Gradations of Democracy? Empirical Tests of Alternative Conceptualizations

Zachary Elkins University of California, Berkeley

A group of influential scholars has argued emphatically that democracy should be measured dichotomously. This position challenges—on both theoretical and methodological grounds—the widespread practice of measuring democracy with graded scales, a practice which has been endorsed by leading methodologists who study democracy. This article proposes several empirical tests that evaluate the competing strategies. The evidence suggests that, on the whole, graded measures have superior validity and reliability. Hence, we should understand that specific cases correspond to the concept of democracy to varying degrees—degrees that can and should be measured.

Should scholars use intermediate categories to measure differences between democratic and nondemocratic regimes? In a series of influential studies, Przeworski, Alvarez, Cheibub, and Limongi describe this practice as "ludicrous" and insist on dichotomous measures (Alvarez et al. 1996, 21; Przeworski et al. 1996; Przeworski and Limongi 1997, 178–179). Their position, which is shared by other prominent scholars (e.g., Linz 1975, 184–185; Huntington 1991, 11–12), is surprising. An insistence on dichotomous measures appears to neglect the advances in data collection and analysis that would allow for the more precise measurement of gradations (Bollen and Jackman 1989, 616–619). Also, their position seems insensitive to the incremental, and sometimes partial, process that characterizes many democratic transitions. Thus, dichotomous measures appear both methodologically regressive and lacking in face validity.

This commitment to dichotomies in light of what seem like clear disadvantages has widened an important division among scholars about the conceptualization of democracy. Przeworski et al.'s argument, which is representative of the dichotomous view, rests on two logically independent claims—one about validity and one about reliability. Their *validity claim* is that democracy is first a question of kind before it is one of degree, and we cannot measure the degree of democracy across different "kinds" of regimes (Alvarez et al. 1996, 21–22).[1] The solution, according to this logic, amounts to the well-known social science maxim to classify before quantifying (e.g., Sartori 1970, 1036–1040). Their *reliability claim* is that, *even if* it made sense to measure gradations of democracy, dichotomous measures would be preferable because they contain less measurement error than do graded measures (Alvarez et al. 1996, 31). In sum, they argue that efforts to

Zachary Elkins is a Ph.D. Candidate, Department of Political Science, University of California, Berkeley, 210 Barrows Hall # 1950, Berkeley, CA 94720-1950 (zelkins@socrates.berkeley.edu).

I am indebted to David Collier for extensive comments and suggestions. Kenneth Bollen, Jake Bowers, Henry Brady, Gregory Caldeira, Jack Citrin, Steven Finkel, Steven Levitsky, Elizabeth Lilliott, Jules Reinhart, Sally Roever, John Sides, Beth Simmons, and the anonymous reviewers were also very helpful. Larry Diamond and Terry Karl graciously invited me to present part of this analysis at the Stanford Democratization Seminar.

[1]Przeworski et al. do not object to measuring the degree of democracy in cases that they classify as "democracies."

American Journal of Political Science, Vol. 44, No. 2, April 2000, Pp. 287–294

©2000 by the Midwest Political Science Association

look for traces of democracy in "nondemocracies" are both invalid and excessively error-prone.

This essay conducts a set of validity and reliability tests, using cross-national data on democracy and its correlates, to evaluate Przeworski et al.'s claims. The background assumption is that these scholars have made an outstanding contribution to the comparative study of democracy, and the objective is to advance and refine this important research program. The results presented here favor graded measures on two counts. First, construct-validity tests suggest that graded measures conform most closely to the explanatory role that social scientists have theorized for democracy. Second, a simulation, together with a clarification of the factors that affect measurement error, demonstrates that graded measures will be more reliable in most circumstances.

