" findings into comparative law.

[8] For example, Carrubba et al. (2012) count 154 articles just in comparative judicial politics from 1990--2009. Isolated examples of empirical comparative law can be found as far back as the 1960s (e.g., Schwartz & Miller 1964), and almost certainly before.

CCP data have already been used extensively in both causal and descriptive studies. Much research focuses on constitution-making. Ginsburg et al. (2009a) found that constitution-making processes (such as the involvement of a constitutional assembly or the requirement of ratification by referendum) correlate with constitutional features such as the number of constitutionally guaranteed rights. Ginsburg et al. (2009b) documented that constitutions last on average only 19 years, but that those that are flexible last longer (cf. Dixon & Ginsburg 2011). Ginsburg (2010) demonstrated that constitutions have been growing in length, scope, and detail. Elkins et al. (2008) showed that constitutions adopted during occupation display surprisingly little resemblance to the occupying power's constitution, yet they rarely survive for long. Other research examines subsequent outcomes. Elkins & Sides (2007) documented that ethnic divisions tend to be higher in countries with federalism and proportional electoral systems. Ginsburg & Garoupa (2009) demonstrated that there is little relationship between judicial council design and judicial quality. Ginsburg et al. (2011) observed that executive term limits are overwhelmingly observed in established democracies. CCP data allow quantification of important phenomena such as the correlation of human rights on the books and in action (Law & Versteeg 2013), and they shed doubt on what used to be considered foundational distinctions, such as those between parliamentary and presidential democracies (Cheibub et al. 2013). They have also inspired a book on comparative constitutional design (Ginsburg 2012).

Other data remain in use. For example, Shulztiner & Carmi (2014) documented the rise of "dignity" (which the CCP also records) with data collected directly from the constitutions of all 193 UN member states. They showed both qualitatively and quantitatively that "dignity" is not necessarily associated with liberal practices and may even be invoked for illiberal purposes. Dreher et al. (2010) used data combining human rights on the books and in action from the CIRI Human Rights Data Project (humanrightsdata.com) to show that terrorism tends to be followed by a significant decrease in the respect for human rights. Using data derived from the US State Department and Amnesty International reports, Goderis & Versteeg (2012) showed that such decreases are smaller in countries with independent judicial review (as measured by La Porta et al. 2004).

3.2. Law and Finance, Doing Business, and Legal Origins

Perhaps the largest literature in ECL to date is known as law and finance. In the eponymous paper, financial economists La Porta, Lopez-de-Silanes, Shleifer, and Vishny (La Porta et al. 1998) introduced additive indicators of certain shareholder and creditor protection rules in 49 countries, respectively known as the antidirector rights index and the creditor rights index. They showed that the antidirector rights index positively correlated with equity market outcomes such as market capitalization and ownership dispersion (La Porta et al. 1997; 1998; 1999; 2000; 2002a,b). Some of these early results subsequently yielded to better data (Holderness forthcoming, Spamann 2010c). But newer, more refined studies on even larger samples have upheld, refined, and added other results with new indices of public and private securities laws (La Porta et al. 2006), rules against managerial self-dealing (Djankov et al. 2008b), duration of and recovery in bankruptcy (Djankov et al. 2008a), and a revised creditor rights index (Djankov et al. 2007). A voluminous follow-up literature has documented correlations of these measures with various financial market outcomes, developed supporting theory, and tested corollary

hypotheses on comparative and domestic data. In a recent survey, La Porta et al. (2013, p. 450) concluded that this literature had established that "better [legal] investor protection…is associated with improved financial development, better access to finance, and higher ownership dispersion." Several empirical studies, however, dispute this claim (e.g., Armour et al. 2009b, Cheffins et al. 2013, Holderness forthcoming) as well as the regularity assumptions embedded in the quantitative methods used to support it (e.g., Milhaupt & Pistor 2008, Pistor 2013).

The main authors of law and finance soon exported their approach to other areas of law. They showed that, as measured by the legal indices they specifically designed and collected for these studies, procedural formalism "is associated with higher expected duration of judicial proceedings, less consistency, less honesty, less fairness in judicial decisions, and more corruption" (Djankov et al. 2003, p. 453); "judicial independence and constitutional review are associated with greater freedom" (La Porta et al. 2004, p. 445); "[c]ountries with heavier regulation of entry have higher corruption and larger unofficial economies, but not better quality of public or private goods" (Djankov et al. 2002, p. 1); "[h]eavier regulation of labor is associated with lower labor force participation and higher unemployment, especially of the young" (Botero et al. 2004, p. 1339); and "[p]ublic disclosure [of politicians' financial and other conflicts], but not internal disclosure to parliament, is positively related to government quality, including lower corruption" (Djankov et al. 2010, p. 179). A general and controversial theme of this literature is that, as measured by these studies, "interventionist" policies, such as government ownership of banks (Barth et al. 2006,[9] La Porta et al. 2002a), correlate with worse outcomes such as corruption.

The World Bank financed these and the later law and finance studies and used them as foundation for its Doing Business project (doingbusiness.org) (World Bank 2014). Since 2004, Doing Business has been collecting annual legal data in 10 or 11 areas of business law and regulation: starting a business, dealing with construction permits, getting electricity, registering property, paying taxes, trading across borders, getting credit, protecting minority investors, enforcing contracts, resolving insolvency, and labor market regulation. The standard template is to collect data on the cost, duration, and number of procedures involved in a paradigmatic case; some indicators instead code the presence of particular legal provisions. A dedicated team at the World Bank draws on a large network of respondent lawyers and other experts around the world to collect, verify, and improve its data. Over time, the World Bank has gradually refined its methodology and weeded out mistakes. As a result, some key results had to be corrected (e.g., Spamann 2010b). The method and the perceived neoliberal bias of Doing Business have been extremely controversial (cf. Manuel et al. 2013). Used with care, however, Doing Business and related indicators such as the Rule of Law Index from worldjusticeproject.org are extremely valuable pieces of information for empirical comparative research (cf. Davis 2014) (Section 5.1, below).

Law and finance and its progeny were successful in economics and finance and controversial in law in large part because of their solution to the endogeneity problem: Does the law on the books cause the observed social phenomenon (e.g., financial market size), or is it the converse (see generally Section 4)?

[9] Barth et al. (2013) present an expanded and updated version of the underlying data on bank regulation and supervision in 180 countries from 1999 to 2011.

La Porta et al. (1998) observed that their indices of investor protection were on average significantly higher in common law than in civil law systems. Arguing that membership in a legal family ("legal origin") was determined long before contemporary market outcomes and plausibly influences investor protection laws, La Porta et al. (1998) concluded that causation must run from the laws to the market outcomes. All the other studies cited above found the same correlation between common law and more "market-friendly" regulation, even including a lower likelihood of using the draft (Mulligan & Shleifer 2005). This raised the question why the common law countries were more market friendly on average. Drawing on Merryman (1969) and other legal comparatists, Glaeser & Shleifer (2002) and others conjectured that the answer was to be found in the civil law's ostensible preference for statutes and less independent judges.

This conjecture drew a number of responses. First, the correlation between legal origins, laws on the books, and outcomes may be confounded by other factors. Even though the country-level data ruled out many factors (e.g., religion) (La Porta et al. 2008), it could not address others. In particular, Klerman et al. (2011) pointed out that legal origin is almost perfectly correlated with colonial origin, which could have influenced subsequent developments through various other channels.[10] Bubb (2013) validated such concerns by zooming in to the local level. He showed that de facto property rights differ little between either side of the border separating Ghana and Côte d'Ivoire, whereas other economic outcomes do. Michalopoulos & Papaioannou (2014) investigated similar phenomena across Africa. Second, even if laws at the country level were responsible for the diverging outcomes, intrinsic differences between common and civil law would not be the only plausible explanation. In particular, Spamann (2010a) showed that peripheral countries continue to copy legal materials from their origin countries, such that random variation in the origin countries England and France unrelated to the common/civil law distinction could generate the observed pattern. Third, and crucially, there was little direct evidence for the claim that differences in the role of case law (or for that matter any other phenomena traditionally linked to legal origins) explained the cross-country pattern of regulation. For example, Roe (2006) pointed out that the legal rules in question were overwhelmingly statutory in all jurisdictions, not just the civil law countries, whereas Jackson & Roe (2009) showed that common law countries spent more money on public enforcement of securities laws. Similarly, in civil procedure---the one area where most modern comparative lawyers still see pronounced differences between common and civil law (Michaels 2009, p. 781)---corrected data showed no performance differences between the two legal families (Spamann 2010b). Finally, Klerman & Mahoney (2007) and Roe (2007) undermined the historical narrative in Glaeser & Shleifer (2002).

