Ideological bias in democracy measures

Thiago Marzagão*
*PhD candidate, Dept. of Political Science, Ohio State University, marzagao.1@osu.edu

Abstract

In this paper I show that, unlike what previous research has led us to believe, we cannot make any claims about the nature of the ideological biases that contaminate existing democracy measures. For instance, I show that the Freedom House data, often believed to have a conservative bias, may actually have a liberal bias instead. I do that by replicating previous research on the subject (Bollen and Paxton 2000) but replacing real-world data by simulated data in which I manipulate democracy levels and the ideological biases of hypothetical raters. The results of these Monte Carlos show that even though we can confidently assert the existence of bias in *some* democracy measures we cannot say anything about *which* measures are biased or in *what* ways.

1. Introduction

What do we know about ideological bias in democracy measures?[1] Bollen and Paxton (2000), using structural equation modeling, find that several indicators from the Freedom House dataset (Sussman 1982; Gastil 1988) and from Arthur Banks' Cross-National Time Series Archive - CNTS (Banks [1971], updated through 1988) are compromised by ideological bias: the coding is sensitive to a number of variables that (conceptually) have nothing to do with democracy, such as economic policy (whether the polity is Marxist-Leninist), religion (whether the polity is predominantly Roman Catholic or predominantly Protestant), and form of government (whether the polity is a monarchy or a republic). Bollen and Paxton's article has been highly influential in political science and it appears in numerous discussions of democracy measurement (e.g.: Munck and Verkuilen [2002], Treier and Jackman [2008], and Pemstein, Meserve and Melton [2010]).

[1]This work was funded by the Fulbright (grantee ID 15101786); by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES (BEX 2821/09-5); and by the Ministério do Planejamento, Orçamento e Gestão - MPOG (proceed n. 03080.000769/2010-53).

Although highly influential, Bollen and Paxton's findings have never been subject to scrutiny; they are usually taken at face value. In this paper I reassess those findings. I do that with simulated data in which I manipulate the countries' level of democracy and the raters' ideological biases. I find that Bollen and Paxton's method: a) yields incorrect results about which raters are biased; b) yields incorrect results about the nature of those biases; and c) fails to find bias when different raters are biased in similar ways.

Section 2 details Bollen and Paxton's work. Section 3 replicates Bollen and Paxton's work. Section 4 explains the Monte Carlos and presents the results. Section 5 concludes.

2. Bollen and Paxton's analysis

In this section I explain Bollen and Paxton's methodology and results.

2.1 Extracting latent variables

Bollen and Paxton's analysis is based on eight indicators, four from the Freedom House dataset and four from the CNTS dataset. The Freedom House indicators are: "freedom of broadcast media", "freedom of print media", "civil liberties", and "political rights". The CNTS indicators are: "freedom of group opposition", "competitiveness of the nomination process", "chief executive elected", and "effectiveness of the legislative body".

Bollen and Paxton start by using structural equation modeling (SEM) to extract five latent variables from those eight indicators. Two latent variables are assumed to be democracy features ("political liberties" and "democratic rule") and three are assumed to be coder-specific ideological biases (Raymond Gastil's and Leonard Sussman's, who were Freedom House coders, and Arthur Banks', who was the CNTS coder)[2].

[2]Raymond Gastil was responsible for the "civil liberties" and "political rights" indicators. Leonard Suss-

Each of the eight indicators is modeled as being determined by a traits factor, a methods factor, and a random measurement error. For instance, "freedom of broadcast media" is modeled as being determined by the traits factor "political liberties", by the methods factor "Leonard Sussman's bias" (since Leonard Sussman was the researcher responsible for the "freedom of broadcast media" indicator), and by random measurement error. More generally, each indicator is modeled as

$$indicator_{kp} = \lambda_t trait_{tp} + \lambda_m method_{mp} + \delta_{kp} \quad (1)$$

where indicator k for polity p is a linear combination of traits factor t for polity p, methods factor m for polity p, and indicator k's random measurement error for polity p. The complete picture of which indicators load on which factors is provided in Figure 1 below, extracted from Bollen and Paxton (65).[3]

[Figure 1 here]

Figure 1 follows the standard SEM notation, with squared boxes representing indicators (i.e., observed variables) and circles representing factors (i.e., latent variables). The factor-to-indicator arrows show which indicators load on which factors.[4] The factor-to-factor arrows show which factors correlate.[5] The "E" arrows show which indicators have random measurement error.[6]

man was responsible for the "freedom of broadcast media" and "freedom of print media" indicators. Arthur Banks was responsible for the 'freedom of group opposition", "competitiveness of the nomination process", "chief executive elected", and "effectiveness of the legislative body" indicators.

[3]I thank Prof. Bollen for helping me understand some aspects of the model specification.

[4]Based on previous work, Bollen and Paxton model the "freedom of group opposition" as being free from Banks' ideological bias.

[5]The "political liberties" and "democratic rule" factors correlate because they are close concepts. "Sussman" and "Gastil" correlate because they both worked at the Freedom House.

[6]Based on previous work, Bollen and Paxton model the "political rights" and "competitiveness of the nomination process" indicators as having no random measurement error, i.e., $\delta = 0$.

Using country-level data from 1972 to 1988 Bollen and Paxton estimate, for each indicator, the factor loadings (i.e., the lambdas) and the random measurement error. That is done through the usual SEM procedure of finding the estimates that minimize the difference between the observed variance-covariance matrix (S, in SEM notation) and the implied variance-covariance matrix (Σ, in SEM notation). The S is simply the observed variance-covariance matrix of the indicators. The Σ is the theoretical variance-covariance matrix derived from the model, i.e.

$$\Sigma = E[xx'] = E[(\Lambda_x\xi + \delta)(\xi^{'}\Lambda_x^{'} + \delta^{'})] = \Lambda_x E(\xi\xi^{'})\Lambda_x^{'} + \Theta_\delta = \Lambda_x\Phi\Lambda_x^{'} + \Theta_\delta \qquad (2)$$

where x is the matrix of indicators, Λ is the matrix of factor loadings, ξ is the matrix of factors, Φ is the variance-covariance matrix of factors, and Θ is the variance-covariance matrix of random measurement errors (Bollen 1989, 236-237).