Construct Validity of Graded and Dichotomous Measures

We can say a measure of democracy has *construct validity* if the measure is a good predictor of phenomena that are widely hypothesized to be associated with democracy.[2] In the case of graded and dichotomous measures, their relative degree of construct validity will depend upon the predictive power one gains by measuring gradations within would-be dichotomous categories. If changes in the degree of democracy are just as meaningful for "nondemocracies" as they are for "democracies," then it makes sense to measure the degree of democracy even among the "nondemocracies." In order to evaluate the construct validity of the measures, I have chosen two domains of study—international conflict and regime stability—for which we have clear expectations about the behavior of democracies.

Democracy and International Conflict

The war record of democracies is both well known and well theorized. Notwithstanding a healthy degree of skepticism that accompanies anything resembling an empirical law, a large literature attests to the finding that democracies do not fight other democracies (Maoz and Russett 1992; Ray 1995; cf. Layne 1994; Farber and Gowa 1997). International conflict is therefore a promising outcome with which to assess the construct validity of competing measures of democracy. In particular, two questions are relevant. First, which measures predict conflict better, dichotomies or gradations? Second, how does the marginal effect of democracy on conflict vary among countries at different levels of democracy?

Empirical tests of the democratic peace hypothesis are legion. For the purposes of this article I have replicated the analysis of Rousseau et al. (1996), who use a standard graded measure of democracy developed by Gurr (1990). Rousseau et al. specify a model in which the initiation of force depends on three democracy variables: Actor's Democracy, Opponent's Democracy, and an interaction term composed of the two. The authors construct twenty-point scales for the variables Actor's Democracy and Opponent's Democracy by combining Gurr's autocracy and democracy scales from the Polity II dataset. Their interaction term, Actor's and Opponent's Democracy, is the product of Actor's Democracy and a dichotomized Opponent's Democracy (0 if opponent is nondemocratic, 1 if democratic). The product, therefore, equals the value of the variable Actor's Democracy when the actor's opponent is democratic and equals zero when the opponent is nondemocratic. Rousseau et al. include another set of variables in the model which tests alternative propositions generated from realist theories of conflict. In such a model, an insignificant effect for Actor's Democracy and a significant negative effect for the interaction term lend support to the widely held hypothesis that, while democracies may be as conflict prone as nondemocracies, democracies are not likely to fight each other (the dyadic hypothesis). A significant positive effect for Opponent's Democracy suggests that nondemocracies pick on democracies more than they do fellow nondemocracies.

The second column of Table 1 shows the results of a logistic regression of initiation of force on the set of predictors that include the polychotomous democracy variables. Figure 1 presents the logit coefficients as transformed into marginal probabilities; that is, the effect on the probability of initiating force of m̄oving from the lowest value of the independent variable to a given value. The results provide strong support for the dyadic hypothesis: democracies are comparatively peaceful, but only when they pair off against other democracies.

Figure 1 helps us evaluate Przeworski et al.'s claim that varying levels of democracy are not meaningful among nondemocracies. The results indicate that the variance in democracy *does* make a difference at each level of the scale. The effect on the probability of initiating force either increases (Opponent's Democracy) or decreases (Actor's Democracy and Actor's Democracy * Opponent's) across the entire range of each variable.

[2] This idea of validity is closely related to what scholars have termed "predictive" or "nomological" validity (Zeller and Carmines 1980, 78–84).

TABLE 1 Effect of Democracy on Initiation of Force Logistic Regression

Independent Variable	Graded Model	Dichotomous Model
Actor's Democracy	-0.03 (0.02)	-0.25 (0.27)
Actor's Democracy * Opponent's	-0.09** (0.03)	-0.51 (0.66)
Opponent's Democracy	0.05** (0.01)	0.17 (0.24)
Balance of Forces	1.23** (0.38)	1.19** (0.37)
Shared Alliance Ties	-0.01 (0.27)	-0.09 (0.20)
Satisfaction with the Status Quo	-3.36** (0.38)	-3.42** (0.38)
Constant	-0.29 (0.32)	-0.13 (0.25)

