Model Uncertainty in Sociological Research: An Application to Religion and Economic Growth

Cristobal Young
Princeton University

Model uncertainty is pervasive in quantitative research. Classical statistical theory assumes that only one (true) model is applied to a sample of data. In practice, however, researchers do not know which exact model specification is best. Modern computing power allows researchers to estimate a huge number of plausible models, yet only a few of these estimates are published. The result is a severe asymmetry of information between analyst and reader. The applied modeling process produces a much wider range of estimates than is suggested by the usual standard errors or confidence intervals. I demonstrate this using the work of Barro and McCleary on religion and economic growth. Small, sensible changes in their model specification produce large changes in the results: the results are inconsistent across time, and the instrumental variables strategy suffers from a weak instrument set. Also, the observed relationship between religiosity and economic growth does not hold in the West; it is largely a feature of Asian and African countries and of countries whose data is poor quality. In short, empirical findings should be evaluated not just by their significance but also by their robustness to model specification. I conclude with suggestions for incorporating model uncertainty into practice and improving the transparency of social science research.

A high degree of model uncertainty typically besets statistical research (Bartels 1997; Leamer 1983; Sala-i-Martin 1997; Western 1996). In statistical theory, the "true" causal model for an empirical problem is assumed to be known in advance, although in practice, model specification is a matter of considerable doubt. Indeed, authors usually estimate many models but report only a small number of preferred estimates. As a result, readers are left uninformed about the sensitivity of the results to sensible changes in estimation strategy. Moreover, many findings that seem significant can often be overturned by small, sensible changes in model specification.

In this article, I first articulate the problem of model uncertainty, emphasizing how this leads to much greater variance in statistical results than is suggested by usual standard errors and confidence intervals. Second, I illustrate this by replicating the leading research on religion and economic growth (Barro and McCleary 2003; McCleary and Barro 2006). A thorough sensitivity analysis, along the lines suggested by the model uncertainty perspective, finds that the evidence for a connection between religiosity and economic growth is tenuous at best.

MODEL UNCERTAINTY

In the course of statistical analysis, authors estimate a very large number of models but only ever report a handful of the results. These preferred estimates often reflect only "one *ad hoc* route through the thicket of possible models" (Leamer 1985:308). It is unlikely that reported

Direct correspondence to Cristobal Young at (cristo@princeton.edu). The author wishes to thank Robert Wuthnow, Robert Barro, Paul DiMaggio, Martin Ruef, Chris Sims, Bob O'Brien, David Giles, and *ASR* reviewers for helpful suggestions and constructive criticism. The Center for the Study of Religion at Princeton University provided generous financial support.

results are a random or representative sample of all the estimates the authors calculated (Sala-i-Martin, Doppelhofer, and Miller 2004).

Inferential statistics is about quantifying uncertainty. Scholars use sample data to make inferences about an entire population. We are uncertain about our inferences because any given sample might be unrepresentative. This uncertainty about the data is reported in the form of standard errors and incorporated into confidence intervals and *p* values. In this way, we admit that our results may vary, depending on which sample was used. In practice, we also have uncertainty about the model, derived from several factors. In this regard, Leamer (1983:37–38) describes the process of model selection:

> Sometimes I take the error terms to be correlated, sometimes uncorrelated; . . . sometimes I include observations from the decade of the fifties, sometimes I exclude them; sometimes the equation is linear and sometimes nonlinear; sometimes I control for variable z, sometimes I don't.

It is never easy to say exactly which bundle of model attributes is the single best package. Hence, we have model uncertainty. And often, changes in model specification produce nontrivial changes in parameter estimates. In short, model uncertainty leads to variation in estimates that typically goes unreported.

Statistical theory lags behind practice in this arena. In the early days of statistics, it was probably true that only one model was estimated for a given dataset. Calculations were done by hand, and "computers" were people who carried out the thousands of tedious data calculations (Grier 2005). By the early 1970s, most statisticians had access to mainframe systems, to which they could feed their data and statistical code using punch cards and retrieve the results within a week. Estimation was resource intensive and had to be carefully planned in advance. There was still model uncertainty, but the sensitivity of the results was not known. Limited computational ability was, in essence, a "veil of ignorance" (Rawls 1971). Neither analyst nor reader had any knowledge of the tenuousness of their estimates.

The revolution in computing power has led to a growing information asymmetry; it has elucidated model uncertainty for the analyst but not for the reader. Journal space does not permit the reporting of the hundreds or thousands of different statistical models that are estimated in the course of modern data analysis. Moreover, modern statistics lacks the formal procedures that would "more honestly reflect the observed variation of results" (Bartels 1997:643; Chatfield 1995).

A modest formalization of this may help illuminate the discussion. Classical statistics takes the view that there are K possible samples $\{S_1, \ldots, S_K\}$, each of which yields a unique estimate $\{b_1, \ldots, b_K\}$ for the unknown parameter β. The information in a given sample S_K fully determines the resulting estimate b_K. However, we do not treat any resulting b_K as a definitive value for β because the estimates vary from sample to sample. In repeated sampling, the mean of the estimates is denoted as $\bar{b}$, and the variance is

$$V_S = \frac{1}{K}\sum_{k=1}^{K}(b_k - \bar{b})^2.$$

This sampling variance (V_S), or standard error $(\sqrt{V_S})$, indicates how much an estimate is expected to change if we take a new sample. Although scholars rarely undertake repeated sampling, inferential statistical exercises normally provide an estimate of V_S, which is used to gauge the extent to which b_K is a reliable estimate of β.[1]

Although sampling variance is the foundation of inferential statistics, it is by no means the only source of variation in our estimates. For any given sample, there is also a range of plausible models $\{M_1, \ldots, M_J\}$ that could be applied to the data, each of which will yield its own unique estimate $\{b_1, \ldots, b_J\}$. The average of these estimates is denoted as $\bar{b}$, and the variance of the estimates is

$$V_m = \frac{1}{J}\sum_{j=1}^{J}(b_j - \bar{b})^2.$$

We can think of V_m as the model variance; the square root of V_m might be called the "cross-specification standard deviation" (Granger and Jeon 2004:332). Because researchers generally do not provide an estimate of the model variance (V_m), we end up treating any given estimate

[1] The general rule is if $\frac{b_k}{\sqrt{V_S}} > 2$ then b_k is deemed reliable.

b_j as definitive (given the sample). In many applications, however, model variance may be much larger than sampling variance.

To fully measure the overall variance of our estimates, we need to take each sample $\{S_1, \ldots, S_K\}$ and estimate all plausible models $\{M_1, \ldots, M_J\}$, yielding $K \times J$ estimates b_{kj}. We then take the mean of these estimates ($\bar{\bar{b}}$) and compute the total variance as

$$V_t = \frac{1}{KJ}\sum_{k=1}^{K}\sum_{j=1}^{J}(b_{\mathrm{kj}} - \bar{\bar{b}})^2.$$

This expression for V_t captures all the possible sources of variation in our estimates and includes all reasons why different competent statisticians might arrive at different conclusions: they either used a different sample, used a different model, or both. When reporting just a few preferred estimates, authors should report (and construct *t* statistics and confidence intervals using) total variance, rather than just the sample variance. Hunter and Schmidt (2004:205–06) contrast confidence intervals—based on sampling variance—with "credibility intervals" that are based on estimates of total (sampling + model) variance.