In response, La Porta et al. (2008, pp. 286, 308; 2013, pp. 427, 457) adopted a generic characterization of legal families as "style[s] of social control of economic life (and maybe of other aspects of life as well)," where "common law stands for the strategy of social control that seeks to support private market outcomes, whereas civil law seeks to replace such outcomes with state-desired allocations." More specifically, they conjectured that common and civil law differ in their "toolkits" or the "beliefs about

[10] Oto-Peralías & Romero-Avila (2014) argue that legal and colonial origin interacted, in particular because England pursued a strategy of "indirect rule" in thickly settled or otherwise hostile territories.

how the law should deal with social problems…incorporated in legal rules, institutions, and education." This conjecture awaits refinement into hypotheses that are both falsifiable and not obviously false. Which tools are supposed to be lacking in either family, and which aspects of the law are supposed to incorporate these beliefs in such durable fashion, and how? In particular, even proponents of legal origin differences affirm that civil law systems do not lack the "tool" of judicial precedent, i.e., case law (e.g., Merryman 1969, pp. 48--49).

3.3. Diffusion and Legal Transplants

Another active area of research has dealt explicitly with the extent to which, and the reasons why, law "diffuses" from one country to another.[11] One aspect of diffusion is the reception of formal legal materials from another jurisdiction, an idea popularized as "legal transplants" by Watson (1974). Formal materials include statutes, precedents, and treatises. Spamann (2010a) showed that periphery countries continued to import such materials from core countries in the same legal family through the second half of the twentieth century. In a study of ten European high courts, Gelter & Siems (2014) showed, inter alia, that their citations to one another cluster within legal families as well.

Detecting outside influence is much harder when it does not involve literal copying or explicit references. In these cases, ECL can infer diffusion only from the timing of reforms (but see note 17 below). In this spirit, Goderis & Versteeg (2015) showed that constitutional rights in 180 countries after World War II tend to track other countries with the same legal origin, particularly the former colonizer, but also countries with the same religion or aid donor. By contrast, Ginsburg & Versteeg (2014) found no such pattern for the adoption of constitutional review, which correlates only with domestic political developments. There is a large literature on similar questions in political science. For example, Linos (2013) examined "how health, family, and employment laws spread across countries," emphasizing the mechanism of foreign role models in democratic discourse.

Inversely, even literal copying need not imply substantive convergence. For example, Greenhill et al. (2009) found that, although formal labor laws in 90 developing countries tended to resemble those of their main export destinations in the years 1986--2002, their labor practices did not. In general, recipient systems seem to be less "functional" as measured by rates of change and flexibility of corporate law in 10 countries (Pistor et al. 2003) and measures of effective legal institutions (e.g., absence of corruption) in 49 countries (Berkowitz et al. 2003a,b), at least where the local population had no prior experience with the foreign law.

3.4. Other Examples

There are many more examples in virtually all areas of law. Most studies examine cross-sectional correlations, i.e., correlations of two or usually more features across countries at a given point in time. For reasons discussed in Section 4.2, however, some studies examine the correlation of changes across

[11] There is also research on the more basic question whether (written) law in different countries is converging (e.g., Armour et al. 2009a, 2009b; Gahan et al. 2012).

time and space (as in the diffusion studies mentioned in the previous subsection). For example, Armour & Cumming (2008) showed that changes toward more "forgiving" bankruptcy laws tend to be followed by increases in self-employment ("entrepreneurship") in 15 Western countries from 1990 to 2005. Calderón & Chong (2009) showed that increases in labor regulation are associated with decreases in income inequality in a large sample of countries from 1970 to 2000. Gonzalez & Viitanen (2009) showed that the introduction of no-fault or unilateral divorce is associated with an increase in the divorce rate in European countries in the second half of the twentieth century. Klick et al. (2012) showed that abortion liberalization is associated with changes in sexual behavior (as proxied by gonorrhea incidence) in 41 countries from 1980 to 2000. This list could easily be extended (e.g., on patent law and innovation, see Moser 2005, Lerner 2009; on creditor rights and lending, see Haselmann et al. 2010).

4. Causation: The Identification Problem

In the previous section, I reported the results of the surveyed studies as mere correlations within a sample. The real interest for policy makers and most authors, however, is to what extent these correlations provide evidence for a causal link between the phenomena under study. I argue here that comparative evidence alone will hardly ever be sufficient to establish a causal claim and that statistical methods that purport to do so are likely to do more harm than good in comparative settings. Even if comparative data cannot identify any single causal theory, however, they are extremely important in narrowing down the set of plausible theories. A thoughtful pursuit of this more modest agenda seems to me the most fruitful avenue for ECL.

4.1. The Idea of Causal Inference

The basic idea of inferring causation from comparative data is a *ceteris paribus* argument. If countries *A* and *B* are alike in all relevant respects except *X* and *Y*, and *X* precedes *Y*, then any difference in *Y* must be caused by the difference in *X*.[12] The large variation of interesting attributes across countries presents a distinct advantage for such an argument, in that many attributes such as judicial review differ only across, but not within, countries. Unfortunately, this large variation is also comparative data's biggest disadvantage: There are never two countries that differ only in the attributes (*X*,*Y*) of interest. How then can comparative work identify the effect of the factor of interest (*X*) from among all others (*X*', *X*'', etc.), if at all?

Qualitative and quantitative studies pursue fundamentally different answers to this question. Qualitative studies examine the countries in question in depth in an attempt to rule out that there were meaningful differences on other possibly relevant characteristics *X*', *X*'', etc., or at least that these

[12] However, one could (and should!) ask how the difference in *X* arises. I hope that one can temporarily suppress this question for purposes of this thought experiment. In any real case, the origin of *X* is indeed a major issue, because whatever caused *X* may also have caused *Y* directly (cf. the discussion in the subsequent paragraphs of the main text).

differences had a confounding effect in these particular countries.[13] By contrast, quantitative work remains agnostic on the causes of outcome *Y* in any individual country and focuses instead on average effects of *X* in groups of countries. The basic argument is familiar from randomized control trials, i.e., true experiments. For example, in a drug trial, it is accepted and in fact assumed that one cannot ascertain how much, if at all, the drug changed the health outcome for any individual patient (or would have changed it, in the case of a person receiving the placebo). Rather, one infers the average effect of the drug from the difference in average health outcomes between the treatment and control groups. This works because randomizing who receives the drug (or, generically, the treatment) makes the two groups identical in all other respects in expectation, and the probability of deviations from this expectation above a certain size can be estimated from the dispersion of individual outcomes (Holland 1986).

Social scientists generally cannot manipulate their study contexts and randomly assign a "treatment." Modern empirical social science increasingly attempts to approximate the experimental ideal, however, by exploiting natural or quasi experiments, i.e., settings where the key independent variable of interest (i.e., the candidate cause) is plausibly as good as randomly assigned (Angrist & Pischke 2009, 2014; Dunning 2012; Ho & Rubin 2011; Imbens & Rubin 2015). Paradigmatic research exploits lotteries or sharp discontinuities at cutoff values of nonmanipulable continuous variables. For example, researchers have estimated the effect of neighborhood quality on personal outcomes by comparing winners and losers in a housing lottery (nber.org/mtopublic/) and the effect of elite schools on personal outcomes by comparing students just above and below the standardized test score required for admission (Dobbie & Fryer 2014).

Views differ on the availability of natural experiments in comparative settings. In my experience, outsiders to ECL tend to be very skeptical, whereas insiders are unsurprisingly more upbeat.[14] All agree, however, that comparative natural experiments are rare. Most research therefore pursues a different strategy. I describe and assess that strategy first before returning to natural experiments.

4.2. Controlling

As in epidemiology and other nonexperimental (i.e., observational) fields, the standard way to isolate the factor of interest in ECL is to "control" for possible confounding factors. For example, the estimation of an effect of shareholder protection on equity market capitalization may allow for the possibility that the latter is also influenced by the country's majority religion or the country's size. This is also sometimes described as "holding constant" the confounding factors (e.g., religion, size). This description

[13] Cf., e.g., King et al. (2001), Hirschl (2006), and Slater & Ziblatt (2013). The details, including the relationship to quantitative work, have been the subject of considerable controversy in political science; see, e.g., Geddes (2003), Brady & Collier (2004), Sekhon (2004), and Mahoney & Goertz (2006).

[14] Tellingly, political scientists and statisticians writing introductions to causal inference for lawyers do not mention comparative evidence. The only trace of comparative materials in Ho & Rubin (2011) is a cite to Bubb (2013). Epstein & King (2002, p. 103) mention cross-country research once, concluding a long list of possible sources of evidence with "*even* cross-country" (emphasis added).

evokes the idea that the technique is supposed to create and then to compare observations that differ only in the factors of interest (e.g., shareholder protection), at least in expectation.