Note: N = 606 dual conflicts

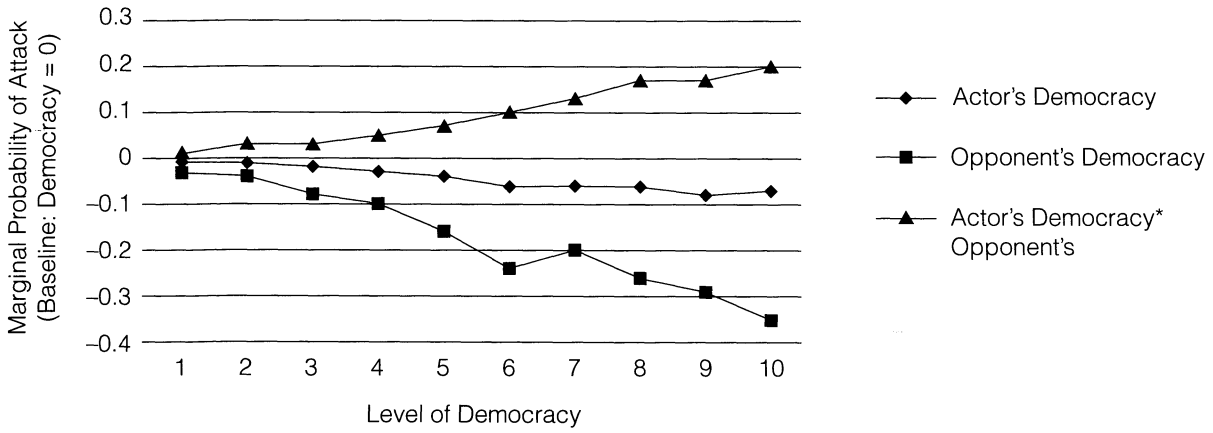
** $p < .01$

* $p < .05$

Standard errors in parentheses

Scale of the Three Democracy Variables
in the Graded model: 0–10
in the Dichotomous model: 0–1

FIGURE 1 Marginal Effects of Democracy on Inititation of Force (Calculated from Logistic Regression)



Note: These probabilities are derived from coefficients from the Graded Model in Table 1 (for a description of the procedure, see Elkins 1999). I report the marginal probabilities for levels of democracy on a ten-point scale and not the twenty-point scale used in the regression only because many of the intervals on the latter are sparsely populated in our sample of 606.

Even among the "nondemocracies" (those at the lower end of the scale), increases in democracy have an impact.

Since the degree of democracy seems to matter, what are the consequences of carrying out Rousseau et al.'s analysis with a dichotomous measure of democracy? The first step towards an answer is to construct a comparable dichotomy. I have chosen to call "democracies" those cases with at least a sixteen on the Rousseau et al. twenty-point scale. Such a cut point maintains the integrity of large groupings at either end of the scale and corre-

sponds closely to Przeworski et al.'s dichotomy. In fact, the new classification agrees with Przeworski et al.'s in 92 percent of the 369 cases for which their samples overlap.

How does this dichotomous version of democracy perform in Rousseau et al.'s model? Results in Table 1, column 2, suggest that the dichotomy obscures any effect democracy might have, either dyadic or monadic. None of the coefficients is statistically different from zero at any defensible level of significance. Furthermore, even if these variables *were* statistically significant, their effects are substantially underestimated in comparison with those in the tests using the graded measure of democracy. The marginal effects of a move from "nondemocracy" to "democracy" are clearly less pronounced than those predicted by the graded measure. For example, the move from 0 to 1 on the dichotomous interaction term lowers the probability of attack by nine percentage points, whereas a move across the full range of the polychotomous interaction term lowers the probability by thirty-five points. The international conflict evidence, then, suggests that graded measures of democracy produce findings that fit well with the way many social scientists expect democracies to behave. In other words, graded measures exhibit superior construct validity.