There are two limiting conditions under which model variance does not matter. One condition is when the "true" model is known, as in the classical assumptions of statistics (i.e., the Gauss-Markov theorem). In this case, the set of plausible models $\{M_1, \ldots, M_J\}$ actually contains only one element (the "true" model); all rival model specifications are regarded as incorrect and misleading. There is neither model uncertainty nor model variance. This is an untestable assumption that few authors would explicitly claim. An alternative condition is that the whole set of plausible models yields the same estimate ($b_{\mathrm{j}} = \bar{b}$, $\forall\ b_1, \ldots, b_J$) so that model variance is zero. This is the tacit assumption underlying most statistical papers. However, this condition (equivalent estimates across the set of plausible models) is testable and provides the central motivation for the rest of this article. Ultimately, the task is to use some form of model sensitivity analysis to place proper confidence intervals around our estimates.

EMPIRICAL EVIDENCE OF MODEL UNCERTAINTY

Classical statistics focuses on the tractable problem of sampling uncertainty. Much evidence suggests, however, that the problem of model uncertainty may be more serious. I draw on three sources of evidence: meta-analyses of existing literature, Bayesian work on model averaging, and econometric field experiments on the diversity of modeling strategies. The evidence suggests that sampling variance is a small component of the total variance in estimates. Using different models leads to much greater variation than does using different samples.

Meta-analysis, like replication, is rare in sociology. The meta-analytic studies available, however, consistently comment on the high degree of excess variation—variance across studies that is not well explained by sampling error (Hsieh and Pugh 1993; Pinello 1999; Stith et al. 2000). Observed variation in results is often many times larger than the standard errors reported in the literature. For example, Pinello (1999) collected 66 comparable estimates of the effect of political party affiliation (Democrat versus Republican) on judges' decision making. Although the average standard error from the studies is .034, the standard deviation across estimates is .275—more than eight times as large. Sampling uncertainty accounts for only 11 percent of the total variation in the estimates. Model specification evidently accounts for the remaining 89 percent.

Bayesian Model Averaging (BMA) offers another approach to examining model uncertainty. The general philosophy of the Bayesian approach is that reporting a single "best" estimate makes little sense (because one is never completely sure which model is "best") (Hoeting et al. 1999; Western 1996). Analysts should run all statistical models that they regard as reasonable. Furthermore, if a model is a priori reasonable enough to estimate, then the results should be reasonable enough to report.

Sala-i-Martin (1997) developed a BMA strategy for macro-level research where the list of social, political, and economic variables is often long, while the datasets are small. The flavor of the approach is well captured in the title of the article, "I Just Ran Two Million Regressions." In a follow-up article, Sala-i-Martin and colleagues (2004) estimate 89 million regressions. Their strategy is to start with 67 variables that

Table 1. Estimates of Model Variance and Sampling Variance

	Field Experiment				Meta-Analysis		
	USA Budget Survey Datasets						
	1941	1950	1960	1972	Minimum Wages	Returns to Education	Judicial Ideology
Average coefficient	.62	.57	.58	.48	–.19	.079	.277
Range	.52, .74	.38, .75	.38, .73	.33, .71	–19.3, 4.8	na	–.16, .87
Cross-specification SD	.059	.119	.125	.143	1.096	.036	.275
Average SE	.022	.017	.015	.017	.160	.015	.034
Ratio: SD / SE	2.7	7.2	8.2	8.6	6.9	2.4	8.1
N studies	6	7	7	7	65	27	36
N estimates	9	10	10	10	1492	96	66
Sqrt(total variance)	.080	.136	.140	.159	1.256	.051	.309
Model portion	73%	88%	89%	90%	87%	71%	89%
Sampling portion	27%	12%	11%	10%	13%	29%	11%

Notes: Field experiment data are from Magnus and Morgan (1999). Eight teams of econometricians volunteered to independently analyze large sample datasets on consumer demand. Results are reported on pp. 73, 106–07, 171, 185, 208, 242, 253, and partially summarized on p. 289. The meta-analysis data are from Doucouliagos and Stanley 2008 (minimum wages), Ashenfelter, Harmon, and Oosterbeek 1999 (returns to education), and Pinello 1999 (judicial ideology).

the published literature on economic growth has found to be significant. They then cycle through millions of possible combinations of these variables and take a weighted average of the OLS regression results. Of the 67 previously "significant" variables, only 18 are "robust." Three more are marginal, and the remaining 46 are weak, having little explanatory power and consistently imprecise estimates. Some are significant in only 10 or 15 out of 1,000 specifications.

This emphasizes the scope that authors have had to publish papers reporting significant findings that are not robust. Almost 70 percent of the variables found to be significant in the growth literature are not robust in a substantial BMA analysis.[2] The results are sobering and highlight the need for rigorous checks on the conclusions of statistical research.

Additional evidence of model uncertainty comes from an econometrics field experiment (Magnus and Morgan 1999). Eight teams of econometricians volunteered to independently analyze several (large sample) datasets on consumer demand. If experienced statisticians tend to gravitate toward the same (best) model specification, these teams would produce similar results. Yet, the diversity of results generated within the scenario above is striking. Table 1 shows the experimental results for the four main datasets. The "cross specification standard deviation" ranges from 2.7 to 8.6 times the magnitude of the sampling standard errors. If we simply add up the model variance and the sampling variance in this study, then the standard errors represent only 10 to 27 percent of the variation in estimates.

For comparative purposes, Table 1 also reports findings from three meta-analyses: employment effects of minimum wages (Doucouliagos and Stanley 2008), the returns to education (Ashenfelter, Harmon, and Oosterbeek 1999), and the effect of party affiliation on judicial decisions (Pinello 1999). These figures are harder to interpret since the dataset is not constant across the meta-analysis estimates. Nevertheless, the conclusions here are much the same. Uncertainty about statistical estimates "mostly arises from the use of alternative model specifications" (Granger and Jeon 2004:332). Based on Table 1, a reasonable rule of thumb is that actual variation across estimates will be about five times the reported standard error.

[2] One caveat should be made. The BMA analysis uses only repeated applications of simple OLS regression. This leaves out a variety of reasonable models in which variables may prove to be more consistently significant. In short, this exercise only shows robustness with respect to the choice of control variables.

Model uncertainty offers researchers substantial leeway to select preferred results from a large menu of model specifications. As a result, the published literature likely contains many nonrobust conclusions. To have confidence in our results, it is not necessary that our conclusions hold in every specification. Nevertheless, given the ubiquity of model uncertainty, standard errors and *p* values patently fail to provide reasonable bounds of confidence around statistical estimates. A significance test should be seen only as an initial starting point; surviving a critical and rigorous sensitivity analysis is a better measure of a strong result. In short, statistical findings should be evaluated as much by their robustness as by their significance.

In the following section, I apply the robustness principle to Barro and McCleary's work on religion and economic growth. This research is a fitting case study for a number of reasons. First, it is a very sophisticated piece of research, employing cutting-edge statistical models that are not well known to sociologists. Second, the article has high visibility; it was recently published in the *American Sociological Review* and its lead author is one of the most widely cited authors in economics today (Kim, Morse, and Zingales 2006). It is important to show that top level competence, sophistication, and screening, are not, in themselves, a solution to the problem of model uncertainty. Finally, the work should be of some intrinsic interest to sociologists given the classical attention to religion and economy.

RELIGION AND ECONOMIC GROWTH

At least since Weber, the relationship between religion and economy has been the source of sociology's most famous theories. Weber argued that the unique theology of Protestantism nurtured a capitalist orientation in which saving, investment, and hard work were valued for their own moral content. The tremendous economic growth of Protestant countries after the reformation, vis-a-vis the rest of the world, was an unintended consequence of Protestant beliefs (Weber 1930). The Weberian thesis has been a subject of great controversy; Barro and McCleary add a fresh perspective by stepping back from the historical debate and focusing on modern economic growth. Influenced by the more recent "trust" literature (e.g., Fukuyama 1995), they focus on religiosity per se, leaving specific religious traditions in the background. Do religious belief and attachment have an influence on economic growth? Their view is that all the major world religions, at least in sacred texts, praise the social virtues of hard work, frugality, and honesty (McCleary 2007). The key distinction is not Protestant versus Catholic (or Buddhist or Muslim or Confucian), but religious versus irreligious, believer versus nonbeliever. Economic prosperity, at some level, depends on social and moral norms that religiosity helps nurture. This is an argument Weber never addressed, but in a world where secularism has become much more salient, it is an important one.