This idea is easiest to understand in an approach called matching. First, each treated observation is matched to an untreated observation that is identical in terms of the pretreatment (control) variables, or at least sufficiently close in some technical sense. The treatment effect is then estimated as the average difference in outcomes between the treated and their respective matched untreated observations. For example, one could find for each country with good shareholder protection a matching country of the same religion and similar size with bad shareholder protection (countries without a match are omitted). Without assuming anything about the interaction of religion and country size, one could then estimate the effect of shareholder protection on equity market size by calculating the average difference in equity market size between the matched countries. This will identify the causal effect of shareholder protection, but only if the matching variables (here, religion and size) include all confounding variables, which presupposes in particular that all confounders are observable. In any event, matching is rarely used in comparative studies because countries are too few and too diverse to find close matches on all but very short lists of variables.[15]

Instead, most studies postulate a model of the interaction between the relevant variables and estimate its coefficients. A typical model may be

$$(\text{equity market size}) = a + b \cdot (\text{investor protection}) + c \cdot (\text{country size}) + \sum_{r \in religions} d_r \cdot r + (\text{residual}),$$

where the residual is assumed to be uncorrelated with the other right-hand-side variables and a, b, c, and the set of d_r will be estimated with some technique such as linear regression. The benefit of this approach over matching is that treated and untreated observations no longer need to be identical in terms of the control variables (here, country size and religion). The cost is that this works only if the variables truly interact only as postulated in the model. (For instance, the example above postulated that the effects of investor protection, country size, and religion are additive and linear.)

In general, "controlling" is model specific and, hence, only as good as the model (e.g., Leamer 1983). In principle, the model need not be as simplistic as in the example above and can include all sorts of candidate interactions and nonlinear effects. In comparative practice, however, there are far too few data points (countries) to estimate anything but simple linear models of the most obviously important observed variables. In addition, and as with matching, many relevant variables cannot be included because, at least in most countries, they are not observed or not measured. As a result, the model can at best provide modest assurance that the relevant factors are even approximately "held constant."

[15] For example, one can never closely match the United States on financial market size because the United States has by far the largest financial market in the world. One would get a closer but still imperfect match on financial market size relative to GDP. The more variables one needs to match on, the fewer and poorer the matches will become. Matching on the estimated propensity score (e.g., Angrist & Pischke 2009, section 3.3.2) can help but is only as good as the model for the propensity score (cf. the next two paragraphs in the main text).

If there are at least two observations per country at different periods in time (panel data), a popular way to deal with unobserved cross-sectional heterogeneity is to remove it all using country fixed effects or similar methods (see examples in Section 3.4).[16] This leads to a comparison of changes rather than levels and, for this reason, is often referred to as differences-in-differences (DD). For example, one could estimate the relationship between shareholder protection and equity market size by examining the cross-country correlation, not of these two variables at some point in time, but of changes in these two variables between two points in time. DD identifies a causal effect if the so-called parallel trends assumption holds, i.e., if in expectation (and conditional on controls) all countries would have experienced identical changes in outcomes (e.g., equity market size) but for the change in the explanatory variable of interest (e.g., shareholder protection). Unfortunately, this assumption is usually untenable in ECL because countries do not implement reforms randomly (cf. Rodrik 2012). Reforms often come in packages or react to changed circumstances unobserved by the researcher.[17] For example, many Asian countries improved shareholder protection in response to the Asian financial crisis of 1997/1998 while they were recovering from a recession and implementing numerous other reforms. Unless such contemporaneous changes are carefully controlled for---and this is usually not possible in ECL owing to lack of data and contextual information---DD estimates of the causal effect can easily be more biased than cross-sectional estimates.[18]

Painful experience in other disciplines has shown that these are not merely abstract concerns. In economics, Levine & Renelt (1992) revealed that almost all conclusions from cross-country growth regressions (as the literature then stood) were fragile to small changes in the selection of control variables. At the time, cross-country growth regressions followed the same template as most ECL today. Worse, LaLonde (1986) showed failure even of DD in a setting with much larger sample sizes and presumably less heterogeneity. LaLonde's test was to compare experimental estimates of the earnings effect of a job-training program, which were significantly positive, with observational estimates for the same program and data, which were mostly negative for men and much higher than the experimental

[16] With sufficiently long time series, even biases from unobserved time-*variant* variables can be removed (Abadie et al. 2010, 2015). There are other ways of using panel data that do not remove all cross-sectional variation; they are affected by a mixture of the problems discussed in this and the previous paragraph, and my sense is that they tend to obscure rather than ameliorate them.

[17] Diffusion studies present a special case of these problems (Spamann 2010a, sect. IV.C). Diffusion studies infer foreign influence from an increase in the probability of adopting a measure after candidate leader countries have adopted the measure. The problem here is that the follower countries might just be reacting to the same common shocks to market organization, technology, security threats, and the like, rather than to legal change in the leader country.

[18] A separate problem is that DD also amplifies the effect of noise if and because the temporal variation in measurement error relative to its level is larger than that in the variables of interest. On a technical note, many studies in empirical comparative law do not account for the fact that repeated observations from the same country are not statistically independent. This omission may severely exaggerate the precision of DD estimates (Bertrand et al. 2004). The standard fix is clustering of the standard errors (by country), but other methods may be necessary if the number of changes is small (Cameron & Miller 2015).

ones for women (cf. Glazerman et al. 2003).[19] In epidemiology, prominent cases include experimental refutation of observational claims that hormone replacement or certain vitamins reduce the risk of coronary heart disease and other ills (e.g., Hartz et al. 2013; Lawlor et al. 2004a,b; Smith & Ebrahim 2002).[20]

Under the impression of these and other failures, modern empiricists tend to be extremely skeptical of cross-country regressions. For example, Klick (2013, 908) commented on Djankov et al. (2003) that "[t]his kind of cross-sectional comparison has no chance of sorting out these issues, and conclusions based on this analysis are close to worthless in terms of having confidence in causality."

4.3. Natural Experiments and Instrumental Variables

Such disillusionment prompted what Angrist & Pischke (2010) called "the credibility revolution in empirical economics" (and, increasingly, political science): the search for natural experiments. Conceptually, natural experiments are controlling in reverse: Rather than attempt to control for confounding factors during estimation, knowledge of the "treatment" assignment mechanism (lottery, admission threshold, etc.) is used to argue that no confounding factors are at work in the first place. This so-called unconfoundedness assumption is partially testable because it implies that the covariates should be balanced between treated and untreated groups.

Importantly, "treatment" *X* (e.g., good shareholder protection) need not be (quasi-) randomly assigned itself. Instead, it is sufficient if

1. ("first stage") some third variable *Z* monotonically affects *X* (this can be tested empirically).
2. ("exclusion restriction") *Z* is not in any way correlated with the outcome *Y* except through its effect on *X*; that is, *Z* is "exogenous" (this is an untestable assumption).

If these two conditions are satisfied, then the causal effect of *X* on *Y* can be estimated as the ratio of the estimated effect of *Z* on *Y* over the estimated effect of *Z* on *X* (e.g., Angrist & Pischke 2009, ch. 4). Here *Z* is called an instrumental variable (IV). For example, La Porta et al. (1998) initially introduced legal origin

[19] That is, the observational estimates were derived from the same data but without knowledge of which applicant was assigned to the treatment group and hence eligible (but not required) to receive the training. The observational estimates were thus confounded by the usual problem that of two people with identical observed earnings and other characteristics, the one to apply for a training program tends to be the one whose job prospects are bleaker for some unobserved reason. A simple comparison of post-training earnings risks misattributing the effect of the initial bleak circumstances to the training program. The candidate "effect" (wages) in fact influences the candidate "cause" (enrolling in a training program). This problem is known as "endogeneity," but it can also be cast as an omitted variable problem because job prospects are not observable to the researcher. An example of an equivalent problem in comparative law is the possibility that high latent crime triggers harsh criminal law, leading to a positive correlation between crime and punishment even though punishment does reduce crime, everything else being equal.

[20] Grodstein et al. (2003) point out that experimental evidence did confirm many other epidemiological estimates. ECL works with far less data than epidemiology, so one should not expect ECL estimates to have the same success rate. That being said, I share the view that observational data remain useful, see subsection 4.4 below.

to the literature as an instrument for shareholder protection in an effort to estimate the latter's effect on equity market outcomes.

IV is the only type of natural experiment that has found wide application in ECL.[21] There are three mutually reinforcing reasons, however, to be very skeptical about IV estimates in ECL.[22] First, the IV estimator is notoriously unreliable---in particular, biased away from zero---in small samples (Bound et al. 1995), and the samples of ECL are extremely small. Second, the exclusion restriction will rarely, if ever, hold in comparative applications. Country-level factors do not cleanly affect, let alone correlate with, only one variable of interest. For example, legal origin correlates with multiple policy measures and outcomes, disqualifying it as an instrument for any one of them (La Porta et al. 2008). The same problem was discovered in many variables that were initially used as instruments in the cross-country growth literature (Bazzi & Clemens 2013, Durlauf et al. 2005). Third, the increasingly far-fetched instruments that researchers have turned to in response to the second problem are a priori weak instruments, i.e., they should be expected to have only a weak first-stage effect. Weak instruments exacerbate the first and the second problem, however, because the small sample bias and the bias from any remaining violation of the exclusion restriction are inversely related to the strength of the instrument (Bound et al. 1995). If an a priori weak instrument yields a strong first stage in the data at hand, this first stage result is more likely a false positive (cf. Section 5.3 below).