Democracy and Regime Persistence

The effects of varying degrees of democracy will not always be linear. Consider democracy's relationship with regime stability, which we can operationalize here as Gurr's (1990) variable "persistence" (the "number of years since the last fundamental, abrupt policy change"). Variation in the degree of democracy could well have different effects on the stability of "lower-level" democracies than on that of "higher-level" democracies (Remmer 1996, 624). One plausible hypothesis would be that democracy's relationship with stability is U-shaped. That is, increases in democracy decrease the probability of survival of "lower-level" democracies, but increase that of "higher-level" democracies. Given such nonlinearity, sorting cases into two classes, instead of degrees of democracy, might make sense.

One way to explore this possibility is to regress a measure of regime stability on both a graded and nongraded measure of democracy. Since Przeworski et al.'s sample matches well with those countries covered by Gurr's (1990) measure of regime "persistence," it is possible this time to use data collected by the scholars who favor dichotomies. One complication associated with this approach, however, concerns the availability of a comparable graded measure. Collapsing graded scales to dichotomous scores (as I did earlier) is one thing; manufacturing polychotomous data from a set of dichotomous scores is considerably more challenging. No simple mathematical transformation will allow us to derive graded categories from dichotomous measures. Nevertheless, even dichotomous measures of democracy are based on multiple attributes, and if the coders have explicitly identified these ingredients, it is possible to generate measures with intermediate categories. Since Przeworski et al. document their coding scheme, we can construct a polychotomous scale on the basis of the criteria that the authors used to produce their dichotomous scores (see Elkins 1999).

To assess the relative performance of the dichotomous and graded measures, I regress regime persistence on the two measures separately (Equations 1 and 2) and then together (Equation 3). Equations 1 and 2, admittedly underspecified, suggest that regimes last an extra six years for every one-unit increase in democracy in the graded measure as opposed to an extra twenty-two years if they are coded as democracies rather than nondemocracies in the dichotomous scheme. Since the graded measure is a four-category scale, the maximum effect (going from 0 to 3) is approximately eighteen years, four less than that predicted by the dichotomous measure. Not only does the dichotomous measure yield a larger effect, but the goodness-of-fit is better as well. Furthermore, the third model which includes both measures not only confirms that the dichotomous measure explains more, but also suggests that within at least one dichotomous category, the levels of democracy may be inversely related to regime persistence. That is, the democracies may be more stable than nondemocracies but, within these categories, an increase in democracy may actually *decrease* the probability of stability.

Equation 4 explores further this pattern of nonlinearity. Here, regime persistence is regressed on a series of dummy variables created from the values of the graded scale, with 0 as the residual category. The regression coefficients for these variables, then, represent the average change in a regime's longevity associated with a move from 0 to a given value on the graded scale. This specification reveals a U-shaped relationship between democracy and stability. Increases in democracy at the lower end of the scale actually decrease a regime's probability of survival, and it is only when regimes arrive at full democracy that their lifespans increase.

In essence, increases in democracy imply a different set of consequences for "democracies" than they do for "nondemocracies." It should be clear, however, that these results should not lead us to abandon graded measures.

TABLE 2 The Effect of Democracy on Regime Lifespan (in Years) OLS Regression

Variable	Equation 1	Equation 2	Equation 3	Equation 4
Constant	19.00** (1.28)	22.49** (0.78)	28.87** (1.43)	32.26** (1.62)
Graded-Przeworski et al.	6.11** (0.59)		-5.13** (0.97)	
Dichotomous- Przeworski et al.		21.80** (1.28)	30.83** (2.13)	
Dummy 1				-13.71** (1.62)
Dummy 2				-12.71** (1.97)
Dummy 3				12.03** (1.91)
R^2	0.02	0.05	0.05	0.06

Note: N = 5593 (country years from 1900 to 1986)

** $p < .01$

* $p < .05$

Standard errors in parentheses

Scale: Dichotomous Przeworski et al. (0–1)
Graded Przeworski et al. (0–3)

Dummy variables above are calculated from the Graded-Przeworski et al. scale with 0 as the residual category. The numbered dummies correspond to their positions on the scale. For example, for the variable Dummy 2, cases with the value 2 on Graded Przeworski et al. are scored 1 and all others are scored 0.