Barro and McCleary find that religiosity matters for economic growth, but it matters in unexpected ways. Two measures of religiosity (church attendance and belief in hell) are significant, but they have opposite signs. Church attendance is "bad" for economic growth, but belief in hell is "good." What matters, they conclude, is the tradeoff between the two: the efficiency of the religious sector.

What does efficiency mean in the context of the religious sector? Suppose that 30 percent of a population believes in hell. If only 20 percent of the population regularly attends church, the religious sector seems quite efficient: more people believe than attend. This kind of efficient religion is good for the economy. If 40 percent of the population attends church, however, the religious sector seems inefficient. More people attend than actually believe. This church inefficiency drags down a nation's economic performance.

Religious beliefs, in this framework, nurture economic growth by instilling people with the moral values (honesty, work ethic, and frugality) that support growth. High levels of church attendance, on the other hand, impair economic growth by diverting resources from more productive uses. In Barro and McCleary's theory, churches act much like a tax-and-spend government. The religious sector taxes the economy by demanding church attendance.[3] It stim-

[3] This may be an indicator of broader resource consumption. Note, however, that even the United States' high church attendance levels only translate into an average of .9 hours per week of formal religious activities (Robinson and Godbey 2000:176)—less than 1 percent of waking hours.

ulates the economy by creating religious belief. If the religious sector imposes a high tax (attendance) but generates a small stimulus (religious belief), then it drags down the economy.

DATA AND MODEL

The dependent variable in this study is the average growth rate of real per capita GDP over three periods, 1965 to 1975, 1975 to 1985, and 1985 to 1995. The main variables of interest are church attendance and belief in hell. Church attendance is measured as the share of the population that attends religious services at least once a month. Belief in hell captures the fraction of the population that says they believe in hell. These are measured only once, primarily in 1990. Both variables are expressed in the form $[x / (1 - x)]$. Also included is the percent of the population affiliated with major religions: Catholic, Protestant, Orthodox, other Christian, Muslim, Jewish, Hindu, eastern religion (including Buddhist), and other religion.[4]

Barro and McCleary generously provided the dataset after my efforts to build it from publicly available sources were unsuccessful. The public datasets from the World Bank and the Penn World Tables contain a nontrivial amount of missing data, which Barro supplemented from other sources.

The data are used to estimate a system of three equations,

$$\Delta Y_{i1} = \alpha_1 + \boldsymbol{X}_1\boldsymbol{\beta}' + \varepsilon_{i1}$$
$$\Delta Y_{i2} = \alpha_2 + \boldsymbol{X}_2\boldsymbol{\beta}' + \varepsilon_{i2}$$
$$\Delta Y_{i3} = \alpha_3 + \boldsymbol{X}_3\boldsymbol{\beta}' + \varepsilon_{i3}$$

where the subscripts $\{1, 2, 3\}$ refer to the time periods {1965 to 1975, 1975 to 1985, 1985 to 1995}.[5] Notice that the vector of coefficients, $\boldsymbol{\beta}'$, is the same in each equation (meaning that variables are constrained to have the same coefficient in each time period). These equations could be stacked and estimated using ordinary least squares (OLS). The concern is that the error terms of the equations may be correlated, thereby violating the classical assumptions of OLS and generating biased estimates. A solution to this problem is to use the "seemingly unrelated regressions" (SUR) model, which accounts for possible correlation in the error terms. In practice, there is little correlation across the error terms, and with these data the OLS and SUR models generally produce very similar estimates (Barro 1997:15). For readers unfamiliar with SUR, it may be simpler to think of the estimator as a variant of OLS.

REVERSE CAUSATION AND INSTRUMENTAL VARIABLES

Barro and McCleary embrace a sophisticated theoretical model in which there may be two very different effects: religiosity influences economic growth (the subject of this article), but growth may also influence religiosity—the classic secularization effect. As economies industrialize and modernize, people become both richer and less religious (Berger 1967). Secularization does not refute a causal effect of religiosity on economic growth, but it does create a difficult identification problem.

It is thus necessary to make the hypothesis more specific: Countries that *for some noneconomic reason* have a greater stock of religious belief will enjoy greater economic growth. How can we identify distinctively noneconomic reasons why some countries are more religious than others? Barro and McCleary's answer is to use instrumental variables (IV) regression. Instrumental variables is a popular technique in economics but is rarely used in sociology.[6]

The IV strategy is an attempt to re-create the conditions of a controlled experiment. An instrument is some variable that is thought (or hoped) to provide a degree of random assignment. A good instrument, in this case, is something that causes differences in religiosity across countries

[4] These data were measured in 1970 using the *World Christian Encyclopedia.* Hsu and colleagues (2008) provide a useful assessment and critique of this data source.

[5] This is a "between effects" model. Identification comes entirely from the cross-sectional variation in growth rates. Longitudinal variation across periods is not captured (this is not technically a problem for the religiosity estimates because these variables enter the model as constants).

[6] In sociology, a more common approach to exploring causality uses structural equation modeling (SEM). References to SEM in the instrumental variables literature are virtually nonexistent. A comparative analysis of the two approaches would be useful.

but has no direct relationship with economic growth. A hypothetical experiment might be to randomly select certain countries to be the targets of very large-scale missionary projects (that have at least some success in raising religious adherence). The instrumental variable in this case would be the dummy variable indicating whether or not a country was (randomly) selected for a mission. One could then test if the extra religiosity caused by the instrument does indeed have an effect on growth.

To do this, one first creates (via standard regression techniques) a new variable made up of just the variation in religiosity caused by the instrument (such as random missions). The idea is that this new variable is free of any reverse causation problems. This new variable (religiosity caused by random selection to a missionary project) is then used to see if religiosity has any causal effect on economic growth.

To make this more explicit, consider the equation,

$$\Delta Y_i = \alpha + \beta X_i + \boldsymbol{W}\boldsymbol{\gamma}' + \varepsilon_i \qquad (1)$$

in which $\boldsymbol{W}$ is a vector of exogenous controls and X_i is an endogenous variable (religiosity) correlated with ε_i so that the OLS estimator of β is biased and inconsistent. To correct this, we find some instrument Z_i (random assignment to missions) that influences X_i but is uncorrelated with ε_i. This variable can be used to strip out the portion of X_i that is problematic, creating a new variable $\hat{X}_i$ that allows for estimation of β that is unbiased in large samples. Using regression decomposition on the religiosity variable, we estimate

$$X_i = \pi_0 + \pi_1 Z_1 + \boldsymbol{W}\boldsymbol{\gamma}' + \upsilon_i \qquad (2).$$

The fitted values from this regression are $\hat{X}_i = (\pi_0 + \pi_1 Z_1 + \boldsymbol{W}\boldsymbol{\gamma}')$, which represents the portion of X_i that is determined by the instrument Z_i (and other exogenous variables $\boldsymbol{W}$). The residual υ_i represents everything else that generates X_i. Note that if economic growth (ΔY_i) has a direct effect on religiosity (X_i) (i.e., there is a secularization effect), this is an omitted variable in Equation 2 and ΔY_i ends up in the error term (υ_i). If the instrument is a good one, then any reverse causation patterns are safely bracketed off in υ_i, effectively stripped out of the new variable $\hat{X}_i$. Finally, we return to our original growth regression, using the "clean" $\hat{X}_i$ in place of X_i:

$$\Delta Y_i = \alpha + \beta \hat{X}_i + \boldsymbol{W}\boldsymbol{\gamma}' + \varepsilon_i \qquad (3).$$

This is the two stage least squares (2SLS) estimator (or, when combined with SUR in a system of equations, the three stage least squares [3SLS] estimator). In this equation, if reverse causation were ever a problem (i.e., if growth causes religiosity), its influence is removed in Equation 2, so that $\hat{X}_i$ represents, in effect, randomly different levels of religiosity.