4.4. Summary: A More Modest Agenda

In summary, comparative data alone can rarely and perhaps never answer nontrivial causal questions. Attempts at causal inference using DD or IV will be grossly misleading if the treacherous conditions of these methods are not met; thus, they can do more harm than good.

That being said, comparative data remain important for assessing causal claims. They may not affirmatively pin down any particular cause, but they can considerably reduce the set of plausible ones. Comparative patterns are more consistent with some theories than with others (Durlauf 2009, Mankiw 1995).[23] Establishing such patterns should be considered a priority, and much remains to be done.[24]

[21] For example, Licht et al. (2007), Givati & Troiano (2012), and Dari-Mattiacci & Guerrero (2015) use language as instruments for culture in order to tease out the latter's effect on law. Besides IV, the other type of natural experiment is to exploit discontinuities around a cutoff, such as an international border (cf. Keele & Titiunik 2015). The only example of this in ECL is Bubb (2013). The reason why discontinuities are difficult to exploit in ECL is that many legal rules change simultaneously at the border, such that the discontinuity in outcomes, if there is one, does not identify the effect of any one of the legal changes. The point of Bubb (2013) was to show that there was no discontinuous change in outcomes at the border, providing evidence against any effect of law (although in theory his results are also consistent with multiple offsetting changes).

[22] Consistent with this skepticism, Albouy (2012) argues that the most famous comparative result using an instrumental variable (settler mortality; Acemoglu et al. 2001) was an artifact of measurement and specification error.

[23] For an example using comparative data in this conservative way, see., e.g., Givati (2014).

Although the patterns will always be subject to omitted variable bias, certain biases are less plausible than others (cf. Oster 2014).

Used carefully, comparative estimates can complement even experimental estimates of causal effects, considered the gold standard of empirical research. Experimental estimates cleanly identify causal effects in a particular setting (internal validity), but they cannot by themselves establish that the effect would be similar in a different setting (external validity) (cf. Rodrik 2009, Sims 2010). Natural experiments also rely on identifying assumptions that may turn out to be wrong (cf. section 4.3 above). Comparative data can thus be important in assessing the generalizability and robustness of (quasi-)experimental findings.

For example, quasi-experimental estimates of prison's deterrence and incapacitation effects vary widely in magnitude (see references in Spamann forthcoming). If the larger estimates were the more representative ones, then the US, which has by far the highest incarceration rate in the world, should have considerably lower crime rates than comparable countries. While no individual country is comparable to the US in this sense, a synthetic comparison can be constructed from cross-country regressions, and it does not have higher crime rates than the US (but a much lower incarceration rate). Section 4.2 above explained why this comparison can only be a rough approximation. But even the rough approximation reveals the inability of some factors (those included in the regression), and allows quantifying the size required of other factors, to reconcile US crime and incarceration rates with strong deterrence and incapacitation effects. Whereas the list of factors that might theoretically increase US crime rates is infinite, the list of plausible ones is arguably short. This list can be further pruned with circumstantial evidence, including US-specific evidence. In this way, cross-country data can contribute to assessing the plausible strength of deterrence and incapacitation across the US criminal justice system, even though cross-sectional comparative data cannot directly identify these effects (Spamann, forthcoming).

5. Other Methodological Issues

I now review certain features of empirical research that assume particular importance in ECL, regardless of whether causal inference is attempted. They divide into collecting, analyzing, and interpreting comparative data. All three depend on the hypothesis under investigation, and all are connected. In particular, the better the measurement and the controls, the stronger will be the conclusions that can be drawn from the data.

5.1. Data Collection: Measurement

Comparative work in general and comparative legal work in particular face special difficulties in designing and collecting consistent measurements. Even unemployment rates were difficult to compare across countries before the OECD created its harmonized unemployment rates. Modern

[24] For example, data on the number, size, and budget of courts have become available only recently and only for the member states of the Council of Europe, and even that only with significant qualifications regarding the comparability of the data (European Commission for the Efficiency of Justice 2014).

communications technology has considerably eased the problem of access to foreign raw information such as statutes or case law. Such technology now includes crowdsourcing sites like nomography.wustl.edu and participedia.net. The real difficulty, however, is to distill the raw information into a measure that achieves a close fit between the facts and the concept (validity) in a reproducible, consistent manner (reliability)

Earlier failures have demonstrated that reliable measurement of alien legal institutions requires a very detailed coding protocol and usually also the involvement of lawyers in collecting and coding the data (Spamann 2010b,c). Some legal institutions may be sufficiently straightforward for lay coding, as is done in the CCP. However, for more complex questions such as the resolution of a particular case or the legality of a transaction, it is hard to imagine that lay coders could correctly combine or even locate all relevant materials. To enable others to verify and replicate the measurement, it is also advisable to post the raw data and coding protocol online (cf. Spamann 2008).

The validity of comparative measures may be compromised by the inconsistency of meaning or importance of certain features across countries. For example, some statutory provision may be very important for shareholder protection in one country but irrelevant in another because of the absence or presence of certain other rules or institutions (Black et al. 2014). Similarly, some institution may be important in one country but redundant in another because of the presence of a functional equivalent (on functional equivalence, see Michaels 2006). Whether the measure should take into account such functional equivalents is determined by the measured concept and, thus ultimately, by the hypothesis under investigation. For example, the hypothesis may be specifically about the effect of *statutory* shareholder protection, perhaps because this is the only concept under policy makers' direct control. In this case, taking into account case law or legal practice in measurement would diminish rather than increase the measure's validity, notwithstanding the fact that case law and practice may be very important for broader concepts of shareholder protection. Conceptual clarity is thus a precondition for valid measurement. Besides, amorphous concepts such as shareholder protection may require further refinement before a discussion of validity can even begin (Bebchuk & Hamdani 2009; see generally Adcock & Collier 2001).

Legal measurement design has made considerable progress. Measuring law is no longer limited to counting the presence or absence of certain statutory rights (as in, e.g., Djankov et al. 2008b, La Porta et al. 1998). One possible improvement is to determine the weight of individual components through factor analysis (Rosenthal & Voeten 2007). A more fundamental improvement is to account for the interaction of different rules by coding, not the rules, but the treatment of a paradigmatic case (cf. Djankov et al. 2008a, World Bank 2014) or, better, several cases (resembling the common core approach in classical comparative law [common-core.org], see Michaels 2009, Spamann 2009). Nonlegal phenomena can serve as an indirect measure if they are plausibly directly and strongly related to the legal aspect of interest. For example, the average discount at which minority shares trade relative to control blocks (Dyck & Zingales 2004) is arguably a direct function of minority shareholder protection, albeit not only legal shareholder protection. When more than one measure is available, all measures can be synthesized into one superior measure (e.g., Pemstein et al. 2010).

In principle, the quality of measurement can be explicitly validated empirically. In particular, one can verify that the measurement correlates with other measurements of the same concept, or with other variables with which the concept is known to correlate (Adcock & Collier 2001). Unfortunately, such opportunities are rarely available in ECL (but cf. Hallward-Driemeier & Pritchett 2015, Ríos-Figueroa & Staton 2012, Spamann 2010b). As a result, studies tend to test joint hypotheses: the substantive hypothesis and the hypothesis that the measurement of the relevant concept is valid (or as many such hypotheses as there are concepts involved). I return to this problem in subsection 5.3.

5.2. Data Analysis: Controlling Revisited

In data analysis, it is important to control as comprehensively, flexibly, and transparently as possible because candidate causes ("treatments") are not randomly assigned in ECL. Without random assignment, other variables may be systematically correlated with the treatment and bias the estimate of the effect of interest. Though it may be impossible to eliminate all potential confounds and identify a particular causal effect (see Section 4.2), sensible controlling will help a great deal in limiting the set of plausible biases and, ultimately, plausible causal relationships. In this respect, ECL has much to learn from modern growth empirics (cf. Durlauf 2001, 2009; Durlauf et al. 2005).