The estimates that minimize the difference between S and Σ are found by minimizing the following function via maximum likelihood:

$$F = log|\Sigma| + trace[S\Sigma^{-1}] - log|S| - k \qquad (3)$$

where $log|\Sigma|$ is the natural logarithm of the determinant of Σ, $trace[S\Sigma^{-1}]$ is the trace of the product of S and the inverse of Σ, $log|S|$ is the natural logarithm of the determinant of S, and k is the number of indicators (Bollen 1989, 254). Bollen and Paxton repeat that for every year between 1972 and 1988 and find the fit statistics shown in Figure 2 below, extracted from their article (67).

[Figure 2 here]

Figure 2 compares two fit statistics - the Incremental Fit Index (IFI) and the Root Mean Square Error of the Approximation (1-RMSEA) - for each year from 1972 to 1988. In both cases (IFI and 1-RMSEA) the larger the statistic, the better the model fit. As we see, the fit of the model improves considerably when we include both traits and methods factors, as compared to when we include only traits factors.[7] The improved fit shows that each of the eight indicators is the product not only of the underlying trait ("political liberties" or "democratic rule", according to the case) and random measurement error, but also the product of a systematic component - the rater's ideological bias.

Next, Bollen and Paxton use those estimates to produce three sets of factor scores. These factor scores are produced using the formula $\hat{L} = \hat{\Sigma}_{LL}\hat{\Lambda}\hat{\Sigma}_{xx}x$, where $\hat{L}$ is the matrix of estimated factor scores, $\hat{\Sigma}_{LL}$ is the estimated variance-covariance matrix of factors, $\hat{\Lambda}$ is the matrix of estimated factor loadings, $\hat{\Sigma}_{xx}$ is the estimated variance-covariance matrix of indicators, and x is the mean-centered matrix of observed indicators. The factor scores are based on the average estimated SEM parameters of 1979 and 1980.[8]

2.2 Regressing factor scores on country-level variables

Bollen and Paxton regress the factor scores produced in the previous step on a number of country-level variables, for two sets of years: 1972, 1975, 1980, 1984, and 1988 (Gastil's factor scores and Banks' factor scores); and 1980 and 1984 (Sussman's factor scores). Table 1 below reproduces the estimates they obtained using 1980 data.

[Table 1 here]

[7] Not all factors are included in every year: Gastil's and Bank's indicators are available for the entire 1972-1988 interval, but Sussman's are only available for the 1979-1981 and 1983-1987 intervals. Hence for 1972-1978, 1980, and 1988 the estimated model is actually a restricted version of the model depicted in Figure 1 above: everything is the same except that the Sussman factor and the corresponding indicators are not included.

[8] Bollen and Paxton also check whether the parameters are stable over time.

As we observe, Leornad Sussman and Raymond Gastil seem to be biased against Marxist-Leninist countries and in favor of Roman Catholic countries, whereas Arthur Banks seems to be biased in favor of Marxist-Leninist countries, Protestant countries, and Roman Catholic countries, and against monarchic countries. In regressions using data from other years, Bollen and Paxton also find positive, statistically significant coefficients for the Protestant variable in the Gastil and Sussman regressions (Bollen and Paxton 2000, 76). As the next section will show, none of these conclusions is warranted.

3. Replication

Before running the Monte Carlos I first replicated Bollen and Paxton's analysis with real-world data, just to make sure I was following the same procedures. I collected the four Freedom House measures and the four CNTS measures and modeled the latent factors exactly as in Bollen and Paxton. I estimated the model for the same years Bollen and Paxton did (1972-1988) and obtained the same fit statistics they did.[9]

[Figure 3 here]

The second part of the replication - i.e., regressing factor scores on country-level variables - was less successful, with several discrepancies in terms of coefficient signs and statistical significance. It is hard to know why. Prof. Bollen and Prof. Paxton no longer have the data, so I had to tabulate everything from scratch, using the same sources. Many

[9] As is often the case with SEM estimation, making the models converge took some doing. I follow Kolenikov's (2009) three-step approach: first I estimate the "traits-only" model, save the estimates, and produce residuals of the eight indicators; second I use those residuals to estimate the "methods-only" model and save the estimates; third I combine both sets of estimates and use them as starting values for the full model (i.e., the model as depicted in Figure 1, with both traits and methods factors). In other words, I use estimates of restricted models as informative starting values when estimating the full model. However, even with informative starting values the models often do not converge (after hundreds of iterations). Thus some of the fit statistics in Figure 3 are based on models that did not converge. But all fit statistics are virtually identical to the fit statistics reported by Bollen and Paxton themselves.

of these were printed sources, so typing errors are a possibility. Some of the data were also available from secondary, electronic sources. I tried these as well, but still could not replicate the original estimates exactly.

I also tried coding the media coverage variable in different ways. Bollen and Paxton report that they built that variable based on how often the country appeared on the New York Times, on the CBS News Index, and on the Facts on File almanac, but they do not report how exactly: was it a mere sum of each country's mentions in each of those three sources? Was it a weighted sum? Was it logged? Were those three sources combined into a single principal components value? I tried each of these possibilities, but still could not obtain the exact same results Bollen and Paxton did.

Finally, I tried dropping different combinations of countries. Bollen and Paxton's 1980 dataset has 81 countries whereas my own 1980 dataset has 112 countries. Bollen and Paxton do not specify what those 81 countries are, so I tried dropping several combinations of 31 countries (112 - 81 = 31) to see if I could obtain the same results, but that did not work either.

Table 2 below reports the regression estimates I obtained for the year 1980, using electronic sources and all the 112 countries, and coding media coverage as the natural logarithm of how often the country was mentioned on the New York Times, CBS News Index, or Facts on File in that year.