Using instrumental variables is a potentially powerful estimation technique and its logic is very appealing, but the econometric literature emphasizes its limitations (Bound, Jaeger, and Baker 1995; Hahn and Hausman 2003; Staiger and Stock 1997; Stock and Yogo 2005). While it is always possible to estimate an IV model, in many cases "the cure can be worse than the disease" (Bound, Jaeger, and Baker 1993). In IV estimation, everything hangs on the validity of the instrument. There are two criteria for a valid instrument: (1) relevance (non-weakness), the instrument causes (nontrivial) changes in the regressor of interest (religiosity); and (2) exogeneity (nonendogeneity), the instrument has no direct relationship with either the dependent variable (GDP growth) or any omitted variables that influence GDP growth. In short, the instrument does not belong in the "true" growth regression.

Instrument relevance can be empirically tested, but instrument exogeneity is ultimately an untestable assumption (because the "true" model, or the presence of omitted variables, can never be known with certainty).[7] If the instrument is invalid, one is usually better off using OLS than IV. 2SLS with weak or endogenous instruments is biased on average toward the OLS estimator. However, 2SLS is also an erratic estimator in these cases, yielding very wide variation in estimates (sometimes even a bi-modal distribution of estimates) (Nelson and Startz 1990; Stock, Wright, and Yogo 2002).

Barro and McCleary use three instruments. The first, and most promising, is state religion. Some countries—hundreds of years ago—established official state religions. If this has any effect on economic growth, it is likely because these countries are generally more religious

[7] The over identifying restrictions test inspects instrument exogeneity, but it still requires the assumption that at least one instrument is exogenous.

Table 2. Regressions for Economic Growth, 1965 to 1975, 1975 to 1985, and 1985 to 1995

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	3SLS	Drop Shares	Drop Religiosity	Drop Hell	Drop Attendance	SUR
Religiosity						
Attendance	–.0147**	–.0054**		–.0047		–.0056**
	(.0049)	(.0019)		(.0035)		(.0019)
Belief in Hell	.0152**	.0038*			.0049	.0067**
	(.0052)	(.0016)			(.0034)	(.0022)
Religion Shares						
Protestant	–.0161		–.0051	–.0120	–.0015	–.0089
	(.0083)		(.0047)	(.0074)	(.0052)	(.0051)
Muslim	–.0484**		.0031	–.0010	–.0093	–.0176*
	(.0184)		(.0050)	(.0067)	(.0105)	(.0085)
Hindu	–.0236		–.0227*	–.0258*	–.0121	–.0149
	(.0144)		(.0111)	(.0126)	(.0119)	(.0105)
Eastern Religion	–.0290		.0489***	.0416*	.0441**	.0225
	(.0278)		(.0127)	(.0169)	(.0155)	(.0152)
Orthodox	–.0106		.0067	.0022	.0075	.0020
	(.0118)		(.0077)	(.0094)	(.0076)	(.0075)
Jewish	.0045		.0167	.0090	.0203	.0120
	(.0143)		(.0102)	(.0125)	(.0103)	(.0099)
Other Christian	–.0225		–.0068	.0050	–.0271	–.0198
	(.0233)		(.0144)	(.0184)	(.0196)	(.0159)
Other Religion	.0119		.0166	.0034	.0386	.0242
	(.0263)		(.0165)	(.0237)	(.0217)	(.0187)
R^2 values for each time period	.61, .62, .35	.58, .63, .35	.68, .50, .61	.67, .56, .52	.67, .53, .55	.68, .60, .52

Notes: Dependent variables are the rates of real per capita GDP growth over the periods 1965 to 1975, 1975 to 1985, and 1985 to 1995. Each system includes a total of 153 observations: 48 countries in period 1, 53 in period 2, and 52 in period 3. Models include time period averages for: trade openness (filtered for scale effects of geographical size and population), terms of trade growth (interacted with trade openness), consumer price inflation, the ratio of investment over GDP, rule of law, and electoral rights and its square. The models also include (but do not report) start-of-period variables for: education (years of male secondary and higher school attainment), life expectancy at age one, fertility, and per capita GDP. Estimation is by three stage least squares. The instruments are state religion; state regulation of religion; religious pluralism; dummy variables for being a colony of Britain, France, Spain/Portugal, and other; and lags of per capita GDP, electoral rights, and its square. In Model 2, religion shares are also used as an instrument. Model 6 is estimated by seemingly unrelated regressions.
* $p < .05$; ** $p < .01$; *** $p < .001$ (two-tailed tests).

today.[8] Another instrument—closely related to the first—is state regulation of religion (such as the political power to appoint or dismiss top church officials). The third instrument is religious pluralism—the substantive diversity of offerings in the market for religion. This is the most questionable instrument because rich countries tend to "import" religious diversity from the poorer regions of the world. "Differences in per-capita income are a key determinant of the size and direction of migration flows" (Borjas 1999:18), suggesting that growing economies attract population from poor countries that often have different religious traditions. Rather than a purely exogenous instrument, religious pluralism is in part a result of GDP growth. In short, Barro and McCleary offer two promising instruments, while the third is somewhat doubtful. After replicating the basic results, the next step will be to test the validity of the instruments.

BASELINE RESULTS

Table 2 replicates the original Barro and McCleary results (Barro and McCleary 2003; McCleary and Barro 2006). In Model 1, the

[8] Note that this empirical finding contradicts the economics of religion literature, which expects that a state religion leads to waning religious adherence (e.g., Finke and Stark 1992).

Table 3. First-Stage Regressions for Instrumental Variables

	Church Attendance	Belief in Hell
State religion	.2027	.3830**
	(.1628)	(.1283)
Regulation of religion	.1122	.0523
	(.1242)	(.0978)
Religious pluralism	.4941	.4861
	(.4430)	(.3491)
Partial R^2	.013	.016
F statistic	2.39	8.78
R^2 values for each time period	.83, .79, .81	.94, .92, .93

Notes: The first stage regression includes the full instrument list. The exogenous variables from the second-stage regression are: start of period per capita GDP, fertility, and life expectancy; average education, investment ratio, trade openness, terms of trade growth, rule of law, and electoral rights and its square. Inflation is excluded from the first-stage regression. Dummy variables for being a colony of Britain, France, Spain/Portugal, and other are considered by Barro and McCleary as instruments for inflation (thus included in the first-stage regression), although they also serve as (significant) instruments for church attendance and belief in hell.
** $p < .01$ (two-tailed tests).

estimated effect of church attendance is negative for economic growth (–.0147), while the effect of belief in hell is positive (.0152). Hence, the key factor seems to be the efficiency of the religion sector—that is, the positive effect of belief after subtracting out the negative effect of church attendance. When the religiosity variables are entered individually, neither are significant (Models 4 and 5). "If church attendance and religious beliefs move together in their usual manner, the overall relation with economic growth tends to be weak" (Barro and McCleary 2003:777). While attendance and belief are strongly correlated ($r \approx .65$), divergences between the two appear to have an effect on growth.

In Model 1, the only religious tradition that is significant for growth is percent Muslim (–.0484), which is associated with weaker economic performance. This result, however, is not robust. In Models 3, 4, and 5, the Muslim share is small and nonsignificant, while the Hindu and Eastern religion shares are (generally) significant. In any event, there is no evidence of a Protestant ethic supporting economic growth; the coefficient for Protestant share is consistently negative and nonsignificant. In short, it does not seem to consistently matter, in these data, whether a country is more Catholic, Protestant, Muslim, Orthodox, Hindu, or Buddhist.