Current practice in ECL is to select a small number of controls ad hoc. This practice is motivated by the fact that ECL samples are small, such that it is not possible to estimate precisely a larger number of parameters with classical methods. Ad hoc selection merely gives a semblance of precision, however, by neglecting model uncertainty. Three improvements are available. First, missing data for individual observations can be imputed to increase sample size. Imputation not only avoids wasting information, but also reduces selection bias (Honaker & King 2010, Little & Rubin 2002). Second, model-averaging techniques can explicitly account for model uncertainty (e.g., Magnus et al. 2010). Third, in some applications, principled selection among controls is possible even when the number of possible controls is larger than the sample size (Belloni et al. 2014). The latter assumes that the number of truly relevant factors is ultimately small. This so-called sparsity assumption is strong, but if it is considered false, then comparative research should arguably be abandoned because there is no hope of identifying complex connections with few data points.[25]

Yet, there is also a danger of controlling too much. To be more precise, the use of certain controls implies assumptions that may not be plausible or may change the interpretation of the results. In particular, if a regression is supposed to approximate a causal relationship, using a variable as a control implicitly assumes that the variable is exogenous, i.e., not affected by the outcome variable. If it were affected (i.e., endogenous), then the estimates for all of the regression's independent variables would be biased in generally unknown ways relative to their true causal effects. Although the importance of the exogeneity assumption is well known in the abstract, its consequences are not always fully

[25] This is the "bet on sparsity" principle coined by Hastie et al. (2009, p. 611): "Use a procedure that does well in sparse problems, since no procedure does well in dense problems." But see Gelman (2011), who argues that sparsity is inapposite in social science.

appreciated. For example, most studies, including all of the literature in Law and Finance as well as legal origins, control for GDP per capita. The resulting estimates are unbiased only if GDP is exogenous, i.e., if the outcome variables, such as financial market size or the quality of judicial procedures, have no effect on GDP. This is possible, but it would make the estimates much less policy relevant.

5.3 Interpretation

Last but not least, it is important to interpret results sensibly in light of prior information, including the study design. I have discussed the obstacles to causal inference (Section 4) and the steps that should be taken to at least reduce the number of alternative causal interpretations (Section 5.2). That discussion was mostly concerned with bias, i.e., the possibility that the estimate would systematically be higher or lower than the true effect because of confounding with other effects such as selection. I now turn to spuriousness, i.e., the possibility that the estimate on a particular sample is fortuitously higher or lower than the true effect because of sampling error. For example, the treatment group in a drug trial may fortuitously contain a disproportionate number of subjects with hidden health problems, which would make the drug's efficacy appear less than it truly is.

Unlike bias, sampling error will differ from sample to sample. As a result, replication on a new, independent sample could address suspicions that the finding is spurious. In ECL, this is cold comfort. Usually, there is only one sample, which is the set of existing countries on Earth, or perhaps some relevant subset thereof.[26] Replication is therefore not an option in ECL. Consequently, ECL must pay particular attention to spuriousness in interpreting its results.

The standard way of dealing with sampling error is to derive an estimate of the sampling variation (the standard error) from the data to calculate the probability of (erroneously) estimating an effect of equal or greater size under the null hypothesis of no effect---the p-value.[27] A p-value of 10% or perhaps 5% is commonly considered statistically significant and tends to be required for publication. However, as is well known in theory and increasingly appreciated in practice in other disciplines, low reported p-values are insufficient to address spuriousness (see, e.g., Pashler & Wagenmakers 2012).[28] There are two reasons for this.

First, because of multitesting, the true probability of falsely rejecting the null hypothesis tends to be much higher than the reported p-value. It is common for individual researchers to try many variables and specifications and report only the "successful" ones. In any event, researchers collectively try many more variables and specifications, and only the "successful" researchers publish their findings. The problem here is not multitesting per se, as extensive testing and even filtering of promising results is

[26] Cf. notes 1 and 2 above.

[27] Estimating standard errors can be tricky. One issue of particular importance to comparative studies is that no country is literally independent from all others, as would be required for standard methods of calculating standard errors. This issue has not received attention in ECL, presumably on the assumption that it is minor. The latter assumption is in tension with the findings of the diffusion literature (see section 3.3. above).

[28] An additional problem is that an exclusive focus on statistical significance does not take into account the respective consequences of erring on one side or another. See, e.g., Ziliak & McCloskey (2007).

desirable. Rather, the problem is that the reported p-values are grossly understated. Reported p-values assume that only a single study was performed. But the greater the number of (unreported) studies, the greater the probability of finding a spurious result above a certain size.[29]

Second, by definition, p-values are not equal to the probability that the null hypothesis is correct, nor is one minus the p-value equal to the probability that the alternative hypothesis is correct. Rather, the odds for the alternative hypothesis after seeing the data (the posterior odds) are equal to the odds prior to seeing the data multiplied by the Bayes factor, which is the ratio of the prior probabilities of the data under the alternative and the null hypotheses, respectively (e.g., Kass & Raftery 1995). Of this formula's two factors, only the Bayes factor is loosely related to the p-value (e.g., Strnad 2007, sect. 2.2). The other factor (the prior odds) means that prior plausibility matters even after seeing the data. A wildly implausible theory may become less implausible after seeing the data, but unless the result is extremely strong, the theory will remain implausible. Importantly, multitesting presumably implies that any of the tested models/theories has a low prior probability of being true, or else fewer models/theories would have been tested (Cox 2006, p. 88).

The formula for the posterior odds also emphasizes that a test can be informative only to the extent the predictions of the null and the alternative differ. At first sight, this may not seem very important because a particular point estimate is naturally much more likely to arise if the true effect is equal to or close to the point estimate (the usual interpretation) than if the true effect is zero. But an effect of that size may not be plausible, and an effect of plausible size may not yield very different predictions from the null (cf. Gelman & Carlin 2011). In particular, when measurement is very noisy, the expected estimate under the alternative hypothesis will be strongly biased toward zero and thus very similar to the prediction of the null hypothesis. Besides, the alternative hypothesis is rarely if ever specified as a precise number, let alone the one actually later estimated. This issue would require a longer detour into statistics and is beyond the scope of this paper.[30] The important takeaway, however, is that limitations of the data and data analysis remain important for interpretation of the results even if the latter are "statistically significant."

An approximate litmus test is to what extent an estimated coefficient of zero would be considered as evidence against the alternative hypothesis. The less this is the case, the less the predictions of the alternative hypothesis differ from the null when allowance is made for measurement error and other design issues; hence, the less one can learn from the evidence. For example, if estimates with a crude index would be considered uninformative if they were close to zero, they should be considered similarly uninformative if they happen to be large and "significant."

[29] In theory, p-values can be adjusted to account for such multi-testing (e.g., Benjamini & Hochberg 1995). In practice, this does not work to the extent the multi-testing is done by different researchers unaware of each others' work.

[30] See, e.g., Strnad (2007, sect.2.2). Bayesian statistics formally integrates data and prior beliefs, including about aspects of the study design (e.g., Gelman et al. 2013). It allows precise treatment of, e.g., doubts about the strength and exogeneity of instruments discussed in Section 4.3 (Conley et al. 2012).

As a practical matter, the foregoing precludes credible tests of effects that must be small relative to the noise. In particular, comparative data cannot be sensibly used to test the effect of technical rules on big picture outcomes such as GDP growth that are the product of a large number of factors. Instead, focus should be directed to the technical rules' effects on less distant outcomes. For example, to test the effect of culture on property protection, Dari-Mattiacci & Guerriero (2015) collected data on one directly pertinent and easily measurable variable, namely the number of years, if ever, after which an illegally dispossessed owner of a moveable good loses her property rights to a bona fide purchaser.

Because comparative evidence is limited, it is imperative to test the theory's assumptions or implications also in domestic settings. For example, Linos (2013) used survey evidence from the United States to bolster her claim that foreign and international models legitimate policy options and, hence, diffuse. This evidence is particularly powerful because the size, geopolitical dominance, and geographic isolation of the United States made US voters least likely to be so influenced (this argument for the power of the US evidence is known as most difficult case logic; Hirschl 2006). Similarly, Cassar et al. (2014) bolstered claims that well-functioning legal institutions increase trust with experimental evidence. By contrast, the claim that legal origin matters became much less plausible when domestic evidence did not fit the theory that archetypical differences between common and civil law, such as reliance on case law, caused the tested outcomes (for example, the driving force of US investor protection turns out to be statutes, not case law).

6. Conclusion

Comparative information is important to assess causal claims. Nevertheless, this article cautions against drawing overly strong conclusions from comparative data alone. From an individual researcher's perspective, it is tempting to brush aside these concerns and "just do it." After all, unlike in other disciplines, there is no risk of being proven wrong by a controlled experiment. In fact, there is not even a risk that another researcher will obtain different results on a different sample---there is only one planet Earth. But from the perspective of the discipline as a whole, the inability to weed out errors through replication is all the more reason to look critically at empirical findings. Otherwise, erroneous findings will pile up and blur our vision, and the incentive to publish such findings will divert attention from higher-value targets.

We will do better if we are clear about the strengths and weaknesses of comparative data. Comparative data will rarely if ever sort out causal questions by themselves. That being said, they can be an extremely important piece in a broader empirical and theoretical analysis. Theories gain strength if they fit the comparative facts, and lose it if they do not. Thus, establishing comparative facts through high-quality data collection should be the first priority.

