The Vodka is Potent, but the Meat is Rotten[1]:
Evaluating Measurement Equivalence across Contexts

Zachary Elkins
Department of Government
University of Texas at Austin
zelkins@austin.utexas.edu

John Sides
Department of Political Science
George Washington University
jsides@gwu.edu

October 2009

Abstract

Valid measurement in cross-national research depends on the equivalence of the meaning of constructs and indicators across cultures, but thus far political scientists have typically assumed equivalence without any formal tests of such. We review different forms and sources of non-equivalence, and demonstrate methods of diagnosing problems. We develop hypotheses about non-equivalence related to two concepts—political involvement and democracy. Subsequent empirical tests with these items demonstrate non-equivalence, though of differing magnitude and of different form. We conclude with a general set of guidelines regarding the diagnosis and treatment of non-equivalence.

We thank Fritz Drasgow, Colin Elman, Chris Frederico, Gerry Munck, and Jonathan Wand for their comments and suggestions on earlier drafts, Sarah Gleisner for research assistance, and Eunice Chang, Svitlana Chernykh, Aya Kachi, Ika Putri, Zehra Toprak, and Sergio Wals for assistance with translation.

[1] Russian (mis)translation of the phrase, "The spirit is strong but the flesh is weak," from a cross-national survey item (Smith 2003).

> Until now, I'd assumed that there was a universal lexicon of meat terms (after all, a leg is a leg is a leg), which, like any other piece of language could be translated from one country to the next. The belief, I was now realizing, had been encouraged by those diagrams of a cow cut in half, the kind you sometimes see in cookbooks telling what a thing is in France, England, and America…One day, wanting to confirm a spelling, I consulted an Italian food encyclopedia…and discovered (under *bovino*) not three or four diagrams but pages of them, thirty in all, none in French or English but only Italian, broken down by region, each one different, no two cuts alike, with few shared terms. The Tuscan chart was dizzying. Every single tissue seemed to be identified. The thigh was a maze, like a road map of an impenetrable medieval city, with more names than there was space on the two-dimensional representational leg to accommodate. I understood why there were no obvious translations of *girello*, *campanello*, or *sottofesa*: because, outside Italy, they don't exist. Outside Tuscany, they rarely exist. I remembered how I'd researched the short rib and been surprised that the terms my butcher in New York used were so different from the ones known to a butcher in Edinburgh or Paris. But I'd understood only half of it: every country—and in Italy, every region and, sometimes, every town—has its own unique way of breaking an animal down to dinner-sized portions. Finally, I was getting it: there is no universal butcher language; *none of it is translatable*.
>
> —Bill Buford, *Heat*, p. 270 [italics in original]

> Sometimes a cigar is just a cigar.
>
> —attributed to Sigmund Freud

Scholars of comparative politics continue to venture into more and more jurisdictions across longer stretches of time. Some cross-national datasets contain the universe of independent states since 1800, and cross-national survey projects now include much of both the developed and developing worlds; the World Values Survey, now in its fifth wave, is administered in more than 80 countries. The benefits of such expansion is clear: added cases can produce more variation in variables of interest, provide more powerful tests of extant theory, and illuminate empirical puzzles that lead to new theories.

However, comparative inquiry grinds to halt if scholars cannot develop concepts and measures that mean the same thing across diverse national and historical contexts. This is true whether the purpose of comparative inquiry is descriptive or inferential. Basic questions of comparability thus arise. Will a survey question about party identification mean the same thing in the United States as it does in Britain? Is the gross domestic product, as reported by individual governments, equally valid in Sri Lanka and Costa Rica? Does a measure of democracy mean the same thing in 1850 as it does in 2000? Answering "no" to these questions—at least an emphatic "no"—seems devastating. If an observed attribute is a product not only of the true

underlying construct but also of a measurement irregularity particular to time or place, then inferences based on those observations are compromised. Thus the importance of equivalent measurement, the challenge we address in this paper.[2]

Concern about measurement non-equivalence is not new, of course. Most of chapter 2 of *The Civic Culture*, for example, is dedicated to justifying the equivalence of the authors' measures (Almond and Verba 1963; see also Anderson 1967; Rokkan, Verba, Viet, and Almasy 1969). Przeworski and Teune (1966) extend these themes, motivated by their participation in a five-country study of the values of local political leaders. But in the years since the initial waves of large-scale comparative projects, the problem of equivalence has receded to the background for most political scientists. As the pace of data collection has increased, the pace of research into equivalence has proceeded much more slowly. There are important exceptions, including Adcock and Collier's (2001) work on measurement validity, Bartels' (1996) work on pooling disparate observations, Brady's (1985, 1989) work on interpersonal incomparability, Herrera and Kapur's (2007) work on improving data quality, King and co-authors' work on anchoring vignettes (King et al. 2004; King and Wand 2007), and attention to contextual differences in the measures of constructs such as human values (Davidov, Schmidt, and Schwartz 2008), democracy (Bollen 1993; Treier and Jackman 2007), economic development (Sen 1999), ethnic diversity (Fearon 2003), federalism (Rodden 2004), identity (Abdelal et al. 2009), ideological ideal points of political elites (Bailey 2007; Epstein et al. 2007; Poole 1998), income levels and inequality (Smeeding, O'Higgins, and Rainwater 1990), nationalism and patriotism (Davidov 2009), political efficacy (King et al. 2004), political knowledge (Elff 2009; Mondak and Canache 2004) postmaterialism (MacIntosh 1998; Clarke et al. 1999), poverty (Brady 2003; Sen 1976), the proportionality of electoral systems (Gallagher 1991), satisfaction with democracy (Canache, Mondak, and Seligson 2001), and social capital (Paxton 1999)—although not all of these offer formal tests of equivalence. Nonetheless, we suspect that most analysts of secondary data largely set aside concerns about non-equivalence, choosing (understandably) to attend to issues directly under their control, such as estimation and model specification.

[2] The problem of poorly travelling measures goes under labels such as measurement non-equivalence and incomparability. We use these terms interchangeably here. The terms "equivalence" and "invariance" are often used synonymously in the literature, and we follow that convention as well.

A review of extant research finds inattention to questions of measurement (Bollen, Entwisle, and Alderson 1993), which contrasts with the attention paid by psychologists and, in particular, organizational and educational psychologists (see, inter alia, Hambleton, Merenda, and Spielberger 2005; Harkness, Van de Vijver, and Mohley 2003a). Adcock and Collier (2001: 534) write, "The potential difficulty that context poses for valid measurement, and the related task of establishing measurement equivalence across diverse units, deserve more attention in political science."

In sum, we do not fully understand the degree of non-equivalence in widely-used measures or how much non-equivalence affects causal inference. As Adcock and Collier (2001: 536) put it, "Claims about the appropriateness of contextual adjustments should not simply be asserted; their validity needs to be carefully defended." Our intention is to aid in this defense. We begin by conceptualizing the forms that non-equivalence takes. We then develop theory about the sources of such non-equivalence, focusing on non-equivalence across both "space" and "time." We then conduct some basic tests of non-equivalence, focusing on three important concepts in comparative politics: political involvement, democracy, and economic development. In so doing, we demonstrate how analysts can establish the presence of non-equivalence. Our intent is not to provide a methodogical treatise, but a basic overview of diagnostic tools and statistical tests for non-equivalence.

Finally, we offer suggestions for analysts using cross-national and over-time data. Sound advice for addressing non-equivalence is scattered across treatments of conceptualization and measurement. For example, Adcock and Collier (2001) offer a sensible set of solutions to the problem, such as the use of context-specific indicators, and work on anchoring vignettes suggests a clever means of improving comparability for survey items. We integrate these suggestions with others to provide some simple guidelines that constitute a reasonable starting place for researchers using measures across spatial or temporal contexts.

Conceptualizing Equivalence

We can conceptualize the forms of non-equivalence in terms of the parameters of the traditional measurement model (see Bollen 1989: 17). Suppose that a unit *i* (an individual, country, etc.) has an observed

value (x_i) for some indicator, which is linearly related to the underlying, unobserved (latent) attribute (ξ). The strength of the relationship between the latent attribute and the observed indicator is captured by a parameter λ, often referred to as the factor loading. The observed value of x also depends on an additional "uniqueness" parameter (δ, i.e., the residual). An intercept term μ, is the value of x when the latent attribute is equal to 0. To capture better the notion of equivalence, let us suppose that two equations are estimated, one for "Context 1" and another for "Context 2":

$$x_{i1} = \mu_1 + \lambda_1 \xi_{i1} + \delta_{i1} \quad (1)$$

$$x_{i2} = \mu_2 + \lambda_2 \xi_{i2} + \delta_{i2} \quad (2)$$

If the contexts were countries, an important goal would be to use the values of x in Countries 1 and 2 to make inferences about how much the latent attribute varies across countries. If the contexts represented time periods, then the parallel goal would be to make inferences about variation across time. Either task depends crucially on equivalence in measurement.

Equivalence can be conceived in hierarchical terms, starting with the fundamental forms and moving to the more subtle, with each form depending on equivalence at the prior level. Most fundamentally, the construct under investigation should have a similar meaning, however measured, in each context. This sort of equivalence—which sometimes goes under the label *construct equivalence* (Van de Vijver 2003) or *conceptual equivalence* (Hui and Triandis 1985)—pertains to conceptualization and even case selection, both of which precede operationalization and measurement. If construct equivalence does not hold, then, in essence, the latent attributes in equations (1) and (2) are different; instead of ξ_1 and ξ_2, we might imagine that the attributes are really φ_1 and ξ_2—i.e., apples and oranges. Cases of construct non-equivalence will be obvious to those with basic knowledge of the contexts in question. Examples might include efforts to measure attitudes towards the monarchy in societies that have never experienced such rule, or the strength of party organizations in settings in which parties are banned. Such measurement misadventures are cases of *conceptual stretching* (Sartori 1971; Collier and Levitsky 1997). One solution, as Sartori points out, is to climb the "ladder of abstraction" by, to take the examples above, asking respondents about systems of authoritarian rule instead

of monarchies, or political organizations instead of parties. The other solution, of course, is to retain the construct but avoid measuring it where it is meaningless or hopelessly stretched. As one might imagine, if construct equivalence is violated, so too will more subtle forms of equivalence. Thus, construct equivalence is of the utmost importance. Unfortunately, construct equivalence is not something that can be evaluated in any strict empirical sense, although its violation will be apparent in tests of more subtle forms of equivalence. Construct equivalence is best evaluated through careful conceptualization that is guided by theory and grounded in detailed knowledge about particular countries or cultures. As such, we do not address this type of equivalence in much depth, focusing instead on measurement more specifically, although evaluating construct equivalence remains an important task for researchers.

A second kind of equivalence is *structural equivalence*, in which the latent construct has a similar "architecture" across contexts.[3] That is, the observed indicators that measure the concept in one context overlap perfectly with those that do so in another context. At the extreme, a concept may have a unique empirical manifestation in one context, in which case none of the observables that tap that construct elsewhere would be relevant. However, it is more likely that a given set of indicators will introduce surplus meaning where or when the construct takes on a narrower meaning or under-represent the construct where or when it has a broader, multidimensional meaning.

Assume for the moment than a concept is measured with a five-indicator scale administered in two contexts. In equations (1) and (2) x_i is replaced with a 5×1 vector of indicators ($\mathbf{X}$) that is related to the same latent attribute ($\boldsymbol{\xi}$) by a vector of loadings ($\boldsymbol{\Lambda}$) and a vector of uniqueness parameters ($\boldsymbol{\delta}$):

$$\mathbf{X}_1 = \mu_1 + \boldsymbol{\Lambda}_1\boldsymbol{\xi}_1 + \boldsymbol{\delta}_1 \quad (3)$$

$$\mathbf{X}_2 = \mu_2 + \boldsymbol{\Lambda}_2\boldsymbol{\xi}_2 + \boldsymbol{\delta}_2 \quad (4)$$

Assume that in the first context, each of the indicators in $\mathbf{X}_1$ is a valid indicator of the latent attribute $\boldsymbol{\xi}_1$. Structural equivalence would not obtain if, for example, any of the indicators in $\mathbf{X}_2$ is not in fact an indicator

[3] According to Van de Vijver (2003: 154), "An instrument administrated in different cultural groups shows structural equivalence if it measures the same construct in all these groups." Similar notions appear elsewhere in the literature though they are denoted differently—Cheung and Rensvold's (2000) "factor form invariance" and Stark, Chernyschenko, and Drasgow's (2006) "configural invariance." Bollen (1989: 356) discusses invariance in terms of "model form."

of the latent attribute $\boldsymbol{\xi}_2$. The same is true if the indicators in $\mathbf{X}_2$ are actually indicators of two distinct factors rather than a single factor—in other words, if the latent attribute $\boldsymbol{\xi}_2$ is multi-dimensional in the second context but not the first. Przeworski and Teune's (1966) distinction between "equivalent" and "identical" items is instructive here, even though their terms differ somewhat from modern conventional usage. Their basic logic of comparison suggests that items do not have to be "identical" (i.e., exactly the same) across contexts, as long as they capture the same basic phenomenon. So, in their example, a measurement model of political activity in the United States that uses an indicator such as "giving money to campaigns" might be "equivalent" but not "identical" (in the Przeworski and Teune sense of those terms) to a Polish model that substituted "volunteer for social works" for the campaign item. Nevertheless, to return to the standard usage of "equivalent," the measurement model across these contexts would not be structurally equivalent, since the architecture of indicators differs.[4]

The possibility of structural non-equivalence surfaces in Canache, Mondak, and Seligson's (2001) study of the standard "satisfaction with democracy" survey item. They find that this indicator taps multiple dimensions—satisfaction with the current political system, satisfaction with incumbent political authorities, and support for democracy—and that its relationship to these dimensions varies across countries. For example, "satisfaction with democracy" is strongly associated with system support in Uruguay but not so in Costa Rica. In a study of system support, then, the inclusion of the item "satisfaction with democracy" would exhibit structural non-equivalence due to its near-irrelevance in the Costa Rican context.

A third form of invariance is *metric equivalence* (Stark, Chernyshenko, and Drasgow 2006). Metric invariance implies that the relationship between the observed indicators and the underlying attributes is the same across contexts. In equations (3) and (4), metric equivalence means that $\boldsymbol{\Lambda}_1=\boldsymbol{\Lambda}_2$; that is, the factor loadings are equivalent.[5] The notion of metric invariance is closely linked with the concept of *differential item functioning* (DIF), about which there is a large literature deriving mostly from educational testing (see Angoff

[4] In our view, construct non-equivalence and structural non-equivalence will likely go hand-in-hand much of the time. (Van de Vijver and Leung (1997:8-9) actually treat construct and structural equivalence as the same thing.) However, it is possible that a theoretical construct means the same thing in various contexts but that the "structural" relationship between that construct and a set of indicators differs across contexts.