While the effect of democracy on regime persistence is nonlinear, the degree of democracy within both categories has an appreciable effect on a regime's lifespan. Indeed, the interesting debilitating impact of democratization on low-level democracies is apparent only if we measure gradations of democracy across regime types.

Furthermore, a reliance on a dichotomous measure in nonlinear models places great importance on the choice of cut point between categories. The analysis of coups, a phenomenon closely related to regime persistence, illustrates this point. A logistic regression of the probability of "coups"[4] on the same three combinations of variables discussed above reveals a relatively poor fit for Przeworski et al.'s dichotomous measure. Whereas the graded results suggest that the probability ranges fifteen points across the four categories, the dichotomous measure registers a maximum change of only five points. Furthermore, when the two measures are included in the same equation, the graded version maintains its explanatory power, while the dichotomous version is insignificant, both substantively and statistically. As with the regime-stability model, the effects of democracy on coups are decidedly nonlinear. This time, however, Przeworski et al.'s dichotomy, which sets a high (or, more importantly, *different*) cut point for democracy, is an inferior predictor of the outcome. For dichotomous measures, this finding highlights the difficult, but crucial, burden of identifying the relevant threshold between categories—a burden which, by definition, does not impair graded measures.

What, then, can we conclude from the regime persistence data? For one thing, democracy does exhibit threshold effects in some relationships—effects that may incline us to speak of two classes of regimes. However, even within these dichotomous categories, gradations of democracy have meaningful effects that appear to be worth estimating. Moreover, the regime persistence and coup examples show that the location of the threshold between democracies and nondemocracies is critical to making causal inferences. Locating this cut point is not a

[4]Banks defines coups as "an extraconstitutional or forced change in the top government elite and/or its effective control of the nation's power structure in a given year" (Gurr, 1990).

trivial matter at all; indeed, in the case of regime persistence, it is literally a matter of life and death. It is critical that researchers avoid lumping and splitting cases in such a way that masks a causally relevant threshold.

The Reliability of Graded and Dichotomous Measures

Przeworski et al.'s other argument is that a dichotomous measure will be more reliable than a graded measure *even if* the true nature of democracy is continuous. The claim is counterintuitive. If the construct in question really varies continuously, one might expect that a more fine-grained measure would capture this variation with greater reliability. I am convinced that Przeworski et al.'s claim results from an incomplete conception of measurement error, a conception which deserves clarification.

On one level, it is a relatively simple matter to show that a graded measure of a continuous phenomenon will be more reliable than a dichotomous one. Suppose that the "true" scores for democracy across nations can be arrayed on a scale from 0 to 10 and that Gurr's (1990) variable, Institutionalized Democracy, is known to have recorded the "truth" for each and every observation. Two researchers then set out to measure the "truth," one using two categories (a dichotomy), and the other five. Assume for the moment that both researchers are able to score each of the cases correctly. That is, each researcher sorts the cases into the correct category according to their true level of democracy. We now have three scores for each case: the true score and the scores recorded by the two researchers. With this information, we can easily assess the comparative reliability of the two scales with the standard measure of reliability,

$$1 - [\text{Var(Error)}/\text{Var(Total)}],$$

or the ratio of the observed variance to the total variance (see Zeller and Carmines 1980, 49).[5] It is no surprise to find that the five-category measure is more reliable (0.95 to 0.84).[6]