[Table 2 here]

As we see, some results are the same but others are not. As in Bollen and Paxton here too Marxism-Leninism has a negative, statistically significant coefficient in the Gastil and Sussman regressions and a positive, statistically significant coefficient in the Banks regression. Also as in Bollen and Paxton we find here that Roman Catholic has a positive,

statistically significant coefficient in the Gastil and Sussman regressions. The similarities stop there. In Bollen and Paxton Protestant, Roman Catholic, and monarchy all turn out statistically significant in the Gastil regression, but in the replication they do not. (The other variables have little to do with ideological bias so they are of no interest here.)

4. Monte Carlos

4.1 Basic idea

How solid are the results obtained in the previous section? The estimates reported on Table 2 suggest, for instance, that Gastil and Sussman are biased against Marxism-Leninism and that Banks is biased in favor of Marxist-Leninist countries. But what if all three raters are biased in the same direction, only to different degrees? If all three raters are biased in favor of Marxism-Leninism but Banks much more so than Gastil and Sussman, couldn't that produce the opposite coefficient signs we observe? Or, alternatively, if all three are biased against Marxist-Leninist countries but Gastil and Sussman much more so than Banks, couldn't that produce opposite coefficient signs as well?

The same applies to the other three variables of interest - Protestant, Roman Catholic, and monarchy. Table 2 suggests that none of the raters are biased against or in favor of Protestant or monarchic countries. But maybe they all are, and to similar degrees - so the bias becomes "invisible" and the SEM estimates simply cannot capture it. Table 2 also suggests that Gastil and Sussman are biased in favor of Roman Catholic countries whereas Banks is not. But what if Banks is biased in favor of Roman Catholic countries as well, just less so than Gastil and Sussman?

How can we verify all that? We cannot observe a country's "true" level of democracy or a rater's ideological bias - these are latent variables. But we can simulate them, fixing

the parameters of interest at certain levels, and then re-do the whole analysis with the simulated data and check how often we obtain misleading results.

I start by producing simulated data in which I fix the level of democracy and the direction and magnitude of each rater's ideological bias. I then make these simulated factors load on a number of simulated indicators, estimate the structural model, and extract the factor scores.

Finally, I regress those simulated factor scores on (real-world) country-level variables, repeat the process many times, and count how often we obtain misleading coefficients - for instance, how often the coefficient of Marxism-Leninism turns out negative and significant even when the simulated rater actually *favors* Marxist-Leninist countries. That should give us an idea of how reliable the findings on Table 2 - and, by extension, those in Bollen and Paxton - are.

4.2 Model specification

I begin by simulating three factors: each country's level of democracy; the idiosyncrasies (i.e., the systematic measurement error) of a hypothetical rater we are going to call Rater #1; and the idiosyncrasies of a hypothetical rater we are going to call Rater #2. The level of democracy is generated as a uniform random variable ranging from 0 to 20.[10] Rater #1's factor is generated as a normal random variable with mean 5 and standard deviation 5. And Rater #2's factor is generated as a normal random variable with mean 15 and standard deviation 15.[11] For each factor I generate 112 observations (the number of countries in the dataset).

[10] That seems to be the distribution of actual measures of democracy (e.g., the "political rights" index of the Freedom House).

[11] The normal distribution is chosen because structural equation models rely on the assumption that the factors follow a multivariate normal distribution (thus we could not have all three factors follow a uniform distribution).

The second step is to introduce ideological bias into the factors. I do that by making the factors alternately respond to Marxism-Leninism, Protestantism, Roman Catholicism, or monarchy. The nature of the simulated bias is different across these four variables. In the case of Marxism-Leninism Table 2 suggests opposite biases. So I test whether the same result might be obtained even if Rater #1 and Rater #2 were biased in the same direction, but to different degrees. Hence for Marxist-Leninist countries I boost Rater #1's factor by *p1* points and Rater #2's factor by *p2* points, with *p2* always fixed at 0.025 and *p1* taking the following values: 0.5, 1, 3, 5, 7, 10, 15, and 20.

In the case of Protestantism Table 2 would have us believe that none of the raters are biased. But what if all raters are biased in the same direction and to similar degrees? Could that not make the bias become "invisible" in the estimation? To check that, for Protestant countries I boost Rater #1's factor by *c1* percent and Rater #2's factor by *c2* percent, with the *c1*-*c2* pairs being: 30%-35%, 50%-55%, 70%-75%, 130%-135%, 150%-155%, 170%-175%, 190%-195%, and 230%-235%.[12]

In the case of Roman Catholicism Table 2 suggests that Gastil and Sussman are positively biased and that Banks is not biased in any direction. We want to know whether that result might be obtained even if all three raters were positively biased, only with Gastil and Sussman more so than Banks. So here I do the same as in the Marxism-Leninism case: for Roman Catholic countries I boost Rater #2's factor by *p2*=0.025 points and Rater #1's factor by *p1* points, with *p1* taking the following values: 0.5, 1, 3, 5, 7, 10, 15, and 20.

Finally, in the case of monarchy Table 2 suggests no one is biased, but - as in the case of Protestantism - perhaps Gastil, Sussman, and Banks are all biased in the same direction and to similar degrees, which could make the bias "disappear" in the SEM estimations. So

[12]Thus here the bias is multiplicative - unlike in the Marxism-Leninism case, where the bias is additive.

for monarchies, as for Protestant countries, I boost Rater #1's factor by *c1* percent and Rater #2's factor by *c2* percent, with the *c1*-*c2* pairs being, again, 30%-35%, 50%-55%, 70%-75%, 130%-135%, 150%-155%, 170%-175%, 190%-195%, and 230%-235%.