[5] Cheung and Rensvold (2000) refer to this property as "factorial invariance."

1993 for an introduction and King et al. 2004 for a political science application). In the educational testing literature, DIF typically means that two students of equal aptitude have different probabilities of answering the same question correctly (e.g., because the question contains a cultural bias). In the social scientific literature, metric non-equivalence or DIF could mean, for instance, that one of the observed indicators is strongly associated with the latent attribute in one context, but only weakly associated with the attribute in the other context. To take a hypothetical example, if the underlying attribute were "love of sports," a question about how often one watched World Cup soccer matches might be strongly associated with love of sports in England, but only weakly so in the United States. Thus, an American and a Brit, both equally passionate about sports, would answer this question very differently. Metric non-equivalence could also arise in the context of survey research if a particular "response style" were more characteristic of one context than another. For example, in a cross-national survey, if respondents in one country were more likely to use the endpoints of standard Likert-type scales—a phenomenon known as extreme response style (Cheung and Rensvold 2000; Greenleaf 1992)—then this would likely affect the relationships between such indicators and an underlying attribute, relative to the estimated relationships in other countries.

While metric equivalence is distinct from structural equivalence, the two are related hierarchically in that the former presupposes the latter. Indeed, sometimes the border between the two standards of equivalence can be murky and an extreme case of metric non-equivalence might be better understood as structural non-equivalence. The "satisfaction with democracy" example described above is a case in point. Given that the relationship between the item and "system support" in Costa Rica is not quite zero ($r = .05$), one might conclude that the item does measure the latent construct in Costa Rica, albeit not to the extent that it does in Uruguay—thus leading to a conclusion of structural but not metric equivalence.

A fourth form of equivalence is *scalar equivalence* (Cheung and Rensvold 2000; Stark, Chernyshenko, and Drasgow 2006). It means that the "intercepts"—i.e. the value of the observed indicator when the latent attribute is 0—are equivalent across countries. If scalar equivalence does not obtain, we might expect to see a kind of scale displacement, whereby observed values in one context are systematically higher or lower than in the other context, even though the value of the latent attribute in both contexts is the same. This could arise,

for example, if the tendency to agree to survey questions (acquiescence response style) differed systematically across contexts, such as countries (Cheung and Resnvold 2000; Johnson et al. 2005; van Herk, Poortinga, and Verhallen 2004) or groups (Bachman and O'Malley 1984). If the denizens of one country are more likely to be yea-sayers when asked survey questions—tending to agree with whatever is put in front of them—then on average they will appear "higher" on a given indicator relative to respondents other countries. Scalar non-equivalence could also arise if respondents in different contexts have different "reference points" in mind when they evaluate themselves. Indeed, the purpose of the anchoring vignettes is to locate, or "anchor," respondents to a common reference point by asking them to evaluate the individuals described in the vignettes in addition to evaluating themselves (King et al. 2004). For example, the World Values Survey asks respondents if they "discuss political matters frequently, occasionally, or never?" One suspects that "frequently" may mean one thing in the United States, where politics is often a topic to be avoided in social conversation, and quite another in Israel where political conversation is virtually unavoidable. Another example of different reference points—in this case across individuals rather than across countries—comes from the study of "feeling thermometers" as a measure of affect towards societal groups. Wilcox, Sigelman, and Cook (1989) show that some people tend to assign higher, or "hotter," scores regardless of the group. A national-level example of this problem, and one we investigate empirically below, comes from one of the standard measures of democracy. The United States has scored the maximum of 10 on Polity's measure of democracy since the civil war. But is a 10 in 1870 equivalent to a 10 in 2000? Almost undoubtedly not, unless one discounts the inclusion of women and minorities in the political process. There appears, then, to be a serious problem of deflation in the Polity measures that require adjustment.

A final form of equivalence involves the variance of the residuals in each of the measurement equations. Thus, from equations (3) and (4), $\text{var}(\boldsymbol{\delta}_1)=\text{var}(\boldsymbol{\delta}_2)$. This sort of non-equivalence can arise, for example, in cross-national survey research when the translation of an item does not necessarily change its meaning but is so poor that respondents have trouble interpreting the meaning. As such, answers might display more error variance, which is unsystematic and therefore uncorrelated with the factor weights ($\boldsymbol{\Lambda}$). An example concerns the passage from the Gospel of Matthew in this paper's title. Imagine that Americans

were asked (admittedly in florid terms) which takes precedence in human behavior, our strong spirits or weak flesh. Now imagine Russians puzzling through some nonsense about "potent vodka" and "rotten meat." If "fuzzy questions" create measurement error (Achen 1975), then the Russian version is a woolly beast indeed.[6]

The various kinds of equivalence draw our attention to the complexities of measurement across contexts and to the potential difficulties when equivalence does not hold. Standard procedure in the analysis of cross-contextual data—which typically entails simply taking well-known indicators (e.g, Polity democracy scores) "off the shelf," or selecting a set of indicators and constructing a summary scale, perhaps with minimal diagnostics within each context to ensure uni-dimensionality (an exploratory factor analysis) or reliability (e.g., an alpha statistic)—are not sufficient to ensure valid and reliable measurement. In particular, non-equivalence may confound efforts to use context-level averages to discern the true effect of context on the attribute in question. The analogous situation in the educational testing literature is "differential test functioning" (DTF), which results from the systematic biases created by differential item functioning. In such cases, the true effect of group membership on the test score, also known as the "impact," is confounded by DIF. Only if differentially functioning indicators have off-setting biases, with some "favoring" one group and some "favoring" others, will DIF fail to create DTF.

Non-equivalence affects not only the estimated means of attributes, but also the relationships among those attributes. Brady (1985) shows that both metric and scalar non-equivalence can bias estimates of the relationship between observed indicators. For example, if a researcher wanted to understand the relationship between democracy and economic development across countries, non-equivalence is just as important as if this researcher wanted to compare mean levels of each attribute. Equivalence in measurement is critical for both descriptive and causal inference.

---

[6] However, measurement error non-equivalence—as well as that of other parameters in the measurement model, such as the means of and covariances among the latent variables—is arguably less crucial than are other kinds of non-equivalence. As Stark, Chernyshenko, and Drasgow (2006: 1293) summarize, "generally, there is no clear prescription about the order or need for performing these tests [of non-equivalence]." Similarly, Raju, Laffitte, and Bryne (2002: 518) note that testing for this sort of equivalence "is neither necessary nor of particular interest when the observed scale scores are used merely as CFA [confirmatory factor analysis] indicators and not as measures in their own right."

Sources of Non-Equivalence: The Woods-Jordan Problem and the Bonds-Ruth Problem

As important as classifying the forms of non-equivalence is understanding its sources. A lack of theory about these sources is an important shortcoming of research to date. As Vanderberg (2002: 152) notes, "there has been little attempt to predict a priori what factors result in a failure to support invariance." Thus, investigations of equivalence are typically exploratory. Their goal is simply to evaluate the measure, not to understand why particular items are problematic. While there is no way to identify every possible source of non-equivalence *a priori*, a general categorization of the major sources is helpful. We discuss two classes of non-equivalence according to the dimension along which the items are invariant. The first involves cross-sectional units, such as places, groups, or individuals. Non-equivalence in this case gives rise to the "Wood-Jordan problem." The second involves temporal units; non-equivalence over time gives rise to the "Bonds-Ruth problem."

The Woods-Jordan Problem: Non-equivalence across Space

Who is the better athlete: Tiger Woods or Michael Jordan? The *New York Times* devoted most of a Sunday magazine issue to the question, which surely provoked conversations around water coolers nationwide. The question has no straightforward answer because Woods and Jordan played different sports, and the talents needed to excel in one sport are less relevant in the other. Woods does not need a jump shot. Jordan does not need a flop shot. Some measures have this same problem: simply put, they do not travel well from one location to another. These dislocating effects arise largely from differences across contexts in language, custom, and culture—that is, shared meanings, norms, and values. While we focus on non-equivalence at the country level, non-equivalence can be prevalent across other jurisdictional units, ethnic groups, social classes, or other entities defined by cultural, political, or economic markers.

One can imagine various manifestations of place-driven non-equivalence. First, the very manner of data collection may engender different reactions in different cultures. As Verba (1969) notes, the act of providing opinions about political issues to a complete stranger may prove unremarkable in some cultures but foreign in others. Second, as discussed earlier, differences across context may create a basic lack of construct

or conceptual equivalence because a particular attribute takes on different meanings across these contexts. Third, contextual differences also create logistical hurdles, such as the difficulty of constructing equivalent measures in different languages (e.g., Hambleton 2005; Iyengar 1976; Blais and Gidengil 1993).[7] Finally, culturally specific values may imply varying "response styles." Different places will have different norms and values that bear on the task of providing data. One example involves cross-cultural variation in psychological processes and value orientations that bear on survey response. For instance, there is a greater tendency toward a collectivist orientation in East Asian countries, and toward an individualist orientation in Western countries (e.g., Nisbett 2003; Triandis 1995; Lalwani, Shavitt, and Johnston 2006). A collectivist orientation is in turn associated with providing socially desirable responses (Lalwani, Shavitt, and Johnson 2006).

Moreover, we should not assume that response styles affect only citizens interviewed for survey projects. Herrera and Kapur (2007: 378) note that measuring quantities that are also "targets" (e.g., stated goals of governments) can give rise to a sort of yea-saying or nay-saying bias: "When targets are ceilings (such as fiscal deficits), the data are likely to have downward bias. When targets are floors (such as social sector indicators), the data are likely to be biased upward." Generally speaking, any actors charged with producing data—survey respondents, government functionaries in statistical agencies, managers in firms, academic experts, etc.—can exhibit biases of various kinds, and elites are not necessarily any less vulnerable than ordinary citizens (Tetlock 2005).

One may also observe the "Woods-Jordan" problem in employing a measure across groups. By group, we refer to a set of individuals who share common characteristics that are not necessarily defined by geographic boundaries, as is nationality. Examples include gender, ethnicity, and religion. These characteristics create non-equivalence because group members are socialized differently than are non-members. They are exposed to particular experiences, ideas, and expectations. They learn to act and think in

[7] Within cross-cultural survey research, a standard set of protocols has developed that helps minimize distortions, one of which is the practice of back-translation, in which the items, having been translated from the source to the target language, are subsequently (and independently) translated back to the source language (see Harkness 2003; Drasgow and Probst 2005). If the source and target items are equivalent, then the source and back-translated items should also be equivalent under the premise that "what goes in ought to come out" (Harkness 2003: 41)—though a single back-translation does not guarantee linguistic equivalence (see Anderson 1967) and, moreover, linguistic equivalence does not guarantee measurement equivalence (see Blais and Gidengil 1993).

distinct ways. Thus, to approach two different groups—even groups that are literal neighbors—with a common set of measures may be no different than taking that set of measures into two different countries. It is these kinds of group differences that underlie non-equivalence (or differential item functioning) in much of the educational testing literature.

One example concerns the well-known gender gap in knowledge of politics. Mondak and Anderson (2004) argue that this gap arises in part because men are more likely to guess, rather than answering "I don't know," when asked these questions. Citing educational testing literature, Mondak and Anderson note that "the self-confidence and competitiveness of male test-takers inflates their scores relative to those of women" (497)—traits that presumably derive from cultural expectations that are reflected in the socialization experiences of men and women. A second example is race of interviewer effects (e.g., Davis 1997). Here, non-equivalence arises not from group differences considered in relative isolation, but from the interaction between group members in the context of a survey interview. The well-known finding is that the some black and white respondents will give different answers to questions about racial topics depending on whether the interviewer is of the same or a different ethnicity.

Related to the concept of equivalence across groups is equivalence across political institutions within a single country. Often scholars want to measure some attribute of actors in these institutions, but lack any common measures. The most noteworthy example involves ideology. Scholars of American politics want to know the ideological ideal point of various elected and appointed leaders: the President, Senators, House members, Supreme Court Justices, other federal judges. But differences across institutions mean that the available measures are quite different. The President does not vote on all of the bills that members of Congress vote on. Senators and representatives also vote on different bills. Judges and justices consider cases and not bills, but not always the same set of cases, given the geographical stratification of federal district and appellate courts. These challenges has led scholars to pursue different strategies—including the use of bridge observations, or actors within one institution who have served in another institution or taken positions on issues confronting another institution (Bailey 2007; Poole 1998)

A final source of cross-section non-equivalence is the individual—the source most relevant to the study of mass behavior. Individuals, like contexts, have distinct values, orientations, and reference points that may affect their responses to measurement instruments, such as surveys (see Brady 1985). Non-equivalence may also arise because of an individual's interaction with the survey interviewer. Those interactions vary considerably in how much rapport the interviewer and respondent develop, how well they communicate with each other, and thus in the quality and content of the response provided. Interviewers are, of course, trained to minimize idiosyncrasies by adhering to protocols. And yet rigid adherence to a protocol could itself create non-equivalence if additional information from the interviewer would enable the respondent to understand and answer the question in the manner intended. Thus, interviewers and therefore researchers face a situation that parallels Przeworski and Teune's (1966-67) distinction between identical and equivalent measures. Should interviewers do precisely the same things (i.e., behave "identically"), no matter what issues may arise in an interview? Or should they deviate occasionally from protocol to ensure the respondent's comprehension, just as researchers might use slightly different, but still "equivalent," measures to tap the same construct in two different countries? Suchman and Jordan (1990: 232) note that this tension is inherent in the interview, which both "relies on a wealth of conventions and resources from ordinary conversation" and yet involves constraints imposed by "standardized procedures." This notion of an "individualized" non-equivalence is probably the source least explored.