SENSITIVITY ANALYSIS I: INSTRUMENT VALIDITY

In their survey paper on "Growth Econometrics," Durlauf, Johnson, and Temple (2006:638) write that "many applications of instrumental variable procedures . . . [are] undermined by the failure to address properly the question of whether these instruments are valid." Instrument validity is a problem in this study.

The standard test for instrument relevance (non-weakness) is built on the partial R^2 of the instruments in the first-stage regression (Equation 2 above). The rule of thumb given by Staiger and Stock (1997) is that the first-stage F statistic should be greater than 10. Stock and Yogo (2005: Table 5.1) provide more precise critical values, ranging from 9.08 (for three instruments) to 11.52 (for 12 to 14 instruments). It is important to recognize that while this test uses the F statistic, it is not a simple test of joint significance as the test has different critical values than what are found in an F table.[9]

The first-stage regressions, where the relevance criteria are established, are shown in Table 3. The results are disappointing. For

[9] Stock and Yogo (2005) also note that when there is more than one endogenous variable, the best test of weak instruments is the Cragg-Donald statistic, which is based on the matrix of F statistics. Unfortunately, Cragg-Donald is not yet implemented for systems of equations (i.e., for the 3SLS estimator).

church attendance, the coefficients on state religion, regulation of religion, and religious pluralism are scarcely larger than their standard errors. The partial R^2 of the three instruments is .013. This means that the IV strategy uses only 1.3 percent of the variation in church attendance to identify its causal effect on growth. If the sample size were notably larger, this might leave "enough" variation for reliable identification of the parameters.[10] However, the *F* statistic is 2.34, far below the critical value of 9.

For belief in hell, state religion achieves significance at the 1 percent level, while regulation of religion and religious pluralism are nonsignificant. The partial R^2 is .016 and the *F* statistic is 8.78. While this result is better than for church attendance, the *F* statistic still indicates a weak instrument set.

This is partly a "too many instruments" problem—which contributes to finite sample bias in the IV estimator (Hahn and Hausman 2003). The overall strength of the instrument set will usually improve if the weakest instruments are dropped. In fact, the partial R^2 is driven almost entirely by the state church instrument, with state regulation and religious diversity making little contribution. The *F* statistics for state church are 7.8 for the church attendance equation (still weak) but 26.3 for the belief in hell equation (much greater than the critical value). State church is thus strong at least as an instrument for belief in hell. This is not a solution, however, because with two endogenous variables we need at least two instruments to identify the system. There is still no instrument for church attendance. The weakest instrument seems to be state regulation. If this term were dropped (leaving state church and religious pluralism as instruments), the *F* statistics are 13.5 for belief in hell but still only 3.9 for the church attendance equation. Better practice would be to drop religious pluralism because this variable is a priori doubtful as an exogenous instrument. This set (state church and state regulation) yields similar, although somewhat smaller, *F* statistics (12.3 for belief in hell but only 1.3 for church attendance). In short, there is no set of non-weak instruments available for both endogenous variables.

What does the instrumental variables procedure do in practice? Table 2 offers a comparison of the IV model (Model 1) with the non-IV SUR estimator (Model 6). Compared with the SUR model, the IV model generates estimates that are larger in magnitude. The effect of IV estimation is to push the coefficients away from zero: the negative coefficient becomes more negative, and the positive coefficient becomes more positive. This is important to emphasize because it is not what one would expect if the IV were correcting for reverse causation.

Reverse causation (secularization) would be expected to have a consistent effect on both coefficients. The presumption is that economic growth lowers religiosity (for both religious belief and church attendance). This creates a negative correlation between growth and religiosity that the IV method is attempting to remove. If the IV were correcting for reverse causation, it should increase both coefficients—in other words, it should *add a positive value* to both. The church attendance coefficient should move closer to (not further from) zero.[11]

Evidently, the instrumental variables model is not correcting for reverse causation. This is not surprising given that the instrument set is weak. The IV strategy has made the religiosity estimates larger in magnitude. Yet 2SLS (or 3SLS) IV models based on weak instruments are prone to erratically biased estimates. There is an alternative IV technique—the Limited Information Maximum Likelihood (LIML) estimator—that is more robust to weak instruments (Stock et al. 2002). It is not available, however, for estimating systems of equations.[12]

[10] A sample size of 510 for these formula inputs would yield an *F* statistic of 9.08. In short, the sample size (154) would need to be more than three times larger (other things being equal). Note that this also assumes that the instruments are perfectly exogenous.

[11] In personal correspondence, Robert Barro raised this point himself. Barro also suggested that the IV procedure could be correcting for (classical) measurement error, thus removing some attenuation bias. However, because the instrument set is weak, it can hardly be correcting for a problem of white noise in the religiosity variables.

[12] In principle, the Full Information Maximum Likelihood (FIML) estimator is the appropriate counterpart to LIML for systems of equations. The literature has not yet shown, however, that FIML is preferred for the case of weak instruments, nor is it readily implemented in the statistical packages used here (Eviews and Stata).

Table 4. Regressions for Economic Growth, by Period

	Model 7: IV – 2SLS			Model 8: OLS			Model 9: IV – LIML		
Period:	1965–75	1975–85	1985–95	1965–75	1975–85	1985–95	1965–75	1975–85	1985–95
Religiosity									
Attendance	–.0110	–.0195	.0161	–.0017	–.0099*	–.0011	.0038	–.0299†	.0621
	(.0129)	(.0146)	(.0194)	(.0046)	(.0039)	(.0037)	(.0359)	(.0180)	(.0883)
Belief in Hell	.0109	.0207	–.0036	.0020	.0118*	.0058	.0096	.0284†	–.0380
	(.0119)	(.0123)	(.016)	(.0052)	(.0048)	(.0042)	(.0278)	(.0151)	(.0765)
N of observations	48	53	52	48	53	52	48	53	52
R^2	.74	.65	.42	.78	.74	.74	NA	NA	NA

Notes: For general model details, see the note to Table 1. Estimation in Model 7 is by two-stage least squares. Model 9 is by limited information maximum likelihood estimation.
† $p < .10$; * $p < .05$ (two-tailed tests).

Furthermore, in small samples, LIML may not be much better than 2SLS at coping with the weak instrument problem. In the next section, I calculate LIML estimates for each equation and compare them with 2SLS and OLS.

SENSITIVITY ANALYSIS II: DECADE-BY-DECADE REGRESSIONS

Another serious problem for this study is missing data. The religiosity measurements are only available for (approximately) 1990. The model jointly estimates three equations, however, one for each of the three time periods. For the first two time periods (1965 to 1985), the data for religiosity are missing. These missing values are filled in using religion data collected in period 3 (1985 to 1995). The underlying assumption is that religiosity is constant over time. Barro and McCleary (2003:772–74) write that "church attendance and religious beliefs exhibit a lot of persistence over time. Hence later values may proxy satisfactorily for the earlier missing ones." This assumption may not hold well.

Iannaccone (2003:26) has constructed long-term church attendance data for 32 countries that show "widespread secularization" during the twentieth century.[13] During the period of this study, church attendance declined on average by 7 percentage points, with a range of +3 to –21. Only two countries show no change in church attendance, and almost a third of the sample shows declines of 10 percentage points or more. The trend of secularization has been of substantial magnitude overall and has varied widely across countries.

The baseline (conservative) treatment of missing data is listwise deletion—drop any observation with missing values (Allison 2001). This would leave only one equation. Period 3 (1985 to 1995) does not have a missing data problem, so the results are free of any possible imputation bias. One way to check if the treatment of missing data is reasonable is to see if the results are consistent over time. Do the results from earlier periods simply reinforce the findings from period 3, or do they notably disagree?