References

1. Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. Journal of the American Statistical Association 105:493-505.
2. Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. Comparative Politics and the Synthetic Control Method. American Journal of Political Science 59:405-510.
3. Acemoglu, Daron, Simon Johnson, and James Robinson. 2001. The Colonial Origins of Comparative Development: An Empirical Investigation. American Economic Review 91:1369-401.
4. Adcock, Robert, and David Collier. 2001. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. American Political Science Review 95:529-46.
5. Albouy, David. 2012. The Colonial Origins Of Comparative Development: An Empirical Investigation: Comment. Am. Econ. Rev. 102:3059-76.
6. Angrist, Joshua, and Jörn-Steffan Pischke. 2009. Mostly Harmless Econometrics. Princeton, NJ: Princeton Univ. Press.
7. Angrist, Joshua, and Jörn-Steffan Pischke. 2010. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics. Journal of Economic Perspectives. 24(2):3-30i.
8. Angrist, Joshua, and Jörn-Steffan Pischke. 2014. Mastering 'Metrics: The Path from Cause to Effect. Princeton, NJ: Princeton University Press.
9. Armour, John, and Douglas Cumming. 2008. Bankruptcy Law and Entrepreneurship. American Law and Economics Review. 10:303-50.
10. Armour, John, Simon Deakin, Priya Lele, Mathias Siems. 2009a. How Do Legal Rules Evolve? Evidence from a Cross-Country Comparison of Shareholder, Creditor, and Worker Protection. American Journal of Comparative Law 57:579-629.
11. Armour, John, Simon Deakin, Prabirji Sarkar, Mathias Siems, and Ajit Singh. 2009b. Shareholder Protection and Stock Market Development: An Empirical Test of the Legal Origins Hypothesis. Journal of Empirical Legal Studies 6:343-80.
12. Barth, James, Gerard Caprio, Ross Levine. 2006. Rethinking Bank Regulation: Till Angels Govern. Cambridge, UK: Cambridge University Press.
13. Barth, James, Gerard Caprio, and Ross Levine. 2013. Bank Regulation and Supervision in 180 Countries from 1999 to 2011. NBER Working Paper No. 18733.
14. Bazzi, Samuel, and Michael Clemens. 2013. Blunt Instruments: Avoiding Common Pitfalls in Identifying the Causes of Economic Growth. American Economic Journal: Macroeconomics 5(2):152–186.
15. Bebchuk, Lucian, and Assaf Hamdani. 2009. The Elusive Quest for Global Corporate Governance Standards. University of Pennsylvania Law Review 157:1263-1317.
16. Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. High-Dimensional Methods and Inference on Structural and Treatment Effects. Journal of Economic Perspectives 28(2):1-23.
17. Ben-Bassat, Avi, and Momi Dahan. 2008. Social rights in the constitution and in practice. Journal of Comparative Economics 36:103–119.
18. Benjamini, Yoav, and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society (Series B) 57:289-300.
19. Berkowitz, Daniel, Katharina Pistor, and Jean-Francois Richard. 2003a. Economic development, legality, and the transplant effect. European Economic Review 47:165-195.

20. Berkowitz, Daniel, Katharina Pistor, and Jean-Francois Richard. 2003b. The Transplant Effect. American Journal of Comparative Law 51:163-203.
21. Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. How much should we trust Differences-in-Differences Estimates? Quarterly Journal of Economics 119:249-275.
22. Black, Bernard, Gledson de Carvalho, Vikramaditya Khanna, Woochan Kim, and Burcin Yurtoglu. 2014. Methods for Multicountry Studies of Corporate Governance: Evidence from the BRIKT Countries. Journal of Econometrics 183:230-240.
23. Botero, Juan, Simeon Djankov, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2004. The Regulation of Labor. Quarterly Journal of Economics 119:1339-1382.
24. Bound, John, David Jaeger, and Regina Baker. 1995. Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak. Journal of the American Statistical Association 90:443-450.
25. Brady, Henry, and David Collier, eds. 2004. Rethinking Social Inquiry: Diverse Tools, Shared Standards. Lanham, MD: Rowman & Littlefield.
26. Bubb, Ryan. 2013. The Evolution of Property Rights: State Law or Informal Norms? Journal of Law and Economics 56:555-594.
27. Calderón, César, and Alberto Chong. 2009. Labor market institutions and income inequality: an empirical exploration. Public Choice 138:65-81.
28. Cameron, Colin, and Douglas Miller. 2015. A Practitioner's Guide to Cluster-Robust Inference. Journal of Human Resources. J. Hum. Resour. 50:317-72
29. Carrubba, Clifford, Matthew Gabel, Gretchen Helmke, Andrew Martin, and Jeffery Staton. 2012. An Introduction To The Complaw Database. Unpublished Manuscript, Emory University. Available at http://perma.cc/MMA6-UBVR
30. Cassar, Alessandra, Giovanna d'Adda, and Pauline Grosjean. 2014. Institutional Quality, Culture, and Norms of Cooperation: Evidence from Behavioral Field Experiments. Journal of Law and Economics 57:821-863.
31. Cheffins, Brian, Steven Bank, and Harwell Wells. 2013. Questioning 'law and finance': US stock market development, 1930–70. Business History 55:601-619.
32. Cheibub, José Antonio, Zachary Elkins, and Tom Ginsburg. 2013. Beyond Presidentialism and Parliamentarism. British Journal of Political Science 44:515-44.
33. Conley T, Hansen C, Rossi P. 2012. Plausibly Exogenous. Rev. Econ. Stat. 94:260-72.
34. Cox, D.R. 2006. Principles Of Statistical Inference. Cambridge, UK: Cambridge University Press.
35. Dari-Mattiacci, Giuseppe and Carmine Guerriero. 2015. Law and Culture: A Theory of Comparative Variation in Bona Fide Purchase Rules. Oxford Journal of Legal Studies. In Press.
36. Davis, Kevin. 2014. Legal Indicators: The Power of Quantitative Measures of Law. Annual Review of Law and Social Science 10:37-52.
37. Dixon, Rosalind, and Tom Ginsburg. 2011. Deciding Not to Decide: Deferral in Constitutional Design. International Journal of Constitutional Law 9:636-672.
38. Dixon, Rosalind, and Richard Holden. 2014. Constitutional Amendment Rules: The Denominator Problem. In Ginsburg 2014, pp. 195-218.
39. Djankov, Simeon, Oliver Hart, Caralee Mcliesh, and Andrei Shleifer A. 2008a. Debt Enforcement Around The World. Journal of Political Economy. 116:1105-49.
40. Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2002. The Regulation of Entry. Quarterly Journal of Economics 117:1-37.
41. Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2003. Courts. Quarterly Journal of Economics 118:453-517.
42. Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer 2008b. The Law and Economics of Self-Dealing. Journal of Financial Economics 88:430-65.

43. Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2010. Disclosure by Politicians. American Economic Journal: Applied Economics 2:179-209.
44. Djankov, Simeon, Caralee Mcliesh, and Andrei Shleifer. 2007. Private Credit in 129 Countries. Journal of Financial Economics 84:299-329.
45. Dobbie, Will, and Roland Fryer. 2014. The Impact of Attending a School with High-Achieving Peers: Evidence from the New York City Exam Schools. American Economic Journal: Applied Economics 6(3):58-75.
46. Dreher, Axel, Martin Gassebner, and Lars Siemers. 2010. Does Terrorism Threaten Human Rights? Evidence from Panel Data. Journal of Law and Economics 53:65-93.
47. Dunning, Thad. 2012. Natural Experiments in the Social Sciences: A Design-Based Approach. Cambridge, UK: Cambridge University Press.
48. Durlauf, Steven. 2001. Manifesto For A Growth Econometrics. J. Econometr. 100:65-69.
49. Durlauf, Steven 2009. The Rise and Fall of Cross-Country Growth Regressions. History of Political Economy 41:315-333.
50. Durlauf, Steven, Paul Johnson, and Jonathan Temple. 2005. Growth Econometrics. In Handbook Of Economic Growth, Ed. Philippe Aghion and Steven Durlauf, 1A:555-677. Amsterdam: North Holland.
51. Dyck, Alexander, and Luigi Zingales. 2004. Private Benefits of Control: An International Comparison. Journal of Finance 59:537-600.
52. Elkins, Zachary, Tom Ginsburg, and James Melton. 2008. Baghdad, Tokyo, Kabul …: Constitution Making in Occupied States. William and Mary Law Review 49:1139-1178.
53. Elkins, Zachary, and John Sides. 2007. Can Institutions Build Unity in Multiethnic States? American Political Science Review 101:693-708.
54. Epstein, Lee, and Gary King. 2002. The Rules Of Inference. The University Of Chicago Law Review 69:1-133.
55. European Commission for the Efficiency of Justice (CEPEJ). 2014. European Judicial Systems – Edition 2014 (2012 data): Efficiency and Quality of justice.
56. Gahan, Peter, Richard Mitchell, Sean Cooney, Andrew Stewart, and Brian Cooper. 2012. Economic Globalization and Convergence in Labor Market Regulation: An Empirical Assessment. American Journal of Comparative Law 60:703-741.
57. Garoupa, Nuno, and Tom Ginsburg. 2009. Guarding the Guardians: Judicial Councils and Judicial Independence. American Journal of Comparative Law 57:103-134.
58. Geddes, Barbara. 2003. Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics. Ann Arbor, MI: University of Michigan Press.
59. Gelman, Andrew. 2011. Causality and Statistical Learning. American Journal of Sociology 117:955-966.
60. Gelman Andrew, and John Carlin. 2011. Beyond Power Calculations: Assessing Type S (Sign) And Type M (Magnitude) Errors. Perspectives on Psychological Science. 9:641-51.
61. Gelman, Andrew, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. 2013. Bayesian Data Analysis. London: Chapman & Hall/CRC Press. 3rd Ed.
62. Gelter, Martin, and Mathias Siems. 2014. Citations to Foreign Courts – Illegitimate and Superfluous, or Unavoidable? Evidence from Europe. American Journal of Comparative Law 62:35-85.
63. Ginsburg, Tom. 2010. Constitutional Specificity, Unwritten Understandings and Constitutional Agreement. In Constitutional topography: Values and constitutions, Ed. Andras Sajo and Renata Uitz, PP. 69-93. The Hague: Eleven Int.
64. Ginsburg, Tom, ed. 2012. Comparative constitutional design. Cambridge, UK: Cambridge University Press.