Of course, the scores in the example above are, by definition, free from measurement error by the coder. That is, in both scales each case is assigned to its appropriate category. Przeworski et al.'s perfectly reasonable assumption, however, is that cases will be miscategorized and that multi-category measures will miscategorize to a greater extent than will two-category measures. This expectation is appropriate. After all, it is easier to sort cases into two classes than into five. How can we represent such measurement error in this example? One way is to introduce noise into Gurr's measure of Institutionalized Democracy. Suppose, for example, that the closest coders come to ascertaining the true value of Institutionalized Democracy is within one unit. Also, suppose that the error is random; measures consistently come within one unit of the true score, but the deviations to one side or another are unsystematic. I produce this pattern of error by randomly adding or subtracting one unit from the true score for each case. These distorted observations can now be sorted into two- and five-category scales in the same way as before. As expected, the five-category scale miscategorizes many more cases than does the dichotomous scale (34.5 percent as opposed to 3.3 percent). Yet how has this error affected the respective reliabilities of the two measures? Interestingly, the five-category measure maintains its reliability edge (0.87 to 0.82).

Under what circumstances will the two-category scale demonstrate superior reliability? Several scenarios present themselves. First, Przeworski et al. rightly contend that dichotomous measures are advantaged by a bimodal distribution of the true scores. The intuitive explanation for this is simple: given one-unit distortions, dichotomous measures will err only when cases surround the cut point between classes, whereas graded measures may misclassify cases at every point on the scale. Fortunately, the distribution of Institutionalized Democracy is actually bimodal, which means that our previous assessment in fact considers the effect of that type of distribution. Thus, even when circumstances favor dichotomies—a bimodal distribution in this case—graded measures are more reliable.

However, a second condition concerns the magnitude of error in the perception of Institutionalized Democracy. In the previous example, I distorted the measure of the construct by one unit. What happens when the noise increases? A dichotomous measure may be expected to have an increasing comparative advantage in reliability as the magnitude of error rises. Indeed, distorting the scoring of Institutionalized Democracy by two units confirms this suspicion. The reliability of the five-category measure plummets to 0.51, while that of the two-category measure drops only to 0.69.

Nevertheless, my disagreement with Przeworski et al. does not depend on the magnitude of error. Przeworski et al. maintain that dichotomous measures produce less error when the probability that the scores are distorted

[5] This value is R^2 in models that regress the true score on each of the observed scores.

[6] The data in this illustration includes 182 countries from 1800 (or founding date of country) to 1986 for an N of 12,450.

by *one* point is 0.2. In my formulation, the polychotomous scale is more reliable even when *every score is distorted by one point.* These divergent findings most likely result from different conceptions of error. Przeworski et al. calculate the expected error with three factors: the probability of an error of a given magnitude, the magnitude of the error, and the number of such errors. In essence, Przeworski et al. measure the error variance. However, this formulation is incomplete. In order to determine the reliability of a measure, it is essential that one compare the *error variance* to the *total variance.* A polychotomous measure will almost certainly have more error variance than a dichotomous measure (at the limit, an infinitely categorized measure will miscategorize every value). However, a polychotomous measure will also produce more total variance, the critical term in the denominator of the reliability equation.

It bears repeating here that adopting this standard notion of reliability does not necessarily privilege graded measures. The reliability is conditional on the number and magnitude of errors, as well as the corresponding sensitivity of our measure. Increased sensitivity comes at the cost of increased error. If we can assume that a construct reveals itself in gradations when we observe cases, then it makes sense to record these gradations with as sensitive a measure as possible. How does one then achieve the proper balance between sensitivity and error? The example above suggests that if we cannot sort more than 60 percent of cases into the correct five categories, then we are better off with a dichotomous measure. However, in reality, we are not blessed with knowledge of the true values of democracy. The number of categories must in all likelihood be determined by the measurer's judgment as to what constitutes a reasonable balance of sensitivity and error. This judgment, of course, can and should be informed by empirical tests like those above.

Admittedly, there may be other reasons to prefer dichotomies to polychotomies. For example, combining attributes to form an ordinal scale requires assumptions that may prove untenable in some cases (Collier and Adcock 1999; Gleditsch and Ward 1997). With respect to measurement error, however, the point is clear: graded measures are not inherently less reliable.