In Table 4, I estimate a model for each time period separately, using OLS, 2SLS, and LIML. In the 2SLS model, basically nothing is statistically significant. This is expected, given the small sample size (low power) in each wave. The magnitudes and signs of the coefficients, however, show that period 2 (1975 to 1985) is driving the overall results. Period 3 (1985 to 1995) actually yields opposite results with the 2SLS estimator (the signs are reversed) and essentially zero estimates using OLS.[14] The Limited Information Maximum Likelihood (LIML) IV estimator, which is more robust to weak instruments, yields opposite signs for periods 1 and 3.

Using 2SLS, the coefficients on church attendance are –.0110 for the first decade, –.0195 for

[13] It should be noted, moreover, that Iannaccone has been quite critical of secularization theory.

[14] Note that the SUR model is used only when the time period equations are estimated jointly (when the error terms may be correlated across equations).

Table 5. Countries in the Growth Regression Model

Americas	Asia	Africa	Europe	Oceania
Canada	Bangladesh	Algeria	Austria	Australia
United States	India	Egypt	Belgium	New Zealand
Mexico	Indonesia	Ghana	Cyprus	
	Iran	South Africa	Denmark	
Argentina	Israel	Uganda	Finland	
Brazil	Japan	Zimbabwe	France	
Chile	Jordan		Germany, West	
Columbia	South Korea		Greece	
Dominican Rep.	Pakistan		Hungary	
El Salvador	Philippines		Iceland	
Peru	Singapore		Ireland	
Uruguay	Taiwan		Italy	
Venezuela	Turkey		Netherlands	
			Norway	
			Poland	
			Portugal	
			Spain	
			Sweden	
			Switzerland	
			United Kingdom	

the second, and .0161 for the third. The coefficients on belief in hell show the same pattern (.0109, .0207, and –.0036, respectively). Using OLS, period 2 dominates to an even greater degree. The estimates are trivially small in periods 1 and 3. The estimate for church attendance is nine times larger in period 2 (–.009) than in period 3 (–.001). Regardless of whether one uses OLS, 2SLS, or the more robust LIML IV estimation, the results are strongly driven by period 2 (using religiosity data imputed from period 3). The results that are free of potential imputation bias (period 3) do not support the Barro and McCleary findings. Nor, in general, do the results from period 1. This is not simply due to lower statistical power: standard errors aside, the estimates in other periods are generally either zero or are going in the wrong direction.

SENSITIVITY ANALYSIS III: PARAMETER HETEROGENEITY

Critics of large cross-country regression analyses have asked a pointed question: "What do Thailand, the Dominican Republic, Zimbabwe, Greece, and Bolivia have in common that merits their being put in the same regression analysis?" (quoted in Durlauf et al. 2006:616).

The idea behind cross-country regression is that (roughly) the same processes are at work in each country; the underlying relationship between religiosity and economic growth is assumed to be the same in Iran, Zimbabwe, and the United States. If this assumption fails, there is a problem of parameter heterogeneity.

This problem seems particularly severe when studying religion across a wide collection of countries. Economists generally believe that inflation and investment have common meanings in different countries. Indeed, it seems likely that investment creates jobs and growth no matter where it happens.

Church attendance and belief in God and hell do not, by contrast, have a standardized meaning across the world. Why should one assume that religious belief or church attendance will have the same effect on growth (and thus have the same regression coefficient) in these countries? Religion may stifle the market in some countries (Iran) (Kuran 2004) but support market activity in others (United States) (Lindsay 2007).

There is indeed a disparate group of countries in the dataset, as shown in Table 5. Barro and McCleary have commented on the issue, noting that "the meaning of religious beliefs and the significance of formal religious services vary across religions" (McCleary and Barro

2006:69). For this reason, they argue that it is important to include religion shares (e.g., the percent Protestant) as regressors. This does not, however, solve the problem; it only addresses intercorrelation between religiosity and religion shares. For example, Muslims attend religious services more often than do Protestants, so a predominantly Muslim country will tend to have higher church attendance. But this is a separate issue from parameter heterogeneity. The deeper concern is that the relationship between religion and economy may be entirely different in Muslim and Protestant countries. Adding percent Muslim as a control does not capture this difference. There is still only one coefficient relating church attendance to GDP growth, averaged across Muslim and Protestant countries. To address the issue of parameter heterogeneity, one must refrain from averaging across different religious groups. One must ask more specific questions: What is the relationship between religion and economy in Western countries? Does the same relationship hold in non-Western countries?

I test for parameter heterogeneity in two basic ways. First, I drop non-Western countries (i.e., the Asian and African nations). In most of these countries, religious traditions are very different from (Western) Christianity. In many places, asking people (Christian-centric) questions about whether they believe in hell or attend religious services is not very meaningful.[15] Furthermore, the social context and economic consequences of being highly religious may be very different in Iran or Zimbabwe than in the United States or Germany. As well, in many of the Asian and African countries, the quality of data is particularly poor for both GDP and religiosity. Finally, the dataset has poor coverage of Asia and Africa. It includes 13 out of 50 countries in Asia (26 percent) and only 6 out of 54 countries in Africa (11 percent). There is little reason to think these are representative samples. By excluding the Asian and African countries, one can see if the results hold up in the Western world, where the meaning and significance of religious beliefs and church attendance are more comparable, and where the data quality is generally stronger. In this approach, I restrict the country set to Europe, Oceania, and the Americas (N = 35), excluding Asia and Africa (N = 19). I flush this out by using the full dataset with an interaction model that provides separate estimates for Western and non-Western countries.

Table 6 reports the results. Models 10 and 11 (SUR and 3SLS, respectively) show that when the Asian and African countries are dropped, the coefficients fall almost to zero and the sign on church attendance is reversed. Interaction models reinforce this. The interaction term for church attendance is large and highly significant in the SUR model (Model 14), while the coefficients on church attendance and belief in hell are both negligible. In other words, church attendance has a negative effect on economic growth only in the Asian and African countries. Even in the 3SLS model, where none of the coefficients are significant, the coefficients for attendance and belief are twice as large for Asia/Africa than for the West.

As an alternative approach, I classify the countries by the quality of their data. This is partly captured in the difference between Western and Asian/African countries, but only indirectly. Latin America is Western Christian, but the data on both GDP and religiosity are weak. By contrast, some Asian countries have good data quality (such as Japan, Korea, and Israel). The Penn World Tables—the source of the GDP data—provide assessments of data quality ranging from A to D. I constructed a dummy variable in which those receiving A or B grades are high data quality countries; those with C or D grades are low data quality.

Table 6 shows that in countries where data quality is good, the religiosity variables are small and nonsignificant. The interaction between church attendance and poor data quality is negative, large in magnitude, and significant at $p < .01$ in the SUR model (Model 14); the interaction is equally large but nonsignificant in the IV model. In countries with poor data, there is a consistent negative effect of church attendance. In countries with good data, church attendance has no effect on economic growth.

This result is robust to alternative specifications of data quality. In additional analyses not reported, the religiosity results are strongest for

[15] It is important to note, nevertheless, that many sub-Saharan African countries are now predominantly Catholic or Protestant, although still intermixed with traditional, indigenous beliefs.