The third step is to use those simulated factors to generate simulated indicators (i.e., the variables we do observe in SEM estimation). I model them as follows:

$$indicator1 = 14.12 \times rater1 + 69.87 \times democracy + \delta_1/m \quad (4)$$
$$indicator2 = 06.71 \times rater1 + 68.57 \times democracy + \delta_2/m \quad (5)$$
$$indicator3 = 18.31 \times rater1 + 53.45 \times democracy + \delta_3/m \quad (6)$$
$$indicator4 = 31.81 \times rater1 + 28.21 \times democracy + \delta_4/m \quad (7)$$
$$indicator5 = 95.69 \times rater1 + 40.63 \times democracy + \delta_5/m \quad (8)$$
$$indicator6 = 38.13 \times rater1 + 97.69 \times democracy + \delta_6/m \quad (9)$$
$$indicator7 = 70.70 \times rater2 + 21.17 \times democracy + \delta_7/m \quad (10)$$
$$indicator8 = 31.51 \times rater2 + 51.63 \times democracy + \delta_8/m \quad (11)$$
$$indicator9 = 90.83 \times rater2 + 26.09 \times democracy + \delta_9/m \quad (12)$$
$$indicator10 = 12.99 \times rater2 + 55.01 \times democracy + \delta_{10}/m \quad (13)$$
$$indicator11 = 53.06 \times rater2 + 63.13 \times democracy + \delta_{11}/m \quad (14)$$
$$indicator12 = 52.19 \times rater2 + 15.67 \times democracy + \delta_{12}/m \quad (15)$$

As we see, democracy loads on all twelve indicators; Rater #1 loads on the first six indicators; and Rater #2 loads on the last six indicators. There is also a random measurement error, δ, specific to each indicator. In one third of the simulations the m parameter (that divides δ) is simply 1, so the error term does not suffer any transformation. In another third of the simulations the m parameter is 0.001, so we can see what happens to the estimates when the random errors are magnified. And in another third of the simulations the m parameter is 1,000, so we can see what happens to the estimates when the random errors shrink. Each δ is a combination of a normal random variable and a

beta random variable, as follows:

$$\delta_1 = N(\mu = 0, \sigma = 827) + Beta(\alpha = 0.68, \beta = 0.78) \quad (16)$$

$$\delta_2 = N(\mu = 0, \sigma = 4188) + Beta(\alpha = 0.10, \beta = 0.72) \quad (17)$$

$$\delta_3 = N(\mu = 0, \sigma = 228) + Beta(\alpha = 0.58, \beta = 0.67) \quad (18)$$

$$\delta_4 = N(\mu = 0, \sigma = 3237) + Beta(\alpha = 0.50, \beta = 0.60) \quad (19)$$

$$\delta_5 = N(\mu = 0, \sigma = 1965) + Beta(\alpha = 0.06, \beta = 0.83) \quad (20)$$

$$\delta_6 = N(\mu = 0, \sigma = 734) + Beta(\alpha = 0.15, \beta = 0.86) \quad (21)$$

$$\delta_7 = N(\mu = 0, \sigma = 1439) + Beta(\alpha = 0.51, \beta = 0.46) \quad (22)$$

$$\delta_8 = N(\mu = 0, \sigma = 2983) + Beta(\alpha = 0.23, \beta = 0.40) \quad (23)$$

$$\delta_9 = N(\mu = 0, \sigma = 1190) + Beta(\alpha = 0.73, \beta = 0.21) \quad (24)$$

$$\delta_{10} = N(\mu = 0, \sigma = 112) + Beta(\alpha = 0.26, \beta = 0.63) \quad (25)$$

$$\delta_{11} = N(\mu = 0, \sigma = 806) + Beta(\alpha = 0.54, \beta = 0.37) \quad (26)$$

$$\delta_{12} = N(\mu = 0, \sigma = 4299) + Beta(\alpha = 0.13, \beta = 0.46) \quad (27)$$

These modeling choices need justification. The number of factors - three - is the minimum we need to be able to fix each country's level of democracy and to evaluate Bollen and Paxton's assertions about bias direction. The number of indicators (twelve) is somewhat arbitrary; it could have been eight or fourteen, for instance. What matters is that for each factor there are at least three or four indicators, so that there are enough data to estimate the model. The loading coefficients (14.12, 69.87, etc) are entirely arbitrary, except that they are always positive; otherwise the direction of the bias would change between the factors and the indicators[13].[14]

The parameters shown above - the distributional parameters of the factors, the factor

[13]E.g., if Rater #1 is biased *in favor* of Marxism-Leninism, a negative factor loading would make the corresponding indicator be biased *against* Marxism-Leninism.

[14]Initially I generated the random errors (the δs) as purely normal variables. But that resulted in excessive correlations between the errors, even when varying the standard deviations. That resulted in highly correlated indicators, which results in non-invertible matrices and makes SEM estimation impossible. That is why I add the beta component.

loadings of each indicator, and the distributional parameters of the error terms - remain the same across all simulations. But at each simulation the factors and the random errors are redrawn, so the indicators (which are functions of both) change as well. Also, the m parameter, as explained above, assumes three different values (1, 0.001, and 1,000).

The simulations are done separately for each of the four variables of interest, i.e., in any given simulation the hypothetical raters are biased toward only one of the four variables. The basic procedure is: I generate the three factors (democracy, Rater #1, Rater #2), bias Rater #1 and Rater #2, generate the twelve random errors, generate the twelve indicators, estimate the structural equations model, save the two sets of factor scores assumed to represent ideological bias (Rater #1's and Rater #2's), regress each set on country-level variables (same ones used to produce Table 2)[15], and count how often the outcome of interest (i.e., the outcome analogous to that of Table 2) obtains.

I repeat this process 1,000 times for each of the four variables of interest and for each of the *p1-p2* pairs and *c1-c2* pairs discussed above. I also repeat the process 1,000 times using the original, unbiased simulated factors, just to have a baseline. Finally, I repeat the whole process for each of the three values of m discussed before (1, 0.001, and 1,000). Thus in total there are 108 different specifications with 1,000 repetitions each.[16]

In SEM estimation identification is usually achieved by fixing some of the parameters. I do that by fixing the variances of the errors and the variances and covariances of the factors. Thus what changes from one simulation to the next are the estimated factor loadings, and consequently the factor scores and the coefficients obtained from regressing these factor scores on country-level variables. All 108,000 estimations converge, so I do not discard any of them (Paxton et al. 2001, 301-302).[17]

[15]These country-level variables are real-world data, not simulated data.