Conclusion

Democratization studies lead us to believe that there is substantial variation in the *degree* of democracy across both time and space. The empirical tests in this article confirm that such variation is meaningful and can be measured reliably. More specifically, construct-validity tests based on hypotheses focused on international conflict and regime endurance demonstrate that measures of democracy which provide for gradations best fit the behavior that theoretical work on democracy would predict. Furthermore, a close look at the factors that lead to measurement error suggests that graded measures will exhibit superior reliability. In short, looking for traces of democracy in seemingly "nondemocratic" regimes makes good theoretical and methodological sense.

Manuscript submitted November 9, 1998.
Final manuscript received September 28, 1999.

References

Alvarez, Michael, José Antonio Cheibub, Fernando Limongi, and Adam Przeworski. 1996. "Classifying Political Regimes." *Studies in Comparative International Development* 31:3–36.

Bollen, Kenneth A., and Robert Jackman. 1989. "Democracy, Stability, and Dichotomies." *American Sociological Review* 54:612–621.

Collier, David, and Robert Adcock. 1999. "Democracy and Dichotomies: A Pragmatic Approach to Choices About Concepts." In *Annual Review of Political Science,* Vol. 2, ed. Nelson W. Polsby. Palo Alto: Annual Reviews.

Elkins, Zachary S. 1999. "Getting Blood from a Stone: Constructing Intermediate Categories from Dichotomous Measures." APSA Organized Section for Political Methodology Web Site. http://polmeth.calpoly.edu/.

Farber, Henry S., and Joanne Gowa. 1997. "Common Interests or Common Polities? Reinterpreting the Democratic Peace." *Journal of Politics* 59:393–417.

Gleditsch, Kristian S., and Michael D. Ward. 1997. "A Reexamination of Democracy and Autocracy in Modern Polities." *Journal of Conflict Resolution* 41:361–383.

Gurr, Ted R. (1989) 1990. *Polity II: Political Structures and Regime Change, 1800–1986* [Computer File]. Boulder, Colo: Center for Comparative Politics [producer]. Ann Arbor, Mich.: Inter-University Consortium for Political and Social Research [distributor].

Huntington, Samuel P. 1991. *The Third Wave: Democratization in the Late Twentieth Century.* Norman: University of Oklahoma Press.

Layne, Christopher. 1994. "Kant or Cant: The Myth of the Democratic Peace." *International Security* 19:5–94.

Linz, Juan J. 1975. "Totalitarian and Authoritarian Regimes." In *Handbook of Political Science,* ed. Fred Greenstein and Nelson Polsby, 3:175–353. Reading, Mass.: Addison-Wesley.

Maoz, Zeev, and Bruce Russett. 1992. "Alliance, Contiguity, Wealth, and Political Stability: Is the Lack of Conflict among Democracies a Statistical Artifact?" *International Interactions* 17:245–267.

Przeworski, Adam, and Fernando Limongi. 1997. "Modernization: Theories and Facts." *World Politics* 49:155–184.

Przeworski, Adam, Michael Alvarez, José Antonio Cheibub, and Fernando Limongi. 1996. "What Makes Democracies Endure?" *Journal of Democracy* 7:39–55.

Ray, James Lee. 1995. *Democracy and International Conflict: An Evaluation of the Democratic Peace Proposition.* Columbia: University of South Carolina Press.

Remmer, Karen. 1996. "The Sustainability of Political Democracy: Lessons from South America." *Comparative Political Studies* 29:611–635.

Rousseau, David, Christopher Gelpi, Dan Reiter, and Paul K. Huth. 1996. "Assessing the Dyadic Nature of the Democratic Peace, 1918–88." *American Political Science Review* 90:512–534.

Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64:1033–1053.

Zeller, Richard A., and Edward G. Carmines. 1980. *Measurement in the Social Sciences: The Link between Theory and Data.* Cambridge: Cambridge University Press.