Table 6. Interaction Models

	Drop Asia and Africa		Full Interaction Models			
	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15
	SUR	3SLS	SUR	3SLS	SUR	3SLS
Religiosity						
Attendance	.0029	.0045	.0002	–.0078	–.0008	–.0062
	(.0026)	(.0038)	(.0023)	(.0058)	(.0025)	(.0061)
Belief in hell	–.0006	.0003	.0016	.0077	.0022	.0071
	(.0028)	(.0040)	(.0026)	(.0052)	(.0030)	(.0059)
Attendance × Asia/Africa			–.0150***	–.0085		
			(.0035)	(.0067)		
Hell × Asia/Africa			.0060	.0058		
			(.0041)	(.0064)		
Attendance × bad data					–.0128***	–.0121
					(.0038)	(.0076)
Hell × bad data					.0022	.0048
					(.0034)	(.0057)
Number of observations for each time period	33, 35, 34	33, 35, 34	48, 53, 52	48, 53, 52	48, 53, 52	48, 53, 52
R^2 values for each time period	.44, .63, .24	.73, .32, .16	.72, .61, .57	.69, .64, .48	.63, .62, .63	.62, .64, .54

Notes: For general model details, see the note to Table 1. Models 10 and 11 exclude Asian and African countries. In Models 10, 12, and 14 estimation is by seemingly unrelated regressions (SUR).
*** $p < .001$ (two-tailed tests).

the countries with the worst data quality (D grade) and incrementally weaker for broader specifications of poor quality (weaker for D and C combined, and weaker still for D, C, and B combined). In summary, the Barro and McCleary findings are driven largely or entirely by the countries with poor quality data.

CONCLUSIONS

The purpose of this article has been to outline a framework of model uncertainty, and to demonstrate its relevance by replicating Barro and McCleary's prominent research on religion and economic growth. It is clear that these results are highly sensitive to model specification. First, the findings are only supported in period 2 (1975 to 1985) using religiosity data imputed from period 3 (1985 to 1995). Second, the models do not effectively correct for reverse causation. While the authors should be lauded for seriously engaging the issue, the instrument set in this study is weak. In this case, the weak instruments appear to exaggerate the estimated effect of religiosity. Finally, interaction models indicate that support for the efficient religion hypothesis is largely or entirely driven by (1) Asian and African countries and (2) countries where data quality is poor.

It should be emphasized, however, that these findings do not mean that "Weber was wrong" or that religion is unrelated to economic performance. Researchers are sometimes too quick to conclude that "a given variable 'does not matter' when a more accurate interpretation is that its effect cannot be identified using the data at hand" (Durlauf et al. 2006:631). This is particularly true for cross-country growth regressions where the number of observations is very limited and the variables are often "crude proxies for underlying theories" (Brock and Durlauf 2001:252).

One step forward would be to develop better, more theoretically-driven measures of religiosity. Weber, for example, emphasized not simply Protestant affiliation or belief in heaven and hell, but rather a very specific theology of "this world" asceticism. McCleary (2007) emphasizes the degree to which religious people believe that they can, through their own actions, influence their chance of attaining salvation. The data do not capture this kind of specificity. Moreover, it seems that many people who believe in hell also believe that their own chances of going there are quite small; hell is

usually for other people—for example, outsiders who have not accepted Jesus Christ as their personal savior. Ultimately, we want to measure the degree to which people believe that God is watching them and may punish them for their misdeeds.

Another way forward may be more localized and fine-tuned research. Rather than straining both the data and the meaning of religiosity by combining all the statistically observed countries of the world into the same model, one might focus more specifically on salient events like the ongoing Protestant Reformation in Latin America (Martin 1990). Woodberry (2004) has proposed studying small Protestant reformations in specific communities in countries such as Brazil that offer good quality census data to see how neighborhood fortunes change as their local reformations gather steam. Such events allow us, in a sense, to rewind history and observe what happens when poor, Catholic communities experience large-scale conversion to Protestantism.

In methodological terms, this article illustrates how research findings can contain a great deal of model uncertainty that is not revealed in conventional significance tests. A point estimate and its standard error is not a reliable guide to what the next study is likely to find, even if it uses the same data. This is true even if, as in this case, the research is conducted by a highly respected author and is published in a top journal. Below, I outline a number of specific steps that could help improve the transparency and credibility of statistical research in sociology.

1. Pay greater attention to model uncertainty. The more that researchers (and editors and reviewers) are attuned to the issue of model uncertainty, it seems likely that more sensitivity analyses will be reported. Researchers with results they know are strong will look for ways to signal that information (i.e., to report estimates from a wider range of models). Results that depend on an exact specification, and unravel with sensible model changes, are not reliable findings. When this is openly acknowledged, the extensiveness of sensitivity analysis will, more and more, augment significance tests as the measure of a strong finding.

2. Make replication easier. Authors should submit complete replication packages (dataset and statistical code) to journals as a condition of publication, so that skeptical readers can easily interrogate the results themselves (Freese 2007). This is particularly important for methodologically complex papers where it can be quite difficult and time consuming to perform even basic replications from scratch (Glaeser 2006). Asking authors for replication materials often seems confrontational, and authors often do not respond well to their prospective replicators. In psychology, an audit study found that only 27 percent of authors complied with data requests for replication (Wicherts et al. 2006). Barro's openness in welcoming this replication—readily providing the data and even offering encouragement—seems to be a rare quality. Social science should not have to rely on strong personal integrity of this sort to facilitate replication. The institutional structure that publishes research should also ensure that any publication can be subject to critical inspection.[16]

3. Establish routine random testing of published results. Pushing the previous point a bit further, Gerber and Malhotra (2006) suggest establishing a formal venue within journals for randomly selected replications of published articles. The idea is to develop a semiregular section titled "Replications," with its own designated editor, in which several of the statistical papers each year are announced as randomly selected for detailed scrutiny, with wide distribution of the data and code, and the range of findings reported in brief form (as in Table 1). Indeed, this could provide ideal applied exercises for graduate statistics seminars. Even if only a dozen or so departments across the country incorporate it into their classes, this alone would provide a remarkably thorough robustness check. The degree of model variance would quickly become transparent. Moreover, the prospect of such scrutiny would no doubt encourage researchers to preemptively critique

[16] The *American Economic Review* (AER), for example, requires data and statistical code to be submitted prior to publication, and these materials are permanently hosted on the AER Web site. For details on the AER data availability policy, see (http://www.aeaweb.org/aer/data_availability_policy.html). Certainly, exceptions must be made where confidentiality and related legal issues would limit data availability (Freese 2007).

their own findings and report more rigorous sensitivity analyses.

4. Encourage pre-specification of model design. One of the problems in statistics today is that authors have no way to credibly signal when they have conducted a true (classical) hypothesis test. Suppose a researcher diligently plans out her model specifications before she sees the data and then simply reports those findings. This researcher would be strategically better off to conduct specification searches to improve her results because readers cannot tell the difference between a true hypothesis test and a data mining exercise. The situation would be greatly improved if there were some infrastructure to facilitate credible signaling. A research registry could be a partial solution. In medical research, clinical trials must be reported to a registry—giving a detailed account of how they will conduct the study and analyze the data—before beginning the trial.[17] A social science registry would similarly allow authors to specify their models before the data become available (Nuemark 2001). This is feasible for established researchers using materials like time series data or future waves of the major surveys (e.g., NLSY, PSID, and GSS). This will, for the subset of work that is registered, bring us back to a time when model specification had to be carefully planned out in advance. Authors could then report the results of their pre-specified designs (i.e., their true hypothesis tests), *as well as* search for alternative, potentially better, specifications that can be tested again when the next round of data becomes available. Because most data already exist, and authors can only credibly pre-specify for future data, this would be a long-term strategy for raising the transparency of statistical research and reducing the information asymmetry between analyst and reader.

Thirty years ago, model uncertainty existed but computational limitations created a "veil of ignorance"—neither analyst nor reader knew much about how model specification affected the results. Today, authors know (or can learn) much more about the reliability of their estimates—how much results change from model to model—than their readers. As Hoeting and colleagues (1999:399) argue, it seems clear that in the future, "accounting for model uncertainty will become an integral part of statistical modeling." All of the steps outlined here would go far, as Leamer (1983) humorously put it, to "take the con out of econometrics."