65. Ginsburg, Tom, Zachary Elkins, and Justin Blount. 2009. Does the Process of Constitution-Making Matter? Annual Review of Law and Social Science 5:209-223.
66. Ginsburg, Tom, James Melton, and Zachary Elkins. 2009. The Endurance of National Constitutions. Cambridge, UK: Cambridge University Press.
67. Ginsburg, Tom, James Melton, and Zachary Elkins. 2011. On the Evasion of Executive Term Limits. William and Mary Law Review 52:1807-1873.
68. Ginsburg, Tom, and Mila Versteeg. 2014. Why Do Countries Adopt Constitutional Review? Journal of Law, Economics, and Organization. 30:587-622.
69. Givati, Yehonatan. 2014. Legal Institutions and Social Values: Theory and Evidence from Plea Bargaining Regimes. Journal of Empirical Legal Studies 11:867-893.
70. Givati, Yehonatan, and Ugo Troiano. 2012. Law, Economics, and Culture: Theory of Mandated Benefits and Evidence from Maternity Leave Policies. Journal of Law and Economics 55:339-364.
71. Glaeser, Edward, and Andrei Shleifer. 2002. Legal Origins. Quarterly Journal of Economics 117:1193-1229.
72. Glazerman, Steven, Dan Levy, and David Myers. 2003. Nonexperimental versus Experimental Estimates of Earnings Impacts. Annals of the American Academy of Political and Social Science 589:63-93.
73. Glendon, Mary Ann. 1989. The Transformation Of Family Law. Chicago: University of Chicago Press.
74. Goderis, Benedikt, and Mila Versteeg. 2012. Human Rights Violations after 9/11 and the Role of Constitutional Constraints. Journal of Legal Studies 41:131-164.
75. Goderis, Benedikt, and Mila Versteeg. 2015. The Diffusion of Constitutional Rights. International Review of Law and Economics 39:1-19.
76. Gonzalez, Libertad, and Tarja Viitanen. 2009. The Effect of Divorce Laws on Divorce Rates in Europe. European Economic Review 53:127-38.
77. Greenhill, Brian, Layna Mosley, and Aseem Prakash. 2009. Trade-based Diffusion of Labor Rights: A Panel Study, 1986–2002. American Political Science Review 103:669-690.
78. Grodstein, Francine, Thomas Clarkson, and JoAnn Manson. 2003. Understanding the Divergent Data on Postmenopausal Hormone Therapy. New England Journal of Medicine 348:645-650.
79. Hallward-Driemeier, Mary, and Lant Pritchett. 2015. How Business is Done in the Developing World: Deals versus Rules. Journal of Economic Perspectives 29(3):121-40.
80. Hartz, Arthur, Tao He, Robert Wallace, and John Powers. 2013. Comparing Hormone Therapy Effects in Two RCTs and Two Large Observational Studies That Used Similar Methods for Comprehensive Data Collection and Outcome Assessment. BMJ Open 2013;3:e002556. doi:10.1136/bmjopen-2013-002556.
81. Haselmann, Rainer, Katharina Pistor, and Vikrant Vig. 2010. How Law Affects Lending. Review of Financial Studies 23:549-580.
82. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman 2009. The Elements Of Statistical Learning. New York: Springer. 2nd ed.
83. Hirschl, Ran. 2006. The Question of Case Selection in Comparative Constitutional Law. American Journal of Comparative Law 53:125-155.
84. Ho, Daniel, and Donald Rubin. 2011. Credible Causal Inference for Empirical Legal Studies. Annual Review of Law and Social Science 7:17-40.
85. Holderness, Clifford. Forthcoming. Law And Ownership Reexamined. Critical Finance Review.
86. Holland, Paul. 1986. Statistics and Causal Inference. Journal of the American Statistical Association 81:945-960.

87. Honaker, James, and Gary King. 2010. What to Do about Missing Values in Time-Series Cross-Section Data. American Journal of Political Science 54:561-581.
88. Imbens, Guido, and Donald Rubin. 2015. Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge, UK: Cambridge University Press .
89. Jackson, Howell, and Mark Roe. 2009. Public and private enforcement of securities laws: Resource-based evidence. Journal of Financial Economics 93:207-238.
90. Jansen, Nils. 2006. Comparative Law and Comparative Knowledge. In Reiman and Zimmermann (2006), 305-337.
91. Kass, Robert, and Adrian Raftery. 1995. Bayes Factors. Journal of the American Statistical Association 90:773-795.
92. Keele, Luke, and Rocio Titiunik. 2015. Geographic Boundaries as Regression Discontinuities. Political Analysis 23:127-55.
93. King, Gary, Robert Keohane, and Sidney Verba. 2001. Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton, NJ: Princeton University Press.
94. Klerman, Daniel, and Paul Mahoney. 2007. Legal Origin? Journal of Comparative Economics 35:278-293.
95. Klerman, Daniel, Paul Mahoney, Holger Spamann, and Mark Weinstein. 2011. Legal Origin or Colonial History? Journal of Legal Analysis 3:379-409.
96. Klick, Jonathan. 2013. Shleifer's Failure. Texas Law Review 91:899-909.
97. Klick, Jonathan, Sven Neelsen, and Thomas Stratmann. 2012. The Relationship between Abortion Liberalization and Sexual Behavior: International Evidence. American Law and Economics Review 14:457-487.
98. La Porta, Rafael, Florencio Lopez-de-Silanes, Cristian Pop-Eleches, and Andrei Shleifer. 2004. Judicial Checks and Balances. Journal of Political Economy 112:445-70.
99. La Porta, Rafael, Florencio Lopez-de-Silanes, and Andrei Shleifer. 1999. Corporate Ownership Around the World. Journal of Finance 54:471-517.
100. La Porta, Rafael, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2002a. Government Ownership of Banks. Journal of Finance 57:265-301.
101. La Porta, Rafael, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2006. What Works in Securities Laws? Journal of Finance 61:1-32.
102. La Porta, Rafael, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2008. The Economic Consequences of Legal Origin. Journal of Economic Literature 46:285-332.
103. La Porta, Rafael, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2013. Law And Finance After a Decade of Research. In Handbook of the Economics of Finance, George Constantinides, Milton Harris, and Rene M. Stulz eds. Vol. 2A:425-91. Amsterdam: Elsevier.
104. La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny. 1997. Legal Determinants of External Finance. Journal of Finance 52:1131.
105. La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny. 1998. Law and Finance. Journal of Political Economy 106:113.
106. La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny. 2000. Agency Problems and Dividend Policies around the World. Journal of Finance 55:1-33.
107. La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny. 2002b. Investor Protection and Corporate Valuation. Journal of Finance 57:1147-70.
108. LaLonde, Robert. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. American Economic Review 76:604-620.
109. Lambert, Edouard. 1905. Conception générale, définition, méthode et histoire du droit comparé. Le droit comparé et l'enseignement du droit. In Congrès international de droit