The Bonds-Ruth Problem: Non-equivalence across Time

Is Barry Bonds the best home run hitter in the history of Major League Baseball? Measured by the raw number of home runs, he is. But is it appropriate to compare Bonds' performance with that of players whose careers occurred years or even decades ago? Many things have changed in baseball since the era of Babe Ruth or even Hank Aaron: the size of ballparks, the quality of pitching, the composition of baseballs and bats, and the apparent prevalence of steroid use. In short, times have changed, and the question is whether they have changed so much as to render "number of career home runs" a misleading measure for

comparing the prowess of hitters over the history of baseball. This is the "Bonds-Ruth problem," with a nod towards the countless other intergenerational rivals who have featured in barroom arguments.[8]

To date, much research on equivalence is concerned with space rather than time—with Jordan and Woods, rather than Bonds and Ruth (exceptions include Bailey 2007 and Paxton 2009). But as data collection persists in the social science, the passage of time can also create non-equivalence as the meaning of constructs, and the items used to measure those constructs, changes. Such changes may stem from political events, shifts in cultural norms, and the like. As Clarke et al. (1999: 646) write, "Comparative politics specialists long have emphasized the difficulties that arise if one is not sensitive to cultural and linguistic differences that may bias survey data analysis. As argued here, economic contexts also are important, and these have dynamic components—they vary temporally as well as spatially." The problem can be no less acute in international relations; a concept like sovereignty is not temporally invariant (Reus-Smit 1997). As time passes, researchers may need to re-conceptualize the construct—e.g., "What does it now mean to be X?"—and reconsider whether particular items tap that construct.

Two examples will illustrate. The first concerns the concept of prejudice, in particular prejudice toward African-Americans. A variety of indicators suggest that prejudice has declined over time: fewer white Americans oppose interracial marriage, integrated schools, and other kinds of social interactions with blacks. Moreover, many fewer Americans believe that blacks are innately inferior, especially with regard to intelligence (see Schuman et al. 1997). But at the same time, other scholars have argued that these changes do not necessarily indicate a less prejudiced public. Instead, the nature of racism itself has changed to one that (and here we simplify) emphasizes lack of effort rather than lack of ability, a notion called symbolic racism (Sears 1988) or racial resentment (Kinder and Sanders 1996). This argument has met vigorous resistance (e.g., Sniderman and Tetlock 1986). Ultimately, this debate and the more basic question of whether contemporary

[8] With these athletic examples, we join a venerable tradition in the scholarship on conceptualization and measurement. For example, in Gallie's (1957: 170) discussion of "contested concepts," he writes: "We are all acquainted with the concept of 'championship' or of 'the champions' Commonly a team is judged or agreed to be 'the champions' at regular intervals, e.g., annually, in virtue of certain features of its performance against other contesting teams. Then for a certain period, e.g., a year, this team is by definition 'the champions' even though, as months go by, it becomes probably or certain that they will not repeat their success." Note that Gallie is also describing a kind of measurement non-equivalence that arises over time.

America is a less prejudiced America boils down to questions about equivalence. Is the prejudice of yesterday the same as the prejudice of today? Are new, more unobtrusive indicators needed to capture prejudice today? If so, how does that complicate comparisons over time?[9]

A second example involves the measurement of poverty. Is the rate of poverty in a particular country higher today than in years past? This proves a complicated question. Assume that at some previous time point, say 1960, poverty in some country was defined as an income of less than $1,000 per year, and 10% of the population fell below that threshold. What is the "equivalent" percentage in 2000—the percent with an income less than $1,000 per year? Probably not. For one, the threshold itself may need to be adjusted, if median income has changed. Second, the monies that are defined as "income" may need to be adjusted if certain social welfare transfers have been created or eliminated in the intervening year. Third, the tax burden may also have changed, such that more of people's income is being paid in taxes. All of these problems led to a new proposed standard for poverty in the United States in 1995 (see Betson and Warlick 1998).[10] In general, differing approaches to the measurement of poverty over time can lead to drastically different conclusions about the trend (see Jorgenson 1998; Triest 1998).

Both cases demonstrate that measurement equivalence may not only stop at the border, but also decay over time. The meaning of constructs change, as does the relevance of particular measures. Though the example of prejudice involves decay, this is not the only "functional form" that could describe the relationship between time and equivalence. Discrete events may also render certain measures more or less equivalent, perhaps only temporarily so. Much as ensuring cross-cultural equivalence demands attention to the idiosyncrasies of place, including language and culture, ensuring over-time equivalence demands attention to changes within a given geographic context and, in some circumstances, to changes that may occur within a set of contexts. The importance of "local" knowledge—well-known in the study of comparative politics—is

[9] The same set of issues confronts the study of political tolerance. Why has tolerance of such groups as atheists, communists, and socialists increased (Nunn, Crockett, and Williams (1978)? Because the public actually came to support extending liberties even to unpopular groups, or because the groups themselves became less unpopular? A way around this problem is to allow the respondents themselves to specify a group that they do not like (Sullivan, Pierson, and Marcus 1982). But this is not a solution if the strength of disaffection towards the "most disliked group" varies over time. See Mondak and Sanders (2003).

[10] Problems in the measurement of poverty also involve cross-sectional non-equivalence (i.e., the Woods-Jordan problem)—such as the need to take into account different standards of living in different parts of a country and for different kinds of family structures.

the same in the study of "politics over time." To quote Rustow's (1970: 347) paraphrase of Clemenceau, "history is far too important a topic to be left just to historians."

Approaches to Diagnosing Non-Equivalence

Within the literatures on cross-cultural measurement, interpersonal incomparability, and differential item functioning, there are a number of methods devoted to diagnosing non-equivalence. These literatures focus on methods within two data-analytic frameworks, structural equations modeling (SEM) and item response theory (IRT).[11] These approaches, and the specific methods within them, have largely developed in parallel literatures. IRT evolved within educational testing (Lord and Novick 1968), while SEM developed in psychometrics more generally. Despite their different origins, they are quite similar in many respects. (We describe and compare them in the on-line appendix, in the interest of clearing the underbrush.) Our conclusion is that the differences between the two approaches are largely terminological and notational, with one unit-of-measurement assumption separating the two. Given the data in question, the empirical analysis below draws on SEM. The chief liability of both methods is, however, an important one: both are employed when there are multiple indicators of the construct in question. When there are not, these techniques are less helpful. In the conclusion we address the challenge of constructing a single measure that is equivalent across contexts.

Whether one is working in an IRT or the SEM framework, typical tests for equivalence follow the same logic. A researcher conducts a hierarchical series of tests that probe for finer and finer degrees of incomparability (Bollen 1989; Cheung and Rensvold 2000). We describe a series of four specific tests based on the forms of equivalence we describe above. First, does the same measurement model apply equally well in different contexts (structural or factor form equivalence)? This test entails pooling the different contexts and estimating an "unconstrained" model that allows all of the model's parameters to vary across contexts—with the exception of the indicator whose loading is scaled to 1 for the purposes of identification, and one of the intercepts, which must also be constrained equal across groups (see Bollen 1989: 307). If the fit of this

[11] SEM is sometimes referred to as "covariance structure modeling" or as "mean and covariance structure modeling" (MACS).

model is adequate, then structural equivalence is assumed and this model becomes the baseline against which all constrained models are evaluated. If this condition does not hold, then one need go no further, as subsequent tests all assume an invariant structural form.

Second, does the "weight" of different indicators of a particular construct vary across contexts (factorial or metric equivalence)? This question can be tested first with an overall or "global" test of metric equivalence, constraining all loadings to be equal across contexts, again with a single indicator's loading set to 1 for the purposes of scaling.[12] However, tests of metric invariance encounter a "standardization problem" (Rensvold and Cheung 2001). In short, selecting a scaling indicator assumes that this indicator is itself equivalent. If this assumption is not true, then there is a much higher probability of a Type I error, or rejecting the null of equivalence when it is in fact true (Stark, Chernyshenko, and Drasgow 2006: 1304). Rensvold and Cheung (2001) propose a sequence of models that vary the combination of scaling indicator (or "referent") and an indicator (the "argument") whose loading is constrained across contexts.[13] If metric equivalence cannot be established, the choices will depend on the number and nature of available indicators. Researchers could pursue a model that is "partially" equivalent, constraining some factor loadings to be equal across contexts and allowing others to vary (see Byrne, Shavelson, and Muthén 1989). Or researchers may instead choose to eliminate certain (non-equivalent) indicators. Ultimately, there is no one default "solution." We return to this subject below.

If metric equivalence has been established, then one may go on to evaluate scalar equivalence. That is, do different contexts share a common origin on each indicator, i.e., the value of the indicator when the latent construct is 0? Evaluating scalar equivalence entails a similar test constraining the intercepts to be equal across contexts. In many examples, the series of tests stops here, although if a researcher has particular

---

[12] In models with more than one factor, it may also be instructive to conduct subsequent tests of global metric equivalence at the level of each individual factor (Rensvold and Cheung 2001: 36).

[13] More specifically, a model is estimated for each possible pairing of referent and argument—e.g., in a three-indicator model, indicators 1 and 2, 2 and 3, and 1 and 3. If any model's fit is worse than the unconstrained model, then that pair of indicators is flagged. After all pairs have been examined, the set of equivalent items is identified by eliminating any feasible set that includes a flagged pair. So, if 1-2 is the flagged pair, then the only feasible set of indicators is 2-3, since 1-2 and 1-2-3 are eliminated. If no such pairs are flagged, then metric equivalence has been established. In general, in an N-indicator model, there are $N(N-1)/2$ pairs of indicators. The number of feasible subsets of items is a more complicated formula, but the upshot is that the number of such subsets grows much larger with additional indicators. See Rensvold and Cheung (2001) for strategies on dealing with models that have many indicators.

interest in measurement error variance, or any other parameter of the model, he or she may desire further tests (e.g., of the equivalence of measurement error variance).

In estimating this series of models, each constrained model is compared again the fully unconstrained model according to various measures of fit. A standard measure in SEM is the chi-squared statistic. The statistic tests the null hypothesis that the sample covariance matrix is equal to the covariance matrix implied by the model parameters. Thus, a significant chi-squared statistic indicates that there is not a close correspondence between the data and the model's predictions—i.e., there is not a close fit. To test for equivalence, one can evaluate the change in the chi-squared values across nested models to see if imposing constraints worsens the fit. The difference in two chi-squared statistics also follows the chi-squared distribution, with degrees of freedom equal to the difference in the two models' degrees of freedom. A well-known problem with the chi-squared statistic, and with the difference between two chi-squared statistics, is that they are sensitive to sample size. Large samples are more likely to produce significant chi-squared tests. However, as noted by Rensvold and Cheung (2001), there is no other statistic where the difference in that statistic between two models follows a known distribution. Nevertheless, it is important to draw upon multiple indicators of fit, if only to compare visually any differences across models. Below we draw on three other measures: the Tucker-Lewis Index (TLI), the Comparative Fit Index (CFI), and the root mean squared error (RMSEA). For the TLI and CFI, values close to 1 indicate a good fit. For the RMSEA, values close to 0 are a good fit (see Browne and Cudeck 1993).

Three Assessments of Equivalence: Political Involvement, Democracy, and Development

Having outlined a conceptual framework for evaluating equivalence and reviewed the methods appropriate to such tests, we now move to data and constructs in use today in political science. To what degree is non-equivalence manifest in measures of these constructs? How reliably do the methods described above diagnose any non-equivalence? Our approach is to identify, ex ante, methodological or conceptual differences that seem likely to produce non-equivalence, and then to identify particular items that may exhibit non-equivalence. We then evaluate these items using the methods described above. We seek to accomplish

three tasks: (1) identifying items that "function" differently across different groups or countries; (2) understanding the reasons why these items function differently; and (3) determining the practical implications of including and removing these items. We test the effects of sources of non-equivalence across three important constructs, one measured at the individual level (political involvement) and two at the country level (democracy and development). Within all three examples, we develop theoretical expectations of non-equivalence along both spatial and temporal dimensions, and thus identify the potential for both Woods-Jordan problems and Bonds-Ruth problems. In the case of democracy and development, we assess the implications of non-equivalence on causal inference.

Political Involvement

We start with the concept of political involvement, long a concept central to cross-national and cross-time studies of political behavior (e.g., Almond and Verba 1963; Jennings and Markus 1988). We conceive of political involvement as the extent to which an individual is engaged actively and/or psychologically in politics. Using the World Values Survey, we measure the construct with the following six items, all of which are relatively common indicators:

- Politics important. "For each of the following, indicate how important it is in your life…[politics]"
- Political interest. "How interested would you say you are in politics?"
- Discuss politics. "When you get together with friends, would you say that you discuss political matters occasionally, frequently, or never?"
- Follow News. "How often do you follow politics in the news on television, or on the radio, or in the daily papers?"
- Sign Petition. "I'm going to read some different forms of political action that people can take and I'd like for you to tell me, for each one, whether you have actually done any of these things, whether you might do it, or if you would never, under any circumstances, do it…[signing a petition]."
- Belong to Party. "Look carefully at the following list of organizations and say which if any you belong to…[political parties or groups]."

While these items appear to be fairly general items with reasonable face validity, we test two expectations about why they may exhibit non-equivalence.[14] The first concerns attention to political matters

[14] Although we focus on two sources of non-equivalence, we harbor suspicions about others. Consider the two activity items: signing petitions and party membership. These activities could easily vary in relevance and meaning. We know, for example, that party membership ranges from the card-carrying variety in some places to a psychological identification in others. In many contexts, such as the United States, it is not clear what membership entails, exactly. What did the

in the news. The frequency, content, and intensity of political news are highly time-dependent, often rising and falling according to the calendar of political campaigns and elections. Because cross-national surveys are usually administered in the same rough time period, in some places the survey will be conducted during the heat of a campaign and in others during periods of relative calm. For example, interviewing for the WVS in Uganda began and ended a week before and after the country's presidential election of March 12, 2001, whereas the Nigerian fieldwork was carried out almost two years after and three years before the closest national elections. At the height of the political season, the politically attentive public could be a large majority of the public, including the politically involved and otherwise, whereas the same group in the "off season" will likely be limited to political junkies. As such, the "follow news" item might load more weakly on the latent construct in-season than it would off-season; effectively, at the height of the political season the item will not discriminate as effectively political junkies from the politically apathetic. Not only will the slopes, and thus the predictive ability of this item, differ across political seasons, but so will the intercepts, which should be higher in countries surveyed during or soon before or after a campaign. Elections likely stimulate attention to the news, thereby creating an apparent mean-level difference between two people surveyed in and out of season, whose actual level of (latent) political involvement is the same.

Besides the political calendar, there is another potential source of non-equivalence: a translation discrepancy with potentially large distorting effects. In investigating the local-language version of the WVS country instruments, we found that in four of the Spanish-speaking cases (Mexico, Chile, Venezuela, and Colombia), the word "discuss" was translated using the cognate "discutir," which in Spanish takes the meaning of "argue" as opposed to "converse." In the other five cases (Argentina, Spain, Peru, El Salvador, and the Dominican Republic), the translator used "hablar," which is closer to the meaning in the source question.[15] We suggest two hypotheses about how this translation discrepancy might affect equivalence.

---

roughly 20% of the US sample mean when they claimed that they "belonged" to a political party? Petitioning likely suffers from a similar problem, varying in prevalence not only because of citizens' political involvement but also because of local custom. Taken this way, these activities seem highly context-specific. Nonetheless, the other possible items in the WVS that we might have included in the measure—letter-writing, boycotts, demonstrations, etc.—are equally if not more idiosyncratic.

[15] Incidentally, the Portuguese-language questionnaires contain a similar discrepancy: the Brazilian questionnaire uses "discutir" (whose meaning is similar to the Spanish), while the Portuguese questionnaire correctly uses "conversar."

First, the relationship between the "discutir" item and the underlying latent attribute may be weaker than the equivalent relationship involving the "hablar" item. Second, the intercept for the discutir item should be lower than that for the hablar item, as people of equal levels of political involvement would be less likely to report "arguing" about politics than "discussing" politics.[16]

Hablar/Discutir. Using the in-country versions of the survey instrument, we divide the Spanish-speaking cases into the two groups as identified above. Since three of the nine countries did not ask all of the political involvement items, we are left with Argentina, Peru, and Spain in the "hablar" group, and Mexico, Venezuela, and Chile in the "discutir" group.

We begin with an unconstrained model, in which all parameters are allowed to vary across the two groups. Following standard practice in factor analysis, we set the loading and intercept of one indicator (the importance of politics) to 1 and 0, respectively. Figure 1 presents the unstandardized factor loading and intercepts from this model, along with 95% confidence intervals. Fit statistics for the model, presented at the bottom of the figure, suggest that the fit is adequate and thus that the model is structurally equivalent across the two groups. The figure also suggests whether the translation discrepancy induced a significantly different loading or intercept. The top panel of Figure 1 shows that the loading for the discussion item did not differ significantly across the two groups. In fact, with one exception (watching the news), there are no evident differences in the loadings. When we estimate a subsequent model constraining all of the factor loadings to be equal (not shown), the model fit, as measured by the CFI, TLI, and RMSEA is virtually unchanged.[17] Clearly, any distortions in translation were not enough to disturb metric equivalence. It is not clear why the loading for the news item would not be equivalent across these groups, unless the hablar/discutir grouping is correlated with some other difference in measurement, something we can investigate more thoroughly in the subsequent set of analyses.

---

[16] Of course, this is not the only potential problem with the political discussion item. For one, the response categories are unanchored. Better categories would precisely describe frequency—e.g., "1 to 2 times a week, once a month, etc."

[17] When we performed the iterative indicator-by-indicator test, we found that only one indicator exhibits any non-equivalence, the "follow news" item. The loading for the "discuss politics" item actually appears to be equivalent across the two groups for which we expected to see differences.

[insert Figure 1 about here]

The bottom panel of Figure 1 presents the intercepts for the two groups. There are several statistically significant differences across the groups. As we hypothesized, the intercept for "discuss politics" is lower in the "discutir" countries than in the "hablar" countries. The mistranslation may have depressed affirmative responses to that particular question, resulting in scalar non-equivalence. Moreover, several other indicators—interest in politics, watch the news, and sign a petition—unexpectedly manifest this same pattern. When we conduct a global test of intercept invariance, by constraining all the intercepts to be equivalent across the groups, the model's fit worsens somewhat (CFI=.934; TLI=.930; RMSEA=.067), although not to alarming levels.

Overall, this analysis gives us little reason to be concerned with the usage of discutir and hablar. The overall model form and the loading of discuss politics item is invariant. The analysis of the intercepts does suggest that this item exhibits modest scalar non-equivalence, which may or may not be a result of the discutir/hablar distinction, since the intercepts of several other items also differed across the two groups. The prevalence of scalar non-equivalence across these items necessitates further investigation into its implications for the estimation of country means and causal inference.

Proximate vs. Distant Elections. Election timing may affect any of the six indicators. We calculated the proximity (in days) between the closest national election (either presidential, legislative, or referendum) and both the beginning and end of fieldwork. We classify as "election surveys" those in which at least one day of interviewing was within 60 days of a national election, and "non-election surveys" those for which no day of interviewing was within 365 days of such an election. There were 21 non-election surveys and 15 election surveys across the 62 cases for which we had reliable fieldwork and election dates.[18]

Figure 2 reports the standardized loadings and fit indices from an unrestricted model. The fit indices suggest a good fit and therefore structural invariance. The loadings between the two groups are statistically significantly different in three cases: interest in politics, importance of politics, and party membership. In

[18] Data on the timing of elections and the WVS country surveys are available from the authors.

each case, the loading in countries with proximate elections is higher, contrary to our hypothesis, but the substantive differences are not large. Constraining the loadings to be equal across groups has little effect on the model's overall fit; in fact, two fit indices actually show improved fit, suggesting that differentiating these groups by season makes little sense.[19]

Similarly, the intercepts themselves manifest some statistically significant differences across these groups. Two cases, watching the news and signing a petition, confirm our hypothesis: the intercept value is larger in countries with proximate elections. In two cases, the importance of politics and party membership, the difference between the groups fails to support our hypothesis. But again, any differences are substantively modest. A global test of scalar equivalence, constraining each item's intercepts to be equal across groups, produces negligible changes in fit.

Both of these sets of potential contextual differences – the translation issue that substantially intensified the meaning of the question and the seasonal difference – seemed likely to be consequential. It is somewhat reassuring for analysts, therefore, that the measurement models were robust to these variations. Nonetheless, more serious problems of research design can yield striking problems of non-equivalence. In the analysis (not shown) of other constructs from survey research (e.g., internationalism, tolerance, and anti-immigrant attitudes), we have found problems as serious as structural non-equivalence. In the case of internationalism, for example, we sought to measure an individuals' attitude towards other countries and found two indicators that loaded strongly in highly developed countries, but not at all in less developed countries. Keeping in mind these possibilities, we now explore sources of non-equivalence in the context of an important concept measured at the country level: democracy.

Democracy

In investigating the equivalence of democracy measures, we turn first to a potential manifestation of the Bonds-Ruth problem: whether measures of democracy are equivalent across time and, in particular, across "waves." Many scholars take an explicitly historical view of democracy, sometimes attempting to periodize

[19] Imposing constraints on individual pairs of items suggests that most of the loadings are equivalent. The fit indices are quite stable across this series of models.

the variation in democracy across 200 years (e.g., Huntington 1991). It is even common to see references to ancient Athens in discussions of democracy's origins. Certainly, however, the structure of states, their institutions, and political norms have changed since 1800, not to mention 700 BCE. Thus, measures of democracy that presuppose certain institutional arrangements and practices (e.g., the modes of political participation) may not travel well across time.

Another problem concerns reference points: should we measure democracy against a single standard or against the standard at the time, e.g., by constructing context-adjusted measures (Adcock and Collier 2001)? The issue is analogous to inflation in measures based on currency, like GDP per capita. For example, the United States has scored the maximum 10 on Polity's (Marshall, Jaggers, and Gurr 2004) democracy score since the Civil War. For much of this period, however, significant portions of the population (women and, at least indirectly, blacks) were excluded from political participation. Assuming a single (non-contextualized) standard, the Polity participation scores before women and blacks earned the (*de jure* and *de facto*) right to vote must be inflated and, it seems, incomparable to contemporary scores in an absolute sense (Johnson 1999). Tatu Vanhanen's (2000) measures of democracy, on the other hand, proceed from the opposite assumption. Vanhanen proposes two objective indicators—turnout and the winning party's margin of victory—that ostensibly tap participation and competition, two central dimensions of democracy. Turnout (at least with the total population as the denominator) has undoubtedly increased over the years in most countries. In Vanhanen's sample, turnout numbers ranged from 0 to 20 percent (with a mean of less than 3) in 1875, but from 0 to 70 in 2000 (with a mean of 33). Vanhanen's measure of participation does not adjust for inflation, but rather assumes a single standard of democracy and not a contextualized, era-specific, standard. Either strategy is defensible, depending upon whether one's research design calls for a relative or absolute measure of democracy, but in either case analysts of these data should be conscious of the difference. Research designs that call for a contextualized measure of democracy will be subject to measurement non-equivalence with Vanhanen's measure and vice versa with the Polity measure.

In this case, the issue is probably *not* one of structural non-equivalence. The basic components of each of the two scales—participation, competition, and real constraints on the executive (Polity only)—

arguably form a core set of indicators that are relevant across the last two hundred years. Of course, one could argue that this set of measures under-represents the concept of democracy to a certain extent. For example, outside of political participation, there is no attention in either scale to political and civil rights, something many see as a critical dimension of democracy (e.g., Diamond 1999). More likely, however, the problem concerns the slopes and/or the intercepts of at least one of the items in the measurement model. Consider Bonds and Ruth again, briefly. If modern hitters hit more home runs than they did fifty years ago, it could be that: (a) hitters are better than they were then; or (b) it is easier to hit home runs these days (drugs, smaller parks, more lively baseballs, etc. have inflated the rate of home runs), or (c) hitters are just as good they used to be and home runs are just as hard to hit, but home runs are less relevant to being a good hitter than they used to be (e.g., home runs now correlate less highly with batting average and other indicators of good hitting—the rise of the one-dimensional slugger). These possibilities correspond to (a) real differences in the latent variable (good hitting); (b) scalar nonequivalence; and (c) metric nonequivalence. Substitute turnout for home runs and the challenges of comparing democracy between 1875 and 2000 are equivalent to those of judging good hitting across eras.

We can evaluate these forms of non-equivalence empirically. We start by building a measurement model with a set of indicators from both Polity and Vanhanen, each of which have continuous coverage across all three waves. Other data sources have periodic coverage that crosses at least two waves (e.g., Bollen 1990; Alvarez, Cheibub, Limongi, and Przeworski 1996). Using the components from the Polity score (political competition, executive constraints, and executive recruitment) as well as the two indicators that compose the Vanhanen scale (participation and competition), we can construct a measurement model that ranges from 1816 to 2000. Figure 3 plots the factor loadings for each indicator in the model, when estimated yearly. These are unstandardized loadings, with each indicator re-scaled to range between 0 and 1, and the question is whether we observe any vertical movement in the plot over time. In order to identify the model and scale the magnitude of the loadings, we constrain the loading for the Polity indicator, executive recruitment, to 1 in each year (it is not presented in Figure 3).

[insert Figure 3 about here]

In general, the loadings for each of the indicators increase over the years, suggesting an increasing association with the latent variable. In all years, each of the indicators is at least moderately associated with the latent variable, indicating that the model is not subject to the most severe form of non-equivalence (structural). Some periods—in particular the crisis years following the revolutions of 1848 and World Wars I and II—exhibit levels of low reliability for several of the indicators, turnout in particular. Overall, however, the results suggest that while some yearly comparisons (e.g., a comparison of 19th century cases with contemporary ones) will strain the assumption of measurement equivalence, a comparison of points within the modern period (and specifically the often-sampled post-WWII years) do not. As we show below, some indicators of economic development have much more variable loadings than what we observe here.

Consider now the issue of scalar equivalence in the context of possible inflation, particularly in the measure of turnout. The question is whether the score on any given indicator is the same across eras when the value of the latent variable democracy is zero. Figure 3 plots the intercepts for the same measurement model across time. As we may have expected, the intercept for turnout increases steadily and dramatically through the years. A score of zero on the latent score of democracy corresponds to .03 on the turnout measure in 1900, .14 in 1920, and .29 on the measure in 2000. If we are curious about how much the latent construct democracy affects turnout, we must shift the our expectations accordingly across years, in the same way we might use the consumer price index to adjust for inflation. These estimates suggest that highly undemocratic countries in 2000 average 29 percent turnout, versus 3 percent in 1900, implying an average yearly inflation rate of 8.6 percent [((29-3)/3*100)/100].

A second concern with measures of democracy is whether they are equivalent across different geographical contexts even within the same wave—a manifestation of the Woods-Jordan problem. A small scholarly tempest erupted after the breakdown of the Soviet Union and the transitions to democracy among former communist countries. Scholars who had honed their theories and measures of regime type in the Americas and Southern Europe (call this the "south"), where the early stages of the third wave of democratization had occurred, were eager to cut their teeth on new cases in the post-communist world (the

"east"). Some scholars of the east (Bunce 1995) protested that "transitologists" employed models and measurement tools that were ill-equipped to assess political change in post-Soviet countries.

There are multiple strands of argument in this debate, some of which are not relevant to us here.[20] We focus on the conceptualization and measurement claims within Bunce's argument. One claim is that the concept of democracy, however measured, inadequately captured the large-scale changes that countries in the region were undergoing. For Bunce, to call the post-Soviet outcome "democracy," missed the deeper, more fundamental changes in society, markets, and the organization of the state. Whereas, she allowed, the democracy-authoritarianism dimension may have been the focal dimension that structured decision-making in the south, "what is at stake in eastern Europe is nothing less than the creation of the very building blocks of the social order" (118). Bunce's claim evokes the notion of construct non-equivalence, since her conclusion is that the concept, as developed in the southern context, cannot be fruitfully applied to the east. She writes, "The key question, then, is whether the differences constitute variations in a common causal process—that is, transitions from dictatorship to democracy—or altogether different processes—democratization versus what could be termed post-communism" (119). As we describe earlier, these sorts of conceptual decisions—in particular, whether the concepts of democracy and democratization are relevant to a particular context—are more appropriately judged before proceeding to measurement. Nonetheless, it may be illuminating to evaluate borderline cases of construct non-equivalence by consulting the data, since the existence of construct non-equivalence will likely imply non-equivalence in the lower order measurement parameters (such as the loadings, intercepts, and error variances). Such an exploration makes sense here, since Bunce is skeptical of, but not wholly opposed to, measuring democracy in the east.

In fact, Bunce's claims extend beyond the conceptual level. She also charges that cross-national measures of democracy manifest nonequivalence of such a degree that transitologists have mistaken authoritarian cases for democratic ones. Her argument, to put it in our terms, is that the set of instruments used to measure democracy in the south under-represents the concept of democracy as manifested in the east (i.e., a case of structural non-equivalence). She points to indicators that are especially relevant to the post-

[20] In addition to the conceptualization and measurement issues we summarize here, much of the dissent had to do with the incomparability of the background conditions and causal logic of processes of transitions in the two contexts.

communist setting. For example, she suggests that analysts take into account the presence of members of the ancién regime, asking: "if the communists—now ex-communists—continue to occupy important posts in eastern Europe and if the media in most of these countries is still subject to undue control by the government in office, then is it accurate to argue, as Schmitter and Karl do, that these regimes have moved from the transition period to a period of democratic consolidation?" (113) Her claim implies that the architecture of democracy measures varies between regions (i.e., structural non-equivalence). One possibility, then, is that at least some of the standard democracy items—participation, constraints on the executive, and competition—will be irrelevant to the latent concept of democracy in cases from the east.

[insert Figure 4 here]

We test these expectations by building a measurement model with seven indicators of democracy and testing its parameters in the "East" and the "South." In addition to the five measures (three from Polity and two from Vanhanen) that we describe above, we add an overall measure of political and civil rights from Freedom House, and a dichotomous measure of democracy constructed by Przeworski, Alvarez, Cheibub, and Limongi (1996). Figure 4 plots the unstandardized factor loadings and intercepts, respectively, for separate models for the South and East in the year 1995, again with the loading for executive recruitment constrained to one to identify the model. The loadings appear to be roughly equivalent with the exception of Vanhanen's participation measure (essentially, voter turnout). Such an objective measure appears less informative about the strength of democracy in post-Soviet cases and, in part, corroborates Bunce's unease regarding outsiders' understanding of democracy in these cases. Figure 4b suggests that some real differences in the intercept for turnout across regions as well, suggesting that adjustments for inflation may also make sense for comparisons across region. Nevertheless, a reliance on turnout data was probably not responsible for any errors that transitologists may have made in scoring democracy in the east, as they would likely have been focused on indicators other than turnout. Moreover, the other six indicators behave very similarly in the two contexts, which suggests that Bunce's warnings (at least with respect to structural equivalence) may be overstated. These analyses, of course, do not shed any direct light on Bunce's more conceptual point regarding differences in the salience of democracy in the two contexts. We can say that the architecture of

democracy, and the relative weights of individual indicators within that architecture, are reasonably equivalent across south and east (with one significant exception). Indirectly, that finding can have conceptual implications. A finding of structural or metric non-equivalence—although meant to probe for finer irregularities—can indicate deeper conceptual problems of comparability.

Economic Development

The preceding examples suggest reasons for both concern and reassurance, but without any real sense of what is at stake with respect to causal inference. We now turn to the concept of economic development not only to explore measurement equivalence but also to assess the implications of nonequivalence for causal inference.

The relationship between democracy and economic development is undoubtedly one of the most central and enduring subjects of inquiry in comparative politics. Theory supporting a positive association between the concepts goes back at least as far as Aristotle. Seymour Martin Lipset, in his seminal 1959 article, moved the relationship from conjecture to an accepted empirical regularity.[21] Analyzing a sample of cases from Europe and Latin America in the late 1950's, Lipset demonstrated a strong relationship between democracy and each of a set of indicators of development. Since Lipset, many scholars have explored the relationship between democracy and development in order to better understand the causal mechanisms at work (e.g., Acemoglu and Robinson 2006; Przeworski et al. 2000; Londregan and Poole 1996; Boix 2003). Many of these studies pool data across countries and across time in ways that assume the equivalence of measures.

We test whether the simple relationship that Lipset documented in 1959 is evident fifty years later—a period in which the number of independent states has doubled, dictatorships and democracies have come and gone, and significant technological and geopolitical changes have altered what economic success looks like. This sort of comparison relies upon comparable measures of both democracy and development across time.

[21] As one indicator, Lipset's article is the seventh most cited article in the history of the *American Political Science Review* (Sigelman 2006).

As we shall see, an analysis of democracy and development since 1945 illuminates the effects of measurement nonequivalence, mostly due to nonequivalence in various measures of development.

Lipset's original analysis serves as a useful benchmark given its widespread impact and, more notably, his use of multiple measures of development. We begin with Lipset's conceptualization and measurement of democracy. Lipset conceived of democracy in a minimal Schumpeterian sense, defining democracies as those regimes that fill important offices via elections. His sample was limited mostly to Europe and Latin America and he categorized states therein and circa 1959 as either "stable democracies" or "unstable democracies and dictatorships." In his analysis, he often listed the European and Latin American variants separately, perhaps implying graded classes of democracy. Nevertheless, in other passages of the article, he clearly means for regime type, not geography, to be the critical distinction among cases.[22] Lipset's institutional conception of democracy corresponds to later time-series measures (notably, Polity, Przeworski et al., and Vanhanen) and, indeed, his classification correlates reasonably well with each of these measures during Lipset's sample time period. Lipset's measure loads quite strongly (with a standardized loading of 0.61) on the latent construct in a single-factor model that includes measures of democracy from the three sources listed above, each averaged over the ten years between 1950 and 1959 (a time horizon that presumably approximates Lipset's). Since we will evaluate Lipset's hypothesis over time, it is important to know that we can substitute one or more of these measures for his own.

Lipset approached the concept of development with a set of multiple measures, each of which he compared with his measure of democracy. He conceptualized development in terms of four, presumably highly correlated dimensions: wealth, urbanization, education, and industrialization. For each dimension, he identified between two and six indicators, and found that each one correlated highly with his measure of democracy. Assembling these fifteen indicators of development, we replicate Lipset's analysis and found results that effectively match his. Lipset's democracies and non-democracies are different from one another

[22] While the results could be read as supporting categorical differences between the European and Latin American cases, Lipset (1959: 75) states that "if we had combined Latin America and Europe in one table the differences would have been greater."

in the expected direction across the fifteen indicators.[23] Since we will need to substitute a time-varying measure of democracy for Lipset's static measure, we then compared the association between the available democracy measures and the fifteen indicators of development. The three other democracy measures exhibit the same strong relationship to the development indicators as does Lipset's. The correlation between the Polity measure and the fifteen indicators in 1959 hovers around 0.50 for most indicators. The world as Lipset saw it in 1959 looks the same to us today using updated historical measures of democracy. How does the world in 2000 compare, by these same measures?

We have already demonstrated that the democracy measures are reasonably equivalent in the post-WWII era. Recall that the concern with the democracy measures lies mostly with the comparison between the late 19th century and the modern era. We do not have the same confidence in the comparability of the development indicators. It seems probable, for example, that the prevalence of radios and primary school enrollment—neither of which is now as much a privilege of the affluent—are no longer markers of societal wealth. Indeed, if we consider the countries with the highest primary school enrollment in 1959 and 2000, the list looks quite different. Mostly established states top the list in 1959, while less likely suspects such as Libya, Malawi, and Belize do so in 2000. Primary enrollment per capita seems to indicate something very different from what it indicated in 1959. This same is true of urbanization. After a large-scale migration to cities over the last fifty years, many Latin American countries such as Brazil and Mexico are as urbanized as the United States, but few would put the countries in the same level of development. By contrast, gross domestic product, per capita would appear to be fairly comparable across time (assuming that one accounts adequately for inflation and differences in exchange rates).[24] In short, there is a reasonable worry about metric non-equivalence in some indicators but not others. Do we see shifts in their association with the concept over time, and if so, how do these shifts affect inter-temporal estimates of the relationship between democracy and development?

---

[23] The one exception concerned the industrialization indicator, energy consumption per capita, which we found to be higher in Latin American dictatorships than in Latin American democracies. The difference appears to stem from high values on this indicator in our data for Venezuela (a "dictatorship"). When we exclude Venezuela, our results match Lipset's on this indicator as well.

[24] GDP may not be comparable across states (i.e., because of Woods-Jordan problems), but we do not test those here.

We construct a condensed measurement model that includes four of Lipset's key indicators, one for each of his dimensions of development. The indicators (and the relevant dimension) are: GDP per capita (wealth), percent of the population living in cities over 100 thousand (urbanization), primary school enrollment (education), and energy consumption per capita (industrialization). These four indicators constitute an abridged version of Lipset's measurement strategy and, given the deep historical coverage for each indicator, the four make for a reasonable extension of Lipset's model over time. We note, of course, that a model that comprises four instead of fifteen indicators will be more sensitive to any validity and reliability problems attributable to a particular item. Figure 5 plots the unstandardized loadings for each of the four indicators in a one-factor model, with each indicator scaled to range between 0 and 1. To identify the model, the variance of the latent variable is scaled to 1.

[insert Figure 5 about here]

Through 1960 the items behave as Lipset theorized. The standardized loadings (not shown) are all over .60 in that period, suggesting a reasonably high correlation between each indicator and the latent construct. In this period, GDP per capita and energy consumption per capita load exceedingly strongly, with standardized loadings close to 1.0, while urbanization and primary school enrollment are more moderately associated with the construct. Over time, however, we observe a steep decline in the reliability of primary school enrollment and a sizable, but less dramatic, decline in that of urbanization. In Figure 5, we present the unstandardized estimates over time (unstandardized, in order to preserve any differences in the variances of the indicators). Most strinking is the finding that, after 1980, primary school enrollment is actually *negatively* correlated with the latent construct and, by 1990, significantly so. By contrast, GDP per capita, urbanization, and energy consumption per capita appear to be reasonably comparable across time. We note a sharp drop in the loadings for energy consumption in 1970, which upon further investigation, marks a change in the sources used by Correlates of War (COW) researchers to calculate these values.[25] Of course, these sorts of sourcing inconsistencies in cross-national historical data can potentially have significant effects on

[25] COW researchers report that they use UN data starting in 1970 and, before that, data from Mitchell's historical volumes (Correlates of War, National Material Capabilities Documentation V. 3.0: p. 42).

measurement properties, but here the discontinuity in the energy loading —unlike that of primary school enrollment – suggests differences only in degree, not direction.

What is the effect of including in a measurement model an item that varies so dramatically in its connection to the latent construct across time? Recall the motivating research question, which was to assess differences in the association between democracy and development between Lipset's day and ours. Assume then that a researcher constructs an index of development with the four items in question. On its face, this would be a reasonable strategy to pursue. The four indicators exhibit content validity, they are key indicators in Lipset's benchmark measurement model, and they are available across time for the period under evaluation. Accordingly, we naively construct a simple additive scale with these four indicators and, for each year, regress the Polity measure of democracy on that index. Figure 5b, which plots the regression coefficients over time from this equation, suggests that the relationship between democracy and development has changed markedly since Lipset's assessment in 1960. Indeed, the relationship now appears to have reversed starting in the late 1970's, with democracy and development negatively correlated thereafter. This is a startling finding.

But consider an index of development that *excludes* primary school enrollment but retains the other three. Figure 5b plots the coefficients from a regression of Polity on that measure and tells a very different story. In those comparisons, the relationship between democracy and development appears to be alive and well, albeit with a noticeable drop in magnitude in the early 1990's following regime transitions in Eastern Europe. Together, these two figures suggest that metric non-equivalence, at least in the rather acute form afflicting the primary school indicator, can turn causal inferences on their head. It does so in the context of an extremely important question in political science and after following only a moderately naïve measurement strategy. Admittedly, no comparativists to our knowledge hang their hat on the comparability of urbanization rates or primary school enrollment. However, it is not preposterous to think that they would and, certainly, it is quite common to see analysis that relies upon single indicators of concepts that are as suspiciously incomparable across time. The point is that violations of equivalence can have serious effects.

Discussion and Conclusion

Equivalent measurement is imperative for comparative and historical research. Our goal was to familiarize scholars with the concept of equivalence and to suggest how they might engage it in their research. We have delineated various forms and potential sources of non-equivalence, as well as methods for diagnosing it. Our empirical analyses suggest that indicators of important constructs do exhibit non-equivalence. Indicators of democracy and development across time seem particularly vulnerable. In particular, we found that comparisons of democracy across centuries and across periods with varying international stability present potentially serious, if not fatal, issues of metric and scalar non-equivalence. For example, in 1875 democracy explains approximately 3 and 6 percent of the variation in measures of turnout and political competition (Polity), respectively, but roughly 44 and 81 percent, in 2000. We also found shifts across time in the intercepts of some indicators—indicating, in the case of turnout, massive inflation. The same value of "democracy" translates into very different levels of turnout across eras. We then tested the consequences of non-equivalence in the context of economic development, for which we had identified a more acute case of metric non-equivalence. There we found that one indicator, primary school enrollment, became a poor measure of development over time, and that a scale including this measure showed a declining relationship with democracy over time. Without attention to nonequivalence, this finding would overturn a canonical empirical relationship in comparative politics.

The potential for such serious consequences make it imperative that researchers address nonequivalence. How can they do so, even as they grapple with the other demands of research design and data analysis? To adopt medical patois, we might think of the answer in terms of *prevention*, *diagnosis*, and *treatment*.

Prevention

In an ideal world, of course, prevention is preferable and, as the saying goes, worth a pound of cure. This option is available to researchers who are designing research. A key decision is to define the domain

under study—that is, the particular contexts, whether countries, time periods, or some other context. After evaluating the appropriateness of the contexts, researchers can mitigate any future equivalence problems by delimiting the domain of study to those contextual units where equivalence holds. As Adcock and Collier (2001: 535) write, "scholars may need to make context-sensitive choices regarding the parts of the broader policy, economy, or society to which they will apply their concept." To be sure, this strategy may not be optimal for those who want to generalize as widely as possible. But clearly the gains from generalization are chimerical if the comparisons are invalid. Simply because data have been gathered across a wide array of contexts does not mean that researchers must analyze all of these contexts at once.

Prevention can also involve the construction of the measurement instrument itself. As much as possible, researchers should strive to multiply measures of key constructs. The statistical techniques for evaluating equivalence typically demand multiple measures. Equally important is this mundane reality of measurement: scholars often do not know whether measures will "work" until they are employed. Betting the farm on a single measure is thus a risky strategy. To be sure, pilot studies can help refine new measures before they are put into the field, but if the contexts under study are numerous (as in a multi-country survey) then extensive pilot studies may not be practical. Ultimately, with more measures in hand, researchers can be more confident that at least some of them will prove equivalent across contexts.

A second element of instrument construction is who does the constructing. Where relevant, researchers should multiply the individuals charged with constructing the measure(s)—that is, the judges, coders, etc. Having multiple coders is standard practice in some domains—such as content analysis of texts—but it is beneficial in many others. Multiple coders not only allow researchers to evaluate their measure via standard diagnostics (e.g., intercoder reliability), but also enable researchers to evaluate whether there are measurement artifacts associated with particular coders. Structural models that include parameters for these artifacts can be estimated in order to "cleanse" the resulting measures.

A third element of instrument construction involves characteristics of the measures themselves. What kinds of measures can be developed that will themselves mitigate the potential for non-equivalence? When measures involve self-reports of some kind, as in survey instruments, establishing a common reference

point is crucial (Heine et al. 2002). It is particularly crucial when researchers want to compare levels of some attribute across contexts—as researchers do when they present country-level means from a cross-national survey, or intercepts from country-specific models (see, e.g., Jusko and Shively 2005). Extant research involving political surveys suggests that scalar non-equivalence may be prevalent, complicating inferences about levels (see Davidov, Schmidt, and Schwartz 2007). One potential solution is anchoring vignettes (King et al. 2004), which provide a common reference point for respondents (as long as the vignettes themselves do not manifest any equivalence problems). A second strategy is to move away from self-reported indicators entirely, to measures that are behavioral or physiological, involve unobtrusive measurement of some kind, or draw on other signifiers of the phenomenon of interest, such as media, texts, etc.

Diagnosis

However useful prevention may be, it is often impossible because researchers are using extant, and quite valuable, datasets rather than designing their own. They can then only diagnose non-equivalence. In our examples, we have pursued two different kinds of diagnoses. One involves depictions of the results of measurement models across contextual units, as in Figure 3, that allow visual inspection of the data. While such depictions do not provide for bright-line verdicts, they can do quite a lot to illuminate the particular indicators and contexts where non-equivalence may be a problem. The second diagnostic approach, which can easily piggyback on the first, is a set of more formal statistical tests. We have outlined a useful sequence that systematically considers different kinds of equivalence. Another possibility is to estimate a structural model—whether a pure measurement model or a combination of measurement and causal models—that includes method factors. In other words, the indicators of same latent variable would be modeled as functions of that variable as well as method factors. Bollen and Paxton (1998) provide several good examples.

Perhaps even more consequential than identifying non-equivalence is specifying its sources. The common practice of exploratory equivalence testing is unsatisfying on many levels. Such studies contribute little to knowledge about the origins of equivalence or to the assessment of the relative risks of common methodological deviations in the field. Across any two country studies, multiple sources of non-equivalence

will likely exist, with roots in the social or political environment, the measurement instrument or protocol, and potentially other sources. The consequences of these various sources, non-equivalent parameters in the measurement model, may be indistinguishable. If researchers can identify sources of methodological non-equivalence that are uncorrelated with any other sources of non-equivalence, then one can gain some analytical leverage. Two such sources were evident in our investigation of the case of political involvement. Both the translation error and the timing of the survey were largely accidental and, therefore, orthogonal to any other conceivable sources of non-equivalence.

One way to identify the sources of non-equivalence is to investigate the procedures that the data-gathering organization(s) employed—advice that dovetails with that of Herrera and Kapur (2007). For example, was a survey instrument translated into one or more languages? If so, was back-translation or other additional consistency checks employed systematically? Is there any indication in the survey documentation of problems in constructing and translating items? If so, how were they handled? For non-survey data, the questions are similar. How did researchers divide the labor of scoring cases, if at all, among data coders? What standardization measures were in place for teams of coders? Were country experts given helpful reference points to anchor their measures? The answers to these questions may not lead researchers directly to specific instances of non-equivalence, but it may suggest where to start looking. In general, the more researchers interrogate the data they use, rather than simply taking it off the shelf, the better these data will become.[26]

Treatment

What to do once non-equivalence is discovered? The answer depends not only on the researchers' goals but also on the magnitude and substantive consequences of the non-equivalence. If the researchers' goal is strict comparison of latent factors means and covariances—that is, the level and interrelationships of the underlying concepts that the indicators are intended to measure—then non-equivalence of any variety is potentially serious. If a particular contextual unit, or set of units, is problematic, then one strategy is to

[26] For example, Gibson and Caldeira (2009) describe their discovery of miscodings in political knowledge items on the American National Election Study.

narrow the empirical analysis to include only those units where equivalence can be established. This can be a bitter pill, but it may be inescapable. A rosier scenario is that only a particular item or set of items lacks equivalence, such as we discovered in our investigation of economic development. In this case, it may be possible to drop those items and build measures from a smaller, but equivalent, subset of the items.

A second strategy is to employ "context-specific" or "system-specific" indicators (Adcock and Collier 2001; Przeworski and Teune 1970). In a typical formulation, there is a common set of measures across all contexts, supplemented with some context-specific measures where necessary. For example, Adcock and Collier describe Nie, Powell, and Prewitt's (1969) strategy for measuring political participation in their five-country study: they employ four standard measures for all countries, but for the United States they substitute a measure of campaign involvement for party membership, which is assumed to function differently in the United States given the weak presence of party organizations among the mass public. Although context-specific indicators will never by strictly equivalent—party membership and campaign involvement are obviously different—they may be *functionally* equivalent. Diagnostic tests could confirm, for example, that a context-specific indicator is essentially equivalent to its counterpart in other contexts. Ultimately, as Adcock and Collier note, the most important thing researchers can do is justify their decisions: "Claims about the appropriateness of contextual adjustments should not simply be asserted; their validity needs to be carefully defended" (536).

Of course, none of these strategies is mutually exclusive. Researchers might learn the most by trying different approaches—retaining indicators, dropping them, employing context-specific indicators—and then evaluating whether their substantive results change. Just as researchers often conduct sensitivity analyses in multivariate models, e.g., by reporting alternative specifications with potential confounds, they can report similar analyses with regard to measurement.

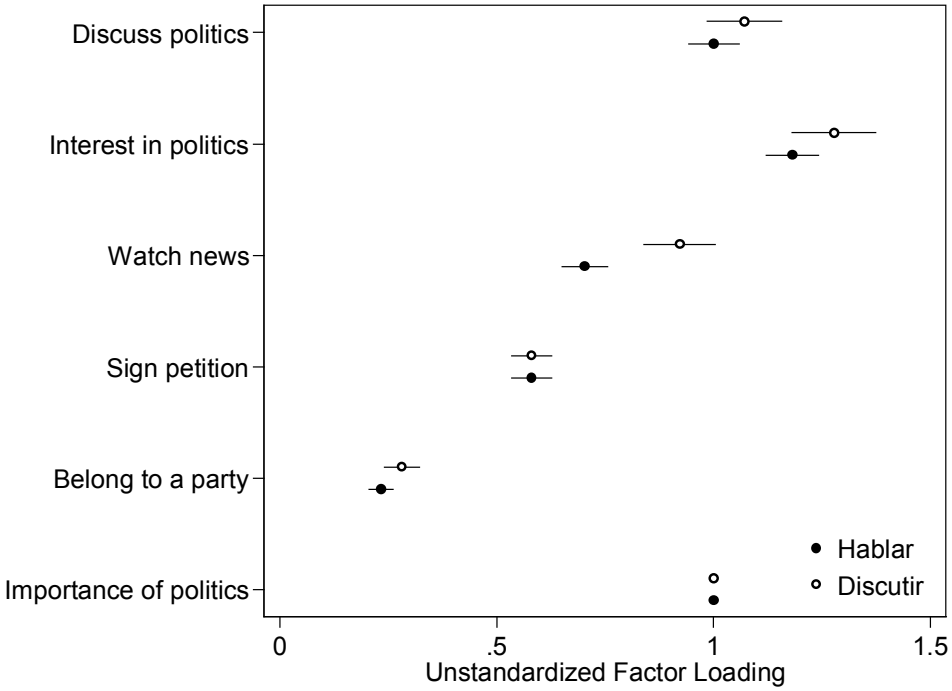
The results of sensitivity analyses might even provide more solace than sorrow. Our call for increased consciousness and even theorizing regarding possible sources of non-equivalence should not stall researchers in their tracks nor should it inject skepticism into past or future analysis of cross-national or time-series data. Indeed, if anything, the examples we present above lead one to be optimistic about the validity of

such data. Sometimes, a cigar really is just a cigar. And even when non-equivalence exists, it may not doom cross-contextual research projects. Even though we are using medical analogies here, methodological and statistical problems are not always serious diseases, with the implication that once you "catch" them, all is lost: "Oh, you have multicollinearity? I'm so sorry to hear that." Statistically significant tests for non-equivalence do not always signal substantive significance, nor do they necessarily alter the general inferences one would draw from the results. Thus, tests for non-equivalence will not always send clear signals. It is again incumbent upon researchers to craft arguments about, and marshal evidence for, their particular interpretation of equivalence tests and the consequences of non-equivalence for both measurement and inference. Transparency should always be paramount. The sciences have a variety of norms about how to report on research design and empirical results (e.g., Altman et al. 2001). Discussion of measurement equivalence should be one such norm.

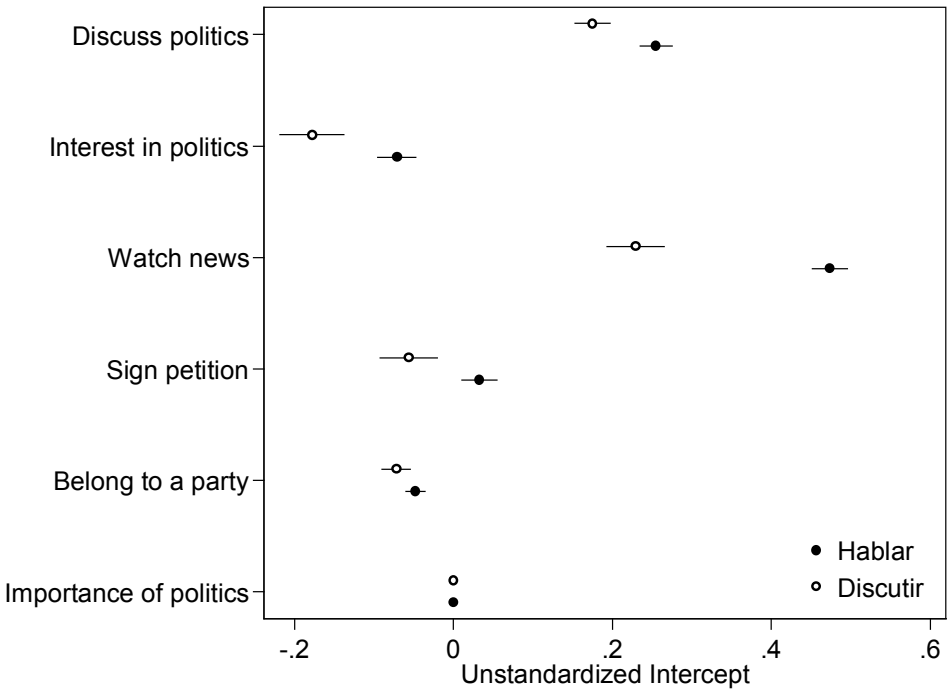
We hope that with more concerted attention to measurement, knowledge within political science would begin to accumulate. Political scientists would know more about which sources of non-equivalence are especially troublesome. They would have diagnostic reports about commonly used batteries of items, and how well these items travel across contexts. The result would be better mid-range theories about which constructs and items function differently in different contexts and why they do so. Researchers would then have an *a priori* sense of whether certain kinds of indicators are likely to be equivalent. We are confident that attention to measurement equivalence will produce important empirical findings and inform broader theoretical debates. In short, getting the measures right can lead to insights about the substantive stories that researchers wish to tell.

Figure 1. Comparing Models of Political Involvement: "Hablar" vs. "Discutir"

(a) Factor loadings



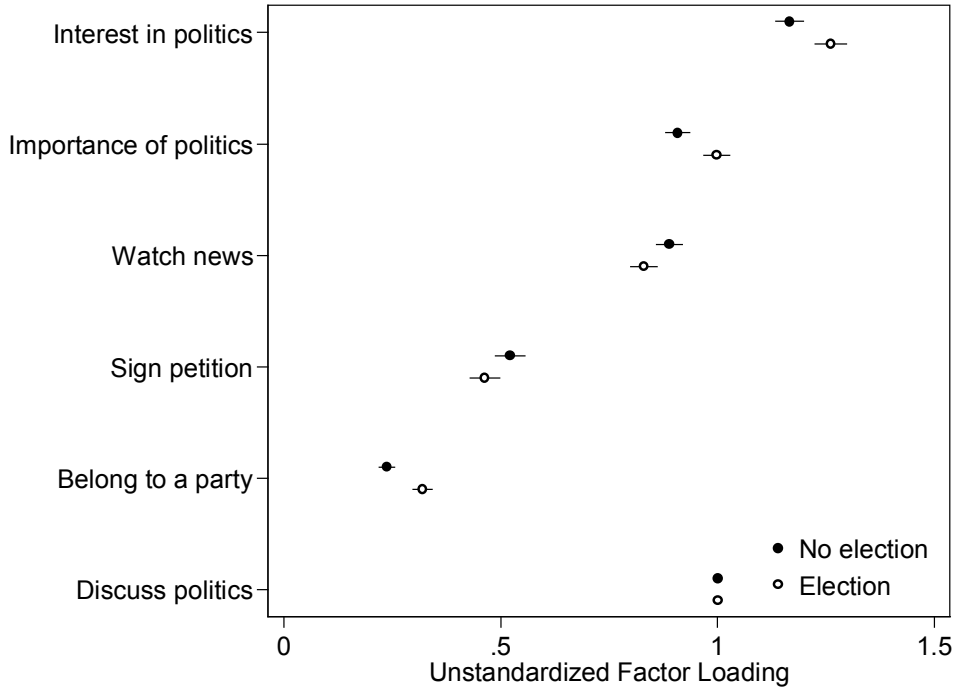
(b) Intercepts



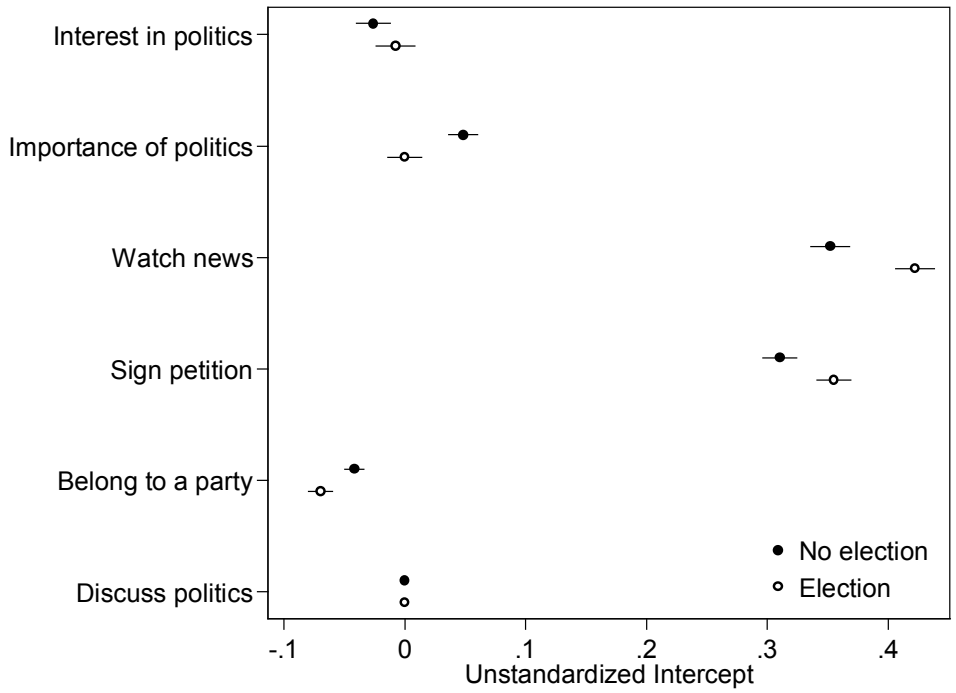
Plots present estimated loadings or intercepts, with 95% confidence intervals. Fit statistics for the model: CFI=.989; TLI=.982; RMSEA=.034. Source: 1999-2001 World Values Survey.

Figure 2. Comparing Models of Political Involvement: Proximate vs. Distant Elections

(a) Factor loadings



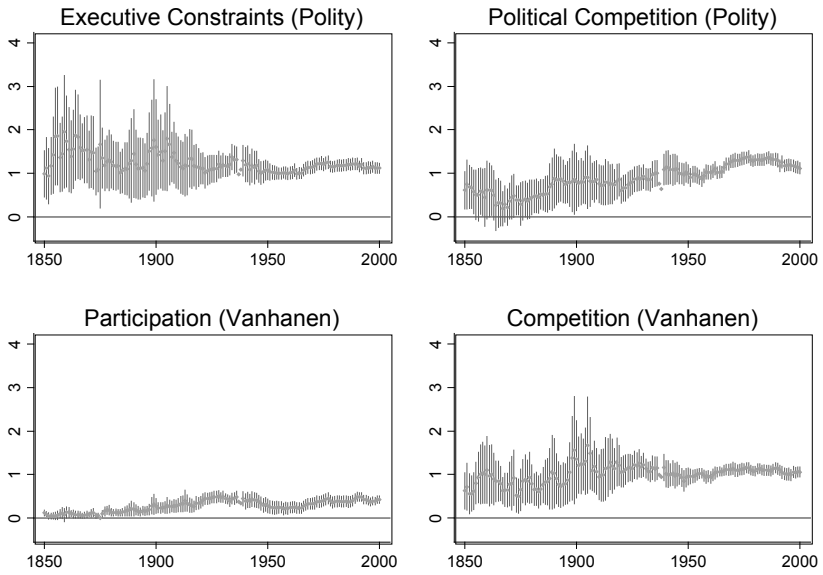
(b) Intercepts



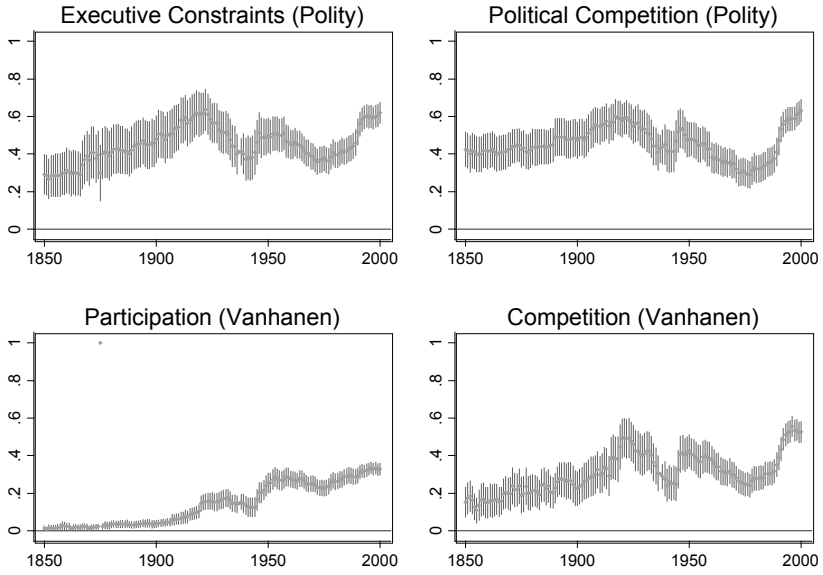
Plots present estimated loadings or intercepts, with 95% confidence intervals. Fit statistics for the model: CFI=.985; TLI=.976; RMSEA=.046. Source: 1999-2001 World Values Survey.

Figure 3. Comparing Models of Democracy across Waves (1850-2000)

(a) Unstandardized factor loadings



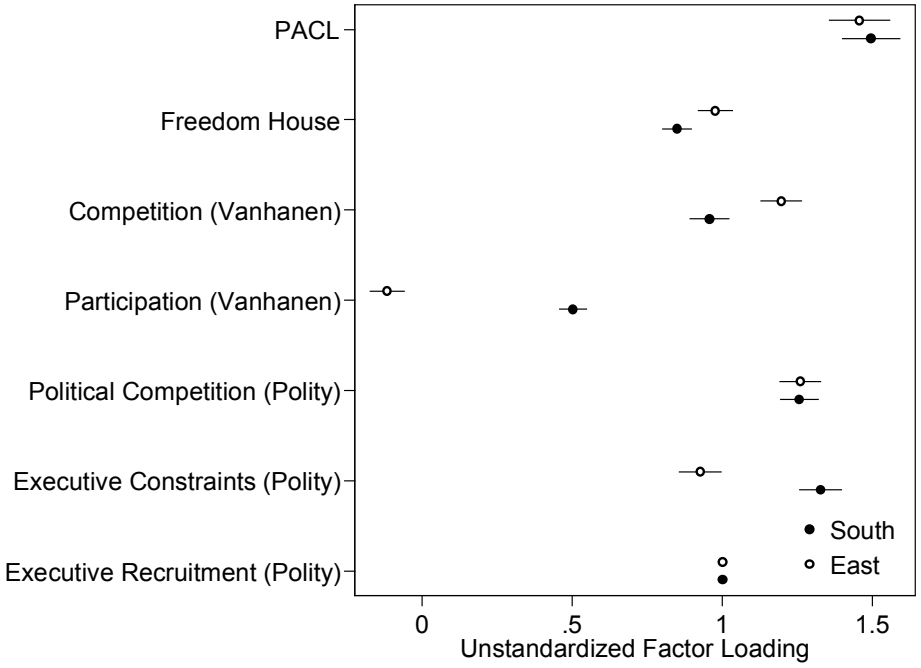
(b) Unstandardized intercepts



The model is a one-factor model, which also includes the Polity variable for executive recruitment, whose loading is constrained to one and so not depicted.

Figure 4. Comparing Models of Democracy between "East" and "South"

(a) Unstandardized Factor Loadings



(b) Unstandardized Intercepts

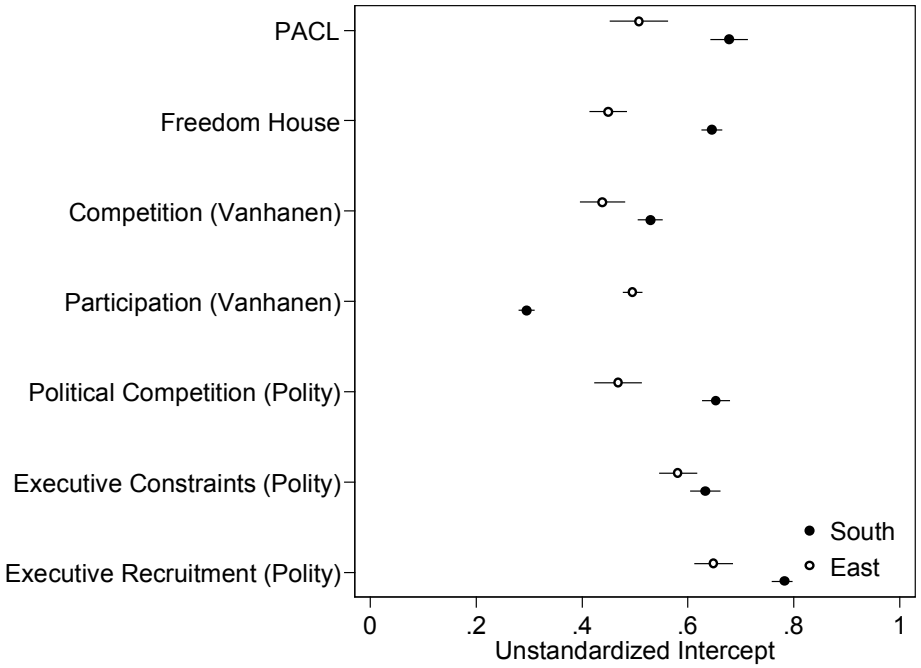
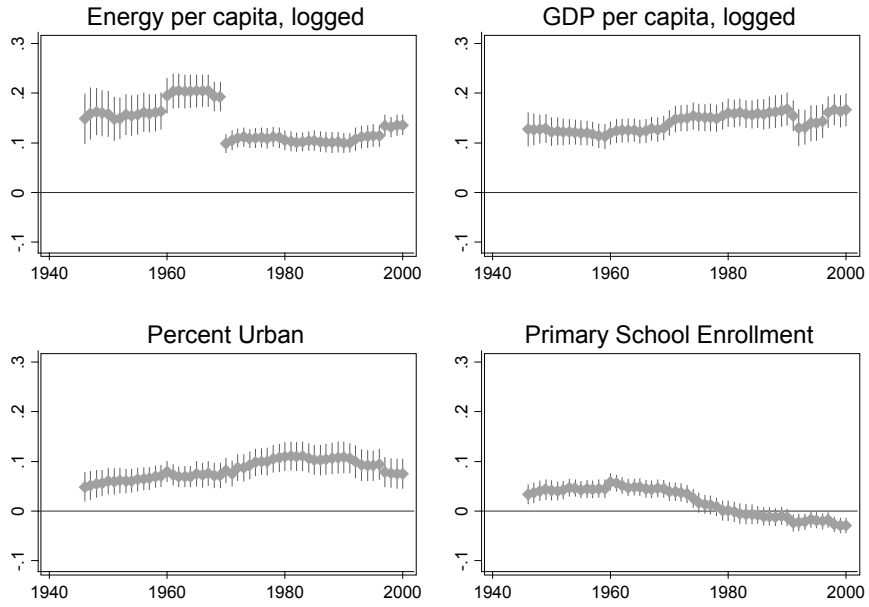
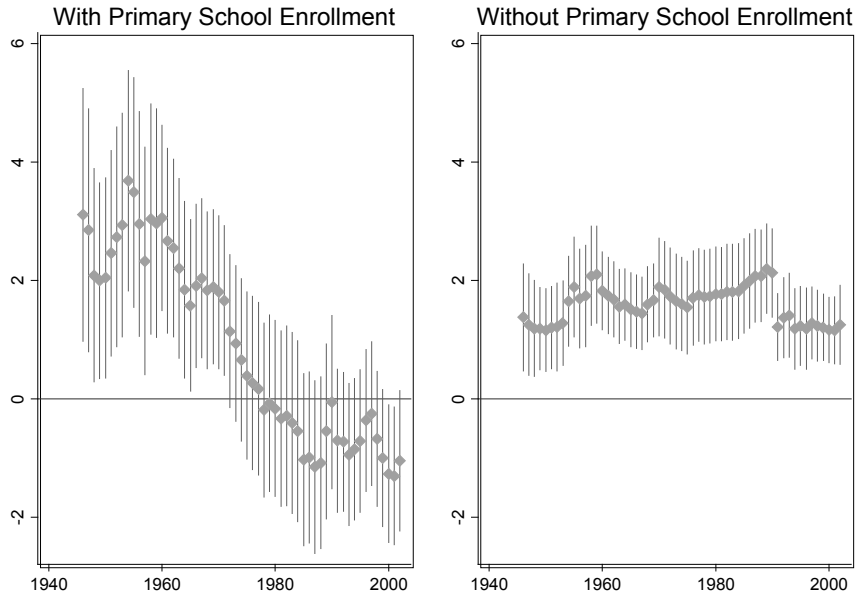


Figure 5. Comparing Models of Development Across Time

(a) Unstandardized factor loadings



(b) Coefficients from yearly regressions of Democracy (Polity) on two indices of development



Bibliography

Abdelal, Rawi, Yoshiko M. Herrera, Alastair Iain Johnston, Rose McDermott. 2009. *Measuring Identity: A Guide for Social Scientists.* New York: Cambridge University Press.

Achen, Christopher H. 1975. "Mass Political Attitude and the Survey Response." *American Political Science Review* 69: 1218-1231.

Acemoglu, Daron and James A. Robinson. 2006. *Economic Origins of Dictatorship and Democracy.* Cambridge: Cambridge University Press.

Ackerman, Terry A. 1992. "A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective." *Journal of Educational Measurement* 29(1): 67-91.

Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3): 529-46.

Almond, Gabriel, and Sidney Verba. 1963. *The Civic Culture: Political Attitudes and Democracy in Five Nations.* Princeton: Princeton University Press.

Altman D.G., Schulz K.F., Moher D., Egger M., Davidoff F., Elbourne D., Gøtzsche P.C., and Lang T. 2001. "The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration." *Annals of Internal Medicine* 134(8): 663-694.

Alvarez, Michael, Jose Antonio Cheibub, Fernando Limongi, and Adam Przeworski. 1996. "Classifying Political Regimes." *Studies in Comparative International Development* 31(2): 3-36.

Anderson, R. Bruce W. 1967. "On the Comparability of Meaningful Stimuli in Cross-Cultural Research." *Sociometry* 30(2): 124-136.

Anderson, Christopher J., and Yuliya V. Tverdova. 2003. "Corruption, Political Allegiances, and Attitudes Toward Government in Contemporary Democracies." *American Journal of Political Science* 47 (1): 91-109.

Angoff, William H. 1993. "Perspectives on Differential Item Functioning Methodology." In Paul W. Holland and Howard Wainer (eds.), *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum. pp. 3-24.

Bachman, Jerald G., and Patrick M. O'Malley. 1984. "Yea-Saying, Nay-Saying, and Going to Extremes: Black-White Differences in Response Styles." *Public Opinion Quarterly* 48 (2): 491-509.

Bailey, Michael A. 2007. "Comparable Preference Estimates across Time and Institutions for the Court, Congress, and Presidency." *American Journal of Political Science* 51(3): 433-448.

Bartels, Larry M. 1996. "Pooling Disparate Observations." *American Journal of Political Science* 40 (3): 905-42.

Blais, André, and Elisabeth Gidengil. 1993. "Things Are Not Always What They Seem: French-English Differences and the Problem of Measurement Equivalence." *Canadian Journal of Political Science* 26(3): 541-555.

Bollen, Kenneth A. 1990. "Political Democracy: Conceptual and Measurement Traps." *Studies in Comparative International Development* 25(1): 7-24.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables.* New York: Wiley.

------. 2003. "Liberal Democracy: Validity and Method Factors in Cross-National Measurement." *American Journal of Political Science* 37(4): 1207-30.

Bollen, Kenneth A., Barbara Entwisle, and Arthur S. Alderson. 1993. "Macrocomparative Research Methods." *Annual Review of Sociology* 19: 321-51.

Bollen, Kenneth A., and Pamela Paxton. 1998. "Detection and Determinants of Bias in Subjective Measures." *American Sociological Review* 63: 465-78.

Brady, David. 2003. "Rethinking the Sociological Measurement of Poverty." *Social Forces* 81 (3): 715-52.

Brady, Henry E. 1985. "The Perils of Survey Research: Inter-Personally Incomparable Responses." *Political Methodology* 11: 269-290.

Brady, Henry E. 1989. "Factor and Ideal Point Analysis for Interpersonally Incomparable data." *Psychometrika* 54 (2): 181-202.

Browne, Michael and Robert Cudeck. 1993. "Alternative Ways of Assessing Model Fit." In *Testing Structural Equation Models,* Kenneth Bollen and J. Scott Long (eds.). Newbury Park, CA: Sage.

Brubaker, Rogers. 1992. *Citizenship and Nationhood in France and Germany.* Cambridge, MA: Harvard University Press.

Bunce, Valerie. 1995. "Should Transitologists be Grounded?" *Slavic Review* 54(1): 111-27.

Boix, Carles. 2003. *Democracy and Redistribution.* Cambridge: Cambridge University Press.

Byrne, Barbara M., Richard J. Shavelson, and Bengt Muthén. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance." *Psychological Bulletin* 105(3): 456-66.

Canache, Damarys, Jeffrey Mondak, and Mitchell A. Seligson. 2001. "Meaning and Measurement in Cross-National Research on Satisfaction with Democracy." *Public Opinion Quarterly* 65:506-28.

Chan, David. 2000. "Detection of Differential Item Functioning on the Kirton Adaption-Innovation Inventory Using Multiple-Group Mean and Covariance Structure Analyses." *Multivariate Behavioral Research* 35(2): 169-99.

Cheung, Gordon W., and Roger B. Rensvold. 2000. "Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equations Modeling." *Journal of Cross-Cultural Psychology* 31 (2): 187-212.

Clarke, Harold D., Allan Kornberg, Chris McIntyre, Petra Bauer-Kaase, and Max Kaase. 1999. "The Effect of Economic Priorities on the Measurement of Value Change: New Experimental Evidence." *American Political Science Review* 93(3): 637-47.

Davidov, Eldad. 2009. "Measurement Equivalence of Nationalism and Constructive Patriotism in the ISSP: 34 Countries in Comparative Perspective." *Political Analysis* 17(1): 64-82.

Davidov, Eldad, Peter Schmidt, and Shalom H. Schwartz. 1998. "Bringing Values Back in: The Adequacy of the European Social Survey to Measure Values in 20 Countries." *Public Opinion Quarterly* 72(3): 420-45.

Davis, Darren W. 1997. "The Direction of Race of Interviewer Effects among African-Americans: Donning the Black Mask." *American Journal of Political Science* 41(1):309-22.

Dawson, Michael C. 2001. *Black Visions: The Roots of Contemporary African-American Political Ideologies.* Chicago: University of Chicago Press.

de Figueiredo, Rui, and Zachary Elkins. 2003. "Are Patriots Bigots?" *American Journal of Political Science* 47 (1): 171-188.

Diamond, Larry. 1999. Developing Democracy: Towards Consolidation. Baltimore, MD: Johns Hopkins University Press.

Drasgow, Fritz, and Charles L. Hulin. 1990. "Item Response Theory." In Marvin D. Dunnette and Leaetta M. Hough (eds.), *Handbook of Industrial and Organizational Psychology* (vol.1). Palo Alto: Consulting Psychologists Press.

Drasgow, Fritz. and Tahira M. Probst. 2005. "The psychometrics of adaptation: Evaluating measurement equivalence across languages and cultures." In R. Hambleton, P. Merenda, & C. Spielberger (eds.) *Adapting educational and psychological tests for cross-cultural assessment.* New Jersey: Lawrence Erlbaum Associates.

Elff, Martin. 2009. "Political Knowledge in Comparative Perspective: The Problem of Cross-National Equivalence of Measurement." Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL.

Epstein, Lee, Andrew D. Martin, Jeffrey A. Segal, and Chad Westerland. 2007. "The Judicial Common Space." *Journal of Law Economics and Organization* 23(2): 303-325.

Fearon, James D. 2003. "Ethnic and Cultural Diversity by Country." *Journal of Economic Growth* 8 (2): 195-222.

Gallie, W.B. 1956. "Essentially Contested Concepts." *Proceedings of the Aristotelian Society*, Vol. 56. London: Harrison and Sons. pp. 167-198.

Gibson, James L., and Gregory A. Caldeira. 2009. "Knowing the Supreme Court? A Reconsideration of Public Ignorance of the High Court." *Journal of Politics* 71(2): 429-41.

Gordon, Stacy B., and Gary Segura. 1997. "Cross National Variation in Political Sophistication of Individuals: Capability or Choice?" *Journal of Politics* 59 (1): 126-147.

Greenleaf, Eric A. 1992. "Measuring Extreme Response Style." *Public Opinion Quarterly*, 56 (3): 328-351.

Hambleton, Ronald K. 2005. "Issues, Designs, and Technical Guidelines for Adapting Tests into Multiple Languages and Cultures." In Hambleton, Ronald K., Peter F. Merenda, and Charles D. Spielberger (eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment.* Mahwah, NJ: Lawrence Erlbaum.

Hambleton, Ronald K., Peter F. Merenda, and Charles D. Spielberger (eds.). 2005. *Adapting Educational and Psychological Tests for Cross-Cultural Assessment.* Mahwah, NJ: Lawrence Erlbaum.

Harkness, Janet A., Peter Ph. Mohler, and Fons J.R. Van de Vijver (eds.). 2003a. *Cross-Cultural Survey Methods.* Hoboken, NJ: Wiley.

----------. 2003b. "Comparative Research." In Janet A. Harkness, Fons J.R. Van de Vijver, and Peter Ph. Mohler (eds.), *Cross-Cultural Survey Methods.* Hoboken, NJ: Wiley. pp. 3-18.

Heine, Steven J., Darrin R. Lehman, Kaiping Peng, and Joe Greenholtz. 2002. "What's Wrong With Cross-Cultural Comparisons of Subjective Likert Scales? The Reference-Group Effect." *Journal of Personality and Social Psychology* 82(6): 903-15.

Herrera, Yoshiko M., and Devesh Kapur. 2007. "Improving Data Quality: Actors, Incentives, and Capabilities." *Political Analysis* 15 (4): 365-86.

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

Iyengar, Shanto. 1976. "Assessing Linguistic Equivalence in Multilingual Surveys." *Comparative Politics* 8(4): 577-589.

Jennings, M. Kent and Gregory B. Markus. 1988. "Political Involvement in the Later Years: A Longitudinal Survey." *American Journal of Political Science* 32: 302-16.

Johnson, Ollie. 1999. "Pluralist Authoritarianism in Comparative Perspective: White Supremacy, Male Supremacy, and Regime Classification." *National Political Science Review* 7: 116-36.

Johnson, Timothy, Patrick Kulesa, Young Ik Cho, and Sharon Shavitt. 2005. "The Relation Between Culture and Response Styles." *Journal of Cross-Cultural Psychology* 36(2): 264-77.

Jorgenson, Dale W. 1998. "Did We Lose the War on Poverty?" *Journal of Economic Perspectives* 12 (1): 79-96.

Jusko, Karen Long, and W. Phillips Shively. 2005. "Applying a Two-Step Strategy to the Analysis of Cross-National Opinion Data." *Political Analysis* 13 (4): 327-44.

Kinder, Donald R., and Lynn M. Sanders. 1996. *Divided by Color.* Chicago: University of Chicago Press.

King, Gary, Christopher J.L. Murray, Joshua A. Solomon, and Ajay Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98 (1): 191-207.

King, Gary, and Jonathan Wand. 2007. "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes." *Political Analysis* 15: 46-66.

Lahav, G. 2004. *Immigration and Politics in the New Europe.* Cambridge: Cambridge University Press.

Lalwani, Ashok K., Sharon Shavitt, and Timothy Johnson. 2006. "What Is the Relationship Between Cultural Orientation and Socially Desirable Responding?" *Journal of Personality and Social Psychology* 90(1): 165-178.

Londregan, John B. and Keith Poole. 1996. "Does High Income Promote Democracy?" *World Politics* 49: 1-30.

Lord, Frederic M., and Melvin R. Novick 1968. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

MacIntosh, Randall. 1998. "Global Attitude Measurement: An Assessment of the World Values Survey Postmaterialism Scale." *American Sociological Review* 63(3): 452-464.

Marshall, Monty G., Keith Jaggers, and Ted Robert Gurr. 2004. Polity IV: Political Regime Transitions and Characteristics, 1800-1999.

Meade, Adam W. 2004. "A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance." *Organizational Research Methods* 7: 361-388.

Mondak, Jeffery J., and Damarys Canache. 2004. "Knowledge Variables in Cross-National Social Inquiry." *Social Science Quarterly* 85(3): 539-58.

Mondak, Jeffery J., and Mary R. Anderson. 2004. "The Knowledge Gap: A Reexamination of Gender-Based Differences in Political Knowledge." *Journal of Politics* 66(2): 492-512.

Mondak, Jeffery J., and Mitchell S. Sanders. 2003. "Tolerance and Intolerance, 1976-1998." *American Journal of Political Science* 47(3): 492-502.

Nie, Norman H., G. Bingham Powell, and Kenneth Prewitt. 1969. "Social Structure and Political Participation: Developmental Relationships, Part I." *American Political Science Review* 63(2): 361-78.

Nisbett, Richard E. 2003. *The Geography of Thought: How Asians and Westerners Think Differently…and Why.* New York: Free Press.

Nunn, Clyde Z., Harry J. Crockett, and J. Allen Williams. 1978. *Tolerance for Nonconformity.* San Francisco: Josey-Bass.

Paxton, Pamela. 1999. "Is Social Capital Declining in the United States: A Multiple Indicator Assessment." *American Journal of Sociology* 105(1): 88-127.

Poole, Keith T. 1998. "Recovering a Basic Space From a Set of Issue Scales." *American Journal of Political Science* 42(3): 954-93.

Przeworski, Adam, and Henry Teune. 1966-67. "Equivalence in Cross-National Research." *Public Opinion Quarterly* 30 (4): 551-568.

Przeworski, Adam, and Henry Teune. 1970. *Logic of Comparative Social Inquiry.* New York: John Wiley.

Raju, Nambury S., Larry J. Lafitte, and Barbara M. Byrne. 2002. "Measurement Equivalence: A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory." *Journal of Applied Psychology* 87: 517-529.

Reckase, Mark D. 1997. "The Past and Future of Multidimensional Item Response Theory." *Applied Psychological Measurement* 21(1): 25-36.

Reise, Steven P., Keith F. Widaman, and Robin H. Pugh. 1993. "Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance." *Psychological Bulletin* 114: 552-566.

Rensvold, Roger B. and Gordon W. Cheung. 2001. "Testing for Metric Invariance Using Structural Equations Models: Solving the Standardization Problem." In Chester A. Schriesheim and Linda L. Neider (eds.), *Research in Management (Vol. 1).* Greenwich, CT: Information Age Publishers. pp.25-50.

Reus-Smit, Christian. 1997. "The Constitutional Structure of International Society and the Nature of Fundamental Institutions." *International Organization* 51(4): 555-589.

Rodden, Jonathan. 2004. "Comparative Federalism and Decentralization: On Meaning and Measurement." *Comparative Politics* 36 (4): 481-500.

Rokkan, Stein, Sidney Verba, Jean Viet, and Elina Almasy. 1969. *Comparative Survey Analysis.* Paris: Mouton.

Schmitter, Phillippe C. and Terry Lynn Karl. 1994. "The Conceptual Travels of Transitologists and Consolidologists: How Far to the East Should They Attempt to Go?" *Slavic Review* 53(1): 173-85.

Schuman, Howard, Charlotte Steeh, Lawrence Bobo, and Maria Krysan. 1997. *Racial Attitudes in America: Trends and Interpretations* (rev. ed.). Cambridge: Harvard University Press.

Schwarz, Norman. 2003. "Culture-Sensitive Context Effects: A Challenge for Cross-Cultural Surveys." In Janet A. Harkness, Fons J.R. Van de Vijver, and Peter Ph. Mohler (eds.), *Cross-Cultural Survey Methods.* Hoboken, NJ: Wiley. pp. 93-100.

Sears, David O. 1988. "Symbolic Racism. In Phyllis A. Katz and Dalmas A. Taylor (eds.), *Eliminating Racism: Profiles in Controversy.* New York: Plenum Press.

Sen, Amartya. 1976. "Poverty: An Ordinal Approach to Measurement." *Econometrica* 44: 219-31.

------. 1999. *Development as Freedom.* New York: Anchor.

Sigelman, Lee. 2006. "American Political Science Review Citation Classics." *American Political Science Review* 100(4): 667-69.

Smeeding, Timothy M., Michael O'Higgins, and Lee Rainwater. 1990. *Poverty, Inequality and Income Distribution in Comparative Perspective.* Washington DC: Urban Institute Press.

Smith, Tom. 2003. "Developing comparable questions in cross-national surveys." In Harkness, Janet A., Peter Ph. Mohler, and Fons J.R. Van de Vijver (eds.), *Cross-Cultural Survey Methods.* Hoboken