[16](4 variables of interest) × (1 baseline + 8 values of p or c) × (3 values of m) × (1,000 repetitions) = 108,000 simulations

[17]The estimations take about five hours to run on a dual core CPU with 2.4GHz and 4GB of memory

4.3 Results

The results are summarized on Tables 3 and 4 below.

[Table 3 here]

[Table 4 here]

The results corroborate the suspicions raised before. For Marxism-Leninism and Roman Catholicism, if the two raters are biased in the same direction but to different degrees that is enough to produce opposite coefficient signs with statistical significance. Granted, for Marxism-Leninism the difference in the magnitude of the bias must be somewhat extreme: a "bonus" of at least 15 points from Rater #1 and of only 0.025 points from Rater #2.[18] But we do not know how much the bias varies across actual, real-world raters - perhaps that extreme difference is realistic.

For Roman Catholicism it takes a much less extreme difference to produce misleading results: if Rater #1 rewards Roman Catholicism with 3 points and Rater #2 rewards Roman Catholicism with 0.025 points, that is enough to produce misleading results 12.2% of the time (122 simulations out of 1,000). As the difference becomes larger, so does the frequency of misleading results: 20.7% if Rater #1's bonus is of 5 points, and 81.6% if Rater #1's bonus is 20 points.

For Protestantism we find that when the two hypothetical raters are biased in the same direction and to similar degrees, we are bound to find no bias whatsoever in our estimations. When the bias is in the vicinity of 30% we find non-significant coefficients

and using the 'sem' package in R.

[18] Moreover, we must keep in mind that Rater #1's and Rater #2's factors have different means and variances (otherwise the indicators would be too strongly correlated, which would lead to non-invertible matrices). Since Rater #2's factor has mean and variance larger than Rater #1's, the difference between a *p1* of 15 and a *p2* of 0.025 is even more extreme than it seems at first.

73.7% of the time - which would mislead the researcher into thinking that neither rater is biased. Even when the bias is so extreme as to be around 170% we still obtain non-significant coefficients 16.4% of the time.

For monarchy the findings are a bit less extreme, but even then the bias only becomes "visible" when it reaches 170% or more. In other words, the bias must be of 170% or higher so we can obtain misleading results less than 10% of the time.

It is clear, on the other hand, that we do obtain the correct result when *none* of the hypothetical raters are biased. Here the misleading result would be evidence of bias when there is in fact no bias. For Marxism-Leninism that happens less than 1% of the time. For Roman Catholicism that happens less than 5% of the time. For Protestantism that happens less than 10% of the time. And for monarchy that happens less than 3% of the time. That provides little solace though - with real-world data we cannot know whether the lack of statistical significance means unbiasedness or whether it means that all raters are biased in similar ways.

5. Conclusion

What do all these results tell us about Bollen and Paxton's analysis? They tell us two things.

First, when Bollen and Paxton find bias, all we can assert is that at least one of the raters is biased, but we cannot know which one(s) or in which direction(s). Consider Protestantism, for instance. Bollen and Paxton claim that Banks is biased in favor of Protestant countries while Gastil and Sussman are unbiased (with 1980 data). But it may be the case that Gastil and Sussman are biased against Protestant countries while Banks is unbiased. Or perhaps all three are biased in favor of Protestant countries, only Banks

more so than Gastil and Sussman. Or, still, perhaps all three are biased against Protestant countries, only Banks less so than Gastil and Sussman. Our simulations show that in any of these scenarios we might obtain the same results that Bollen and Paxton did.

Second, when Bollen and Paxton do not find bias, there is a good chance that there is bias. We know that because when our simulated raters are biased in the same direction and to similar degrees, our results often suggest no bias; depending on the magnitude of the biases, we get wrong results over 80% of the time. In other words, when all raters are biased in a similar way the bias becomes "invisible".

These same warnings apply to other studies that claim to have uncovered bias in existing measures of democracy. Consider Steiner (2012), for instance. He regresses Freedom House data on other democracy measures and then checks for correlations between the residuals and a number of foreign policy indicators (voting behavior in the UN, alliances, rivalries, foreign assistance, and trade). He finds the expected correlations and concludes that the Freedom House "rates countries that have closer political ties and affinities with the U.S. [...] as more democratic" (4).

But what if the Freedom House is unbiased but all other measures are biased against US-friendly countries? Or what if all measures of democracy are biased against US-friendly countries, only the Freedom House less so than the others? In all these scenarios Steiner might observe exactly the same result, so his conclusions are completely unwarranted.[19] Unless we have an unbiased measure of democracy, any statistical attempt to uncover the direction of ideological biases is futile.

Naturally, there are other democracy data - most prominently the Polity dataset (Mar-

[19]It is perfectly plausible that the Freedom House has an anti-market (and as a consequence perhaps an anti-US bias), since it includes "socioeconomic rights" and "freedom from gross socioeconomic inequalities" among its subcomponents (Munck and Verkuilen 2002, 9). Alternatively, it is perfectly plausible that Steiner results simply reveal that countries with closer ties to the US are more democratic, for whatever reasons.

shall, Jaggers, and Gurr 2013). But all democracy data we have today rely on country experts and for any given country there are only so many experts. As Munck and Verkuilen (2002) warn, "for all the differences that go into the construction of these indices, they have relied, in some cases quite heavily, on the same sources and even the same precoded data." (29). Hence for all we know the Polity data are no less biased than the Freedom House data or the CNTS data. True, these democracy measures correlate highly, but that may merely indicate that they are all biased in similar ways (Munck and Verkuilen 2002, 29).

In sum, all we can assert today is that at least some of our democracy measures are contaminated by ideological biases. We cannot say which ones and we cannot say anything about the direction or magnitude of those biases (these are "known unknowns", to use Donald Rumsfeld's famous expression); and it is possible that other biases exist that we have not uncovered yet ("unknown unknowns").

Replication material

I used Stata to replicate Bollen and Paxton's results. The code is available on GitHub, in two parts:

`https://gist.github.com/thiagomarzagao/e49541433d474d11d1fb`

`https://gist.github.com/thiagomarzagao/18fed6a8afbb484e0c9c`

The data used for the replication is available on Amazon S3:

`http://s3.amazonaws.com/thiagomarzagao/data-bollenpaxton.dta`

That is not exactly the same data Bollen and Paxton used, but it is the closest reconstruction I could make. Unfortunately Bollen and Paxton no longer have the original dataset (or code).

I used R to perform the simulations. The code is available on GitHub:

`https://gist.github.com/thiagomarzagao/c916e2a3ce77ea23d9a8`

The simulated data is generated on-the-fly, so there is no dataset here (just run the code).

References

Banks, Arthur. 1971. *Cross-polity time-series data (electronic dataset updated to 1988).* Cambridge, MA: MIT Press.

Bollen, Kenneth. 1989. *Structural equations with latent variables.* New York, NY: Wiley.

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

Gastil, Raymond D. 1988. *Freedom in the World: Political Rights and Civil Liberties 1987-1988.* Washington, DC: Freedom House.

Kolenikov, Stas. 2009. "Confirmatory Factor Analysis Using 'confa'." *Stata Journal*, 9 (3): 329-373.

Marshall, Monty, Ted Gurr, and Jaggers, Keith. 2013. *Polity IV Project: Political Regime Characteristics and Transitions, 1800-2012, Dataset Users' Manual.* Viena, VA: Center for Systemic Peace.

Munck, Gerardo L., and Jay Verkuilen. 2002. "Conceptualizing and Measuring Democracy Evaluating Alternative Indices." *Comparative Political Studies*, 35 (1): 5-34.

Paxton, Pamela et al. 2001. "Monte Carlo Experiments: Design and Implementation." *Structural Equation Modeling*, 8 (2): 287-312.

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

Steiner, Nils. 2012. "Testing for a Political Bias in Freedom House Democracy Scores: Are US Friendly States Judged to Be More Democratic?" Available at SSRN 1919870.

Sussman, Leonard R. 1982. "The Continuing Struggle for Freedom of Information." In *Freedom in the World*, edited by Raymond D. Gastil, 101-119. Santa Barbara, CA: Greenwood.

Treier, Shawn, and Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science*, 52 (1): 201-217.

Tables

Table 1. Bollen and Paxton's regressions for 1980[a]

	Gastil	Sussman	Banks
Marxist-Leninist	-0.976***	-0.504*	1.366**
	(0.294)	(0.3)	(0.53)
Protestant	0.357	0.459	0.670**
	(0.324)	(0.295)	(0.264)
Roman Catholic	0.835***	1.239****	0.676**
	(0.312)	(0.331)	(0.286)
monarchy	0.343	0.131	-1.441****
	(0.263)	(0.282)	(0.268)
ln(energy per capita)	-0.122	-0.063	0.124
	(0.126)	(0.123)	(0.177)
ln(years since independence)	0.152	0.084	-0.515
	(0.153)	(0.153)	(0.16)
coups	0.349	0.380	-0.615***
	(0.268)	(0.287)	(0.202)
internal or interstate war in 1980?	-0.360	0.022	-0.234
	(0.302)	(0.328)	(0.438)
ln(protests)	0.098	0.068	-0.174
	(0.121)	(0.139)	(0.109)
ln(political strikes)	-0.074	-0.080	0.211
	(0.137)	(0.158)	(0.135)
ln(riots)	-0.200	-0.023	0.075
	(0.142)	(0.148)	(0.148)
media coverage	0.019	0.002	-0.034
	(0.036)	(0.037)	(0.032)
ln(population)	0.155	0.029	0.279**
	(0.117)	(0.139)	(0.122)
ln(area in km^2)	-0.050	-0.025	-0.024
	(0.058)	(0.061)	(0.076)
ln(radio sets + TV sets per capita)	0.067	0.022	-0.105
	(0.075)	(0.067)	(0.102)
intercept	-0.817	-0.494	1.922***
	(0.646)	(0.614)	(0.632)
adjusted R^2	0.24	0.26	0.29
N	81	81	81

[a] OLS estimates. Heteroskedastic-consistent standard errors in parentheses. * p <0.10; ** p <0.05; *** p <0.01; **** p <0.001. Data sources: The New York Times, CBS News Index, Facts on File, The World Almanac and Encyclopedia, United Nations Statistical Yearbook, others.

Table 2. Replication of Bollen and Paxton's regressions for 1980[a]

	Gastil	Sussman	Banks
Marxist-Leninist	-0.874***	-0.846**	0.587**
	(0.257)	(0.358)	(0.293)
Protestant	-0.103	-0.081	0.147
	(0.280)	(0.390)	(0.320)
Roman Catholic	0.494**	0.872**	0.372
	(0.238)	(0.332)	(0.272)
monarchy	0.024	-0.473	-0.166
	(0.276)	(0.385)	(0.315)
ln(energy per capita)	-0.170	-0.088	0.368**
	(0.136)	(0.190)	(0.156)
ln(years since independence)	0.294*	0.279	-0.314**
	(0.129)	(0.179)	(0.147)
coups 1976-1980	-0.018	0.054	-0.620**
	(0.159)	(0.222)	(0.182)
internal or interstate war in 1980?	-0.677**	-0.617*	0.352
	(0.263)	(0.366)	(0.300)
ln(protests in 1975-1980)	0.066	0.118*	0.004
	(0.043)	(0.060)	(0.049)
ln(strikes in 1975-1980)	-0.027	0.011	0.031
	(0.038)	(0.053)	(0.043)
ln(riots in 1975-1980)	-0.029	-0.020	0.017
	(0.041)	(0.058)	(0.047)
ln(media coverage)	0.120	0.069	-0.355***
	(0.117)	(0.163)	(0.133)
ln(population)	-0.096	-0.240*	0.296**
	(0.101)	(0.141)	(0.116)
ln(area in km^2)	0.031	0.082	-0.076
	(0.064)	(0.089)	(0.073)
ln(radio sets + TV sets per capita)	0.094	-0.014	-0.188
	(0.137)	(0.192)	(0.157)
intercept	-0.986	0.429	1.004
	(1.132)	(1.578)	(1.292)
N	112	112	112
F	3.98***	3.27***	3.29***
adjusted R-squared	0.2871	0.2344	0.2364

[a] OLS estimates. Heteroskedastic-consistent standard errors in parentheses. * p <0.10; ** p <0.05; *** p <0.01.

Table 3. Simulation results for Marxism-Leninism and Catholicism - frequency of misleading results[a]

	Marxism-Leninism			Roman Catholicism		
	m=1	m=0.001	m=1,000	m=1	m=0.001	m=1,000
$p1$=20; $p2$=0.025	189	212	216	816	821	833
$p1$=15; $p2$=0.025	156	167	167	796	786	786
$p1$=10; $p2$=0.025	94	86	98	605	569	573
$p1$=7; $p2$=0.025	50	47	67	355	341	357
$p1$=5; $p2$=0.025	15	31	31	207	211	250
$p1$=3; $p2$=0.025	15	18	11	122	115	127
$p1$=1; $p2$=0.025	9	3	7	68	76	73
$p1$=0.5; $p2$=0.025	6	4	2	57	47	68
no bias	5	3	2	41	48	46

[a] For Marxism-Leninism the misleading result is a positive, statistically significant coefficient for Rater #1 combined with a negative, statistically significant coefficient for Rater #2. For Catholicism the misleading result is a positive, statistically significant coefficient for Rater #1 combined with a non-significant coefficient for Rater #2. Statistical significance is defined based on a p-value lower than 0.10.

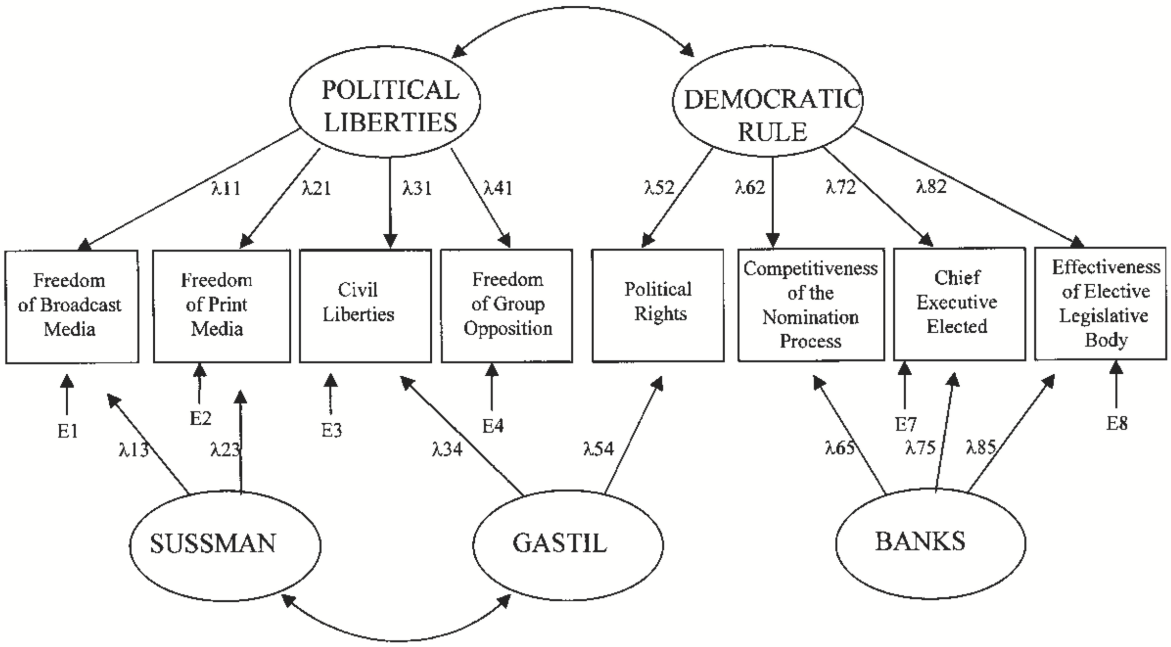
Table 4. Simulation results for Protestantism and monarchy - frequency of misleading results (except for the "no bias" row, which shows correct results)[a]

	Protestantism			monarchy		
	m=1	m=0.001	m=1,000	m=1	m=0.001	m=1,000
no bias	791	816	816	794	795	791
c1=30%; *c2*=35%	737	697	706	661	655	635
c1=50%; *c2*=55%	617	587	589	523	503	532
c1=70%; *c2*=75%	542	499	502	400	354	390
c1=130%; *c2*=135%	265	279	288	135	110	135
c1=150%; *c2*=155%	214	207	205	103	94	79
c1=170%; *c2*=175%	164	177	138	51	55	39
c1=190%; *c2*=195%	119	122	114	40	35	38
c1=230%; *c2*=235%	71	82	74	22	19	9

[a] For both variables the misleading result is a combination of non-significant coefficients for both Rater #1 and Rater #2. Statistical significance is defined based on a p-value lower than 0.10. Unlike the other rows, the "no bias" row does not show misleading results: it simply shows how often we obtain no evidence of bias when there is indeed no bias.

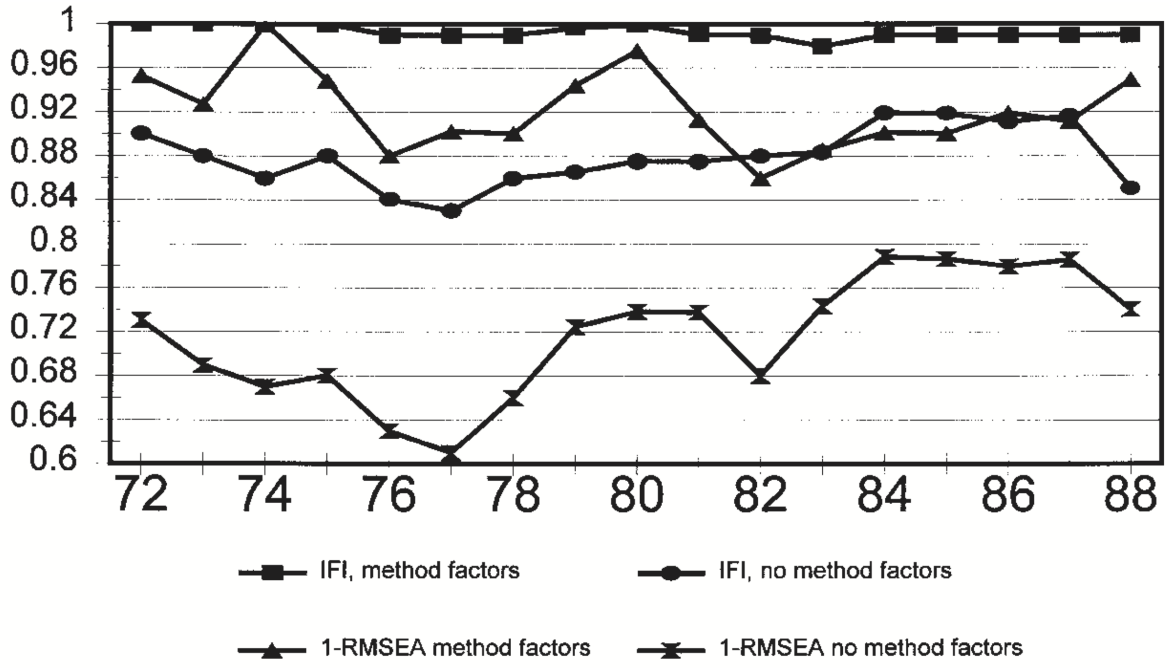
Figures

Figure 1. Bollen and Paxton's model



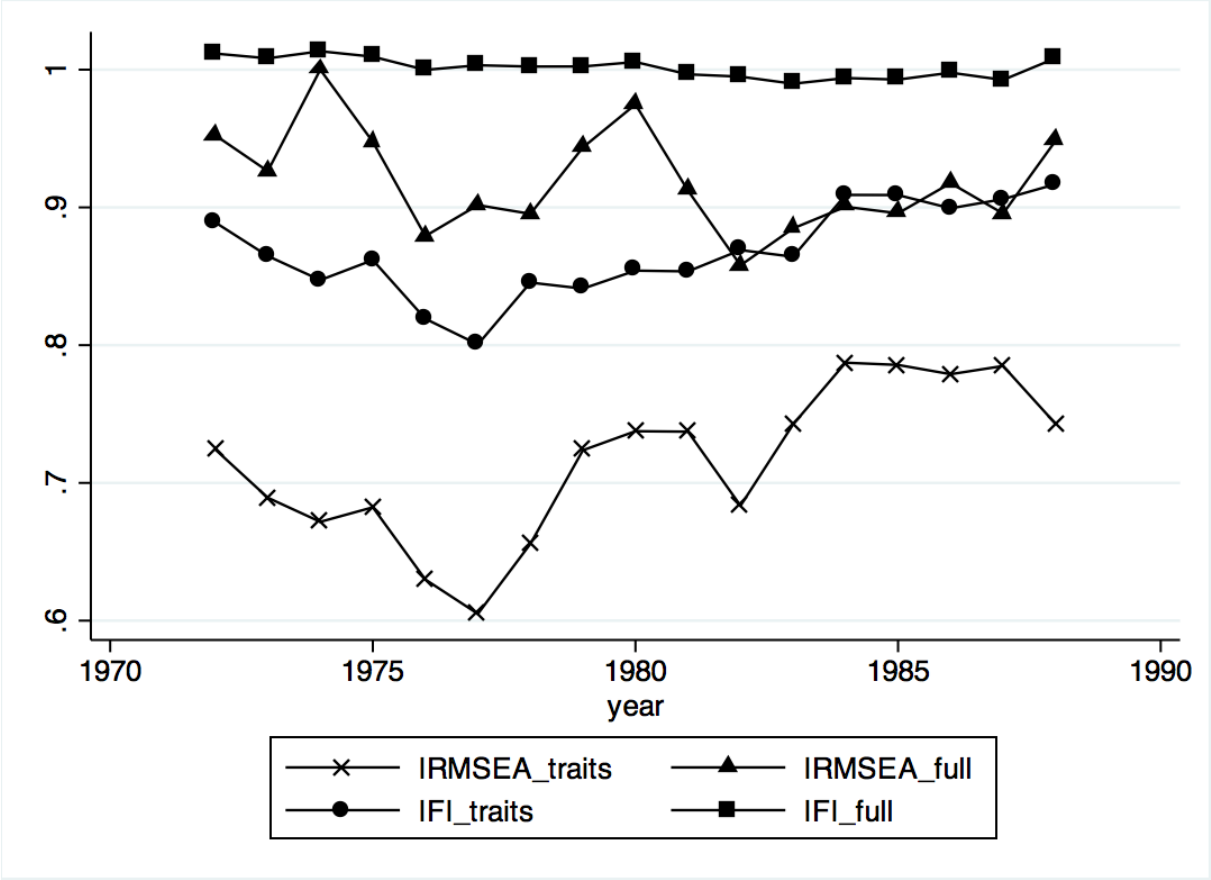
Source: Bollen and Paxton (2000, p. 65).

Figure 2. Bollen and Paxton's fit statistics



Source: Bollen and Paxton (2000, p. 67). IRMSEA stands for 1-Root Mean Square Error of Approximation and IFI stands for Incremental Fit Index.

Figure 3. Fit statistics from my replication of Bollen and Paxton



Source: My own estimations. These estimates are essentially identical to those in Bollen and Paxton (2000, p. 67). IRMSEA stands for 1-Root Mean Square Error of Approximation and IFI stands for Incremental Fit Index.