comparé, tenu à Paris du 31 juillet au 4 août 1900. Procès verbaux des séances et documents, vol. I, 26-61. Paris: Librairie Générale de Droit et de Jurisprudence.
110. Landmann, Todd, and Neil Robinson. 2009. The SAGE Handbook of Comparative Politics. Thousand Oaks, Ca: Sage Publ.
111. Lasser, Mitchel. 2004. Judicial Deliberations: A Comparative Analysis of Judicial Transparency and Legitimacy. Oxford: Oxford University Press.
112. Law David, and Mila Versteeg. 2013. Sham Constitutions. California Law Review 101:863-952.
113. Law, David, and Mila Versteeg. 2013. Sham Constitutions. California Law Review 101:863-952.
114. Lawlor, Debbie, George Davey Smith, Richard Bruckdorfer, Devi Kundu, and Shah Ebrahim. 2004a. Those Confounded Vitamins: What Can We Learn from the Differences Between Observational Versus Randomised Trial Evidence? Lancet 363(9422):1724-27.
115. Lawlor, Debbie, George Davey Smith, and Shah Ebrahim. 2004b. The Hormone Replacement: Coronary Heart Disease Conundrum: Is This the Death of Observational Epidemiology? International Journal of Epidemiology 33:464-67.
116. Leamer, Edward. 1983. Let's Take the Con Out of Econometrics. American Economic Review 73:31-43.
117. Lerner, Josh. 2009. The Empirical Impact of Intellectual Property Rights on Innovation: Puzzles and Clues. American Economic Review Papers & Proceedings 99(2):343–348.
118. Levine, Ross, and David Renelt. 1992. A Sensitivity Analysis of Cross-Country Growth Regressions. American Economic Review 82:942-963.
119. Licht, Amir, Chanan Goldschmidt, and Shalom Schwartz. 2007. Culture Rules: The Foundations of the Rule of Law and Other Norms of Governance. Journal of Comparative Economics 35:659-688.
120. Linos, Katerina. 2013. The Democratic Foundations of Policy Diffusion. Oxford and New York: Oxford University Press.
121. Little, Roderick, and Donald Rubin. 2002. Statistical Analysis with Missing Data (2nd ed.) Hoboken, NJ: Wiley. 2nd ed.
122. Magnus, Jan, Owen Powell, and Patricia Prüfer. 2010. A Comparison of Two Model Averaging Techniques With an Application to Growth Empirics. Journal of Econometrics 154:139-53.
123. Mahoney, James, and Gary Goertz. 2006. A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research. Political Analysis 14:227-249.
124. Mankiw, N. Gregory. 1995. The Growth of Nations. Brookings Papers on Economic Activity 1:275-326.
125. Manuel, Trevor, Carlos Arruda, Jihad Azour, Chong-En Bai, et el. 2013. Independent Panel Review of the Doing Business Report. Independent Panel Review Report, World Bank, Washington, DC. http://perma.cc/YMX5-WP9C
126. Merryman, John. 1969. The Civil Law Tradition. Redwood City, CA: Stanford University Press.
127. Michaels, Ralf. 2006. The Functional Method of Comparative Law. In Reimann & Zimmermann 2006, pp. 339-82.
128. Michaels, Ralf. 2009. Comparative Law by Numbers? Legal Origins Thesis, Doing Business Reports, and the Silence of Traditional Comparative Law. American Journal of Comparative Law 57:765-795.
129. Michalopoulos, Stelios, and Elias Papaioannou. 2014. National Institutions and Subnational Development in Africa. Quarterly Journal of Economics 129:151-213.
130. Milhaupt, Curtis, and Katharina Pistor. 2008. Law and Capitalism: What Corporate Crises Reveal About Legal Systems and Economic Development Around the World. Chicago: University of Chicago Press.

131. Moser, Petra. 2005. How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World's Fairs. American Economic Review 95:1214-1236.
132. Mulligan, Casey, and Andrei Shleifer. 2005. Conscription as Regulation. American Law and Economics Review 7:85.
133. Oster, Emily. 2014. Unobservable Selection and Coefficient Stability: Theory and Validation. Working paper, University of Chicago Booth School of Business, available at http://faculty.chicagobooth.edu/emily.oster/papers/selection.pdf
134. Oto-Peralías, Daniel, and Diego Romero-Avila. 2014. The Distribution of Legal Traditions Around the World: A Contribution to the Legal-Origins Theory. Journal of Law and Economics 57:561-628.
135. Pashler, Harold, and Eric–Jan Wagenmakers. 2012. Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? Perspectives on Psychological Science 7:528-530.
136. Pemstein, Daniel, Stephen Meserve, and James Melton. 2010. Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type. Political Analysis 18:426-49.
137. Pistor, Katharina, Yoram Keinan, Jan Kleinheisterkamp, and Mark West. 2003. Innovation in Corporate Law. Journal of Comparative Economics 31:676-94.
138. Pistor, Katharina. 2010. Statistische Rechtsvergleichung: Eine Kritische Bestandsaufnahme. Zeitschrift für Vergleichende Rechtswissenschaft 109:348-361.
139. Pistor, Katharina. 2013. A Legal Theory of Finance. Journal of Comparative Economics 41:315-330.
140. Reiman, Mathias. 2002. The Progress and Failure of Comparative Law in the Second Half of the Twentieth Century. American Journal of Comparative Law 50:671-700.
141. Reimann, Mathias, and Reinhard Zimmermann. 2006. The Oxford Handbook Of Comparative Law. Oxford: Oxford University Press.
142. Ríos-Figueroa, Julio, and Jeffrey Staton. 2012. An Evaluation of Cross-National Measures of Judicial Independence. Journal of Law, Economics, and Organization 30:104-137.
143. Rodrik, Dani. 2009. The New Development Economics: We Shall Experiment, But How Shall We Learn? In What Works In Development? Thinking Big And Thinking Small, Jessica Cohen and William Easterly eds., pp. 24-47. Washington, DC: Brookings Institution Press.
144. Rodrik, Dani. 2012. Why We Learn Nothing from Regressing Economic Growth on Policies. Seoul Journal of Economics 25:137-151.
145. Roe, Mark. 2006. Legal Origins, Politics, and Modern Stock Markets. Harvard Law Review 120:460-527.
146. Roe, Mark. 2007. Juries and the political economy of legal origin. Journal of Comparative Economics 35:294-308.
147. Rosenthal, Howard, and Erik Voeten. 2007. Measuring Legal Systems. Journal of Comparative Economics 35:711-728.
148. Schwartz, Richard, and James Miller. 1964. Legal Evolution and Societal Complexity. American Journal of Sociology 70:159-169.
149. Sekhon, Jasjeet. 2004. Quality Meets Quantity: Case Studies, Conditional Probability, and Counterfactuals. Perspectives on Politics 2:281-293.
150. Shulztiner, Doron, and Guy Carmi. 2014. Human Dignity in National Constitutions: Functions, Promises and Dangers. American Journal of Comparative Law 62:461-491.
151. Siems, Mathias. 2014. Comparative Law. Cambridge: Cambridge University Press.
152. Siems, Mathias. 2015. Taxonomies And Leximetrics. In The Oxford Handbook of Corporate Law and Governance, Jeffrey Gordon and Wolf-George Ringe. Oxford: Oxford University Press. In Press.

153. Sims, Christopher. 2010. But Economics is Not an Experimental Science. Journal of Economic Perspectives 24(2):47-68.
154. Slater, Dan, and Daniel Ziblatt. 2013. The Enduring Indispensability of the Controlled Comparison. Comparative Political Studies 46:1301-1327.
155. Smith, George Davey, and Shah Ebrahim. 2002. Data Dredging, Bias, or Confounding: They Can All Get You into the BMJ and the Friday Papers. British Medical Journal 325:1437-38.
156. Spamann, Holger. 2008. Appendix to "The 'Antidirector Rights Index' Revisited." Available at http://spamann.net/assets/spamann_adri_revisited_extrafiles.zip?id=374
157. Spamann, Holger. 2009. Large–Sample, Quantitative Research Designs for Comparative Law? American Journal of Comparative Law 57:797-810.
158. Spamann, Holger. 2010a. Contemporary Legal Transplants – Legal Families and the Diffusion of (Corporate) Law. Brigham Young University Law Review 2009:1813-1877.
159. Spamann, Holger. 2010b. Legal Origin, Civil Procedure, and the Quality of Contract Enforcement. Journal of Institutional and Theoretical Economics 166:149-165.
160. Spamann, Holger. 2010c. The 'Anti-Director Rights Index' Revisited. Review of Financial Studies 23:467-486.
161. Spamann, Holger. Forthcoming. The US Crime Puzzle: A Comparative Perspective on US Crime and Punishment. American Law and Economic Review.
162. Spamann, Holger. 2013. Common v. Civil Law, and the Methods of Macro-Comparison. Unpublished Manuscript, Harvard Law School, Cambridge, Mass.
163. Strnad, Jeff. 2007. Should Legal Empiricists Go Bayesian? American Law and Economics Review 7:195-303.
164. Suchman, Mark, and Elizabeth Mertz. 2010. Toward a New Legal Empiricism: Empirical Legal Studies and New Legal Realism. Annual Review of Law and Social Science 6:555-579.
165. Watson, Alan. 1974. Legal Transplants. Edinburgh: Scottish Academic Press.
166. World Bank. 2014. Doing Business 2015: Going Beyond Efficiency. Washington, DC: International Bank for Reconstruction and Development / The World Bank.
167. Ziliak, Stephen, and Deirdre McCloskey. 2007. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. Ann Arbor: University of Michigan Press.
168. Zweigert, Konrad, and Hein Kötz. 1998 [1996]. Introduction to Comparative Law (3rd ed., Tony Weir trans.). Oxford: Oxford University Press.