[17] The aim of this practice is to prevent pharmaceutical companies from funding 20 trials and then suppressing all but the "positive" results.

Cristobal Young *is a PhD candidate in the Department of Sociology at Princeton University. His research focuses on economic sociology and quantitative methods.*

REFERENCES

Allison, Paul D. 2001. *Missing Data.* Thousand Oaks, CA: Sage.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. 1999. "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias." *Labor Economics* 6:453–70.

Barro, Robert J. 1997. *Determinants of Economic Growth: A Cross-Country Empirical Study.* Cambridge, MA: MIT Press.

Barro, J. Robert and Rachel M. McCleary. 2003. "Religion and Economic Growth across Countries." *American Sociological Review* 68(5):760–81.

Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41(2):641–74.

Berger, Peter L. 1967. *The Sacred Canopy: Elements of a Sociological Theory of Religion.* New York: Doubleday.

Borjas, George. 1999. *Economic Research on the Determinants of Immigration: Lessons for the European Union.* World Bank Technical Paper No. 438.

Bound, John, David A. Jaeger, and Regina Baker. 1993. "The Cure Can Be Worse than the Disease: A Cautionary Tale Regarding Instrumental Variables." NBER Working Paper Series, Vol. T0137.

———. 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(430):443–50.

Brock, William and Steven Durlauf. 2001. "Growth Empirics and Reality." *World Bank Economic Review* 15(2):229–72.

Chatfield, Chris. 1995. "Model Uncertainty, Data Mining and Statistical Inference." *Journal of the Royal Statistical Society* 158(3):419–66.

Doucouliagos, Hristos and T. D. Stanley. 2008. "Publication Selection Bias in Minimum-Wage Research? A Meta-Regression Analysis." Working Paper, Department of Economics, Hendrix College, Conway, AR.

Durlauf, Steven, Paul Johnson, and Jonathan Temple. 2006. "Growth Econometrics." Pp. 555–677 in *The Handbook of Economic Growth,* edited by P. Aghion and S. Durlauf. Amsterdam: North Holland.

Finke, Roger and Rodney Stark. 1992. *The Churching of America 1776–1990: Winners and Losers in Our Religious Economy.* New Brunswick, NJ: Rutgers University Press.

Freese, Jeremy. 2007. "Replication Standards for Quantitative Social Science: Why Not Sociology?" *Sociological Methods and Research* 36(2):153–72.

Fukuyama, Francis. 1995. *Trust: The Social Virtues and the Creation of Prosperity.* New York: Penguin Books.

Gerber, Alan and Neil Malhotra. 2006. "Can Political Science Literatures Be Believed? A Study of Publication Bias in the APSR and the AJPS." Working Paper. Society for Political Methodology.

Glaeser, Edward. 2006. "Researcher Incentives and Empirical Methods." Harvard Institute of Economic Research. Discussion Paper Number 2122.

Granger, Clive and Yongil Jeon. 2004. "Thick Modeling." *Econometric Modeling* 21:323–43.

Grier, David Alan. 2005. *When Computers Were Human.* Princeton, NJ: Princeton University Press.

Hahn, Jinyong and Jerry Hausman. 2003. "Weak Instruments: Diagnosis and Cures in Empirical Econometrics." *American Economic Review* 93(2):118–25.

Hoeting, Jennifer, David Madigan, Adrian Raftery, and Chris Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14(4):382–417.

Hsieh, Ching-Chi and M. D. Pugh. 1993. "Poverty, Income Inequality, and Violent Crime: A Meta-Analysis of Recent Aggregate Data Studies." *Criminal Justice Review* 18(2):182–202.

Hsu, Becky, Amy Reynolds, Conrad Hackett, and James Gibbon. 2008. "Estimating the Religious Composition of All Nations: An Empirical Assessment of the World Christian Database." *Journal for the Scientific Study of Religion* 47(4):678–93.

Hunter, John and Frank Schmidt. 2004. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* 2nd ed. Thousand Oaks, CA: Sage.

Iannaccone, Laurence. 2003. "Looking Backward: A Cross-National Study of Religious Trends." Working Paper, Center for Study of Public Choice, George Mason University, Fairfax, VA.

Kim, E. Han, Adair Morse, and Luigi Zingales. 2006. "What Has Mattered to Economics Since 1970?" NBER Working Paper, 12526.

Kuran, Timur. 2004. "Why the Middle East is Economically Underdeveloped: Historical Mechanisms of Institutional Stagnation." *Journal of Economic Perspectives* 18(3):71–90.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73(1):31–43.

———. 1985. "Sensitivity Analyses Would Help." *American Economic Review* 75(3):308–313.

Lindsay, D. Michael. 2007. *Faith in the Halls of Power: How Evangelicals Joined the American Elite.* Oxford, UK: Oxford University Press.

Magnus, Jan and Mary Morgan. 1999. *Methodology and Tacit Knowledge: Two Experiments in Econometrics.* New York: John Wiley & Sons.

Martin, David. 1990. *Tongues of Fire: The Explosion of Protestantism in Latin America.* Oxford, UK: Basil Blackwell.

McCleary, Rachel M. 2007. "Salvation, Damnation, and Economic Incentives." *Journal of Contemporary Religion* 22(1):49–74.

McCleary, Rachel M. and Robert J. Barro. 2006. "Religion and Economy." *Journal of Economic Perspectives* 20(2):49–72.

Nelson, Charles and Richard Startz. 1990. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator." *Econometrica* 58(4):967–76.

Nuemark, David. 2001. "The Employment Effects of Minimum Wage: Evidence from a Pre-specified Research Design." *Industrial Relations* 40(1):121–44.

Pinello, Daniel. 1999. "Linking Party to Judicial Ideology in American Courts: A Meta-Analysis." *The Justice System Journal* 20(3):219–54.

Rawls, John. 1971. *A Theory of Justice.* Cambridge, MA: Belknap Press.

Robinson, John and Geoffrey Godbey. 2000. *Time for Life: The Surprising Ways Americans Use Their Time.* 2nd ed. University Park, PA: Pennsylvania State University.

Sala-i-Martin, Xavier. 1997. "I Just Ran Two Million Regressions." *American Economic Review* 87(2):178–83.

Sala-i-Martin, Xavier, Gernot Doppelhofer, and Ronald Miller. 2004. "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates Approach." *American Economic Review* 94(4):813–35.

Staiger, D. and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65:557–86.

Stith, Sandra, Karen Rosen, Kimberly Middleton, Amy Busch, Kirsten Lundberg, and Russel Carlton. 2000. "The Intergenerational Transmission of Spouse Abuse: A Meta-Analysis." *Journal of Marriage and Family* 62(3):640–54.

Stock, James H., Jonathan H. Wright, and Motohiro Yogo. 2002. "A Survey of Weak Instruments and Weak Identification in Generalized Method of

Moments." *Journal of Business and Economic Statistics* 20(4):518–29.

Stock, James H. and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." Pp. 80–108 in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg,* edited by D. W. K. Andrews and J. H. Stock. Cambridge, UK: Cambridge University Press.

Weber, Max. 1930. *The Protestant Ethic and the Spirit of Capitalism.* Trans. by T. Parsons. New York: Routledge.

Western, Bruce. 1996. "Vague Theory and Model Uncertainty in Macrosociology." *Sociological Methodology* 26:165–92.

Wicherts, Jelte, Denny Borsboom, Judith Kas, and Dylan Molenaar. 2006. "The Poor Availability of Psychological Research Data for Reanalysis." *American Psychologist* 61:726–28.

Woodberry, Robert D. 2004. "Project on Religion and Economic Change: Grant Proposal for the Spiritual Capital Project." Center for the Scientific Study of Religion, Population Research Center, University of Texas at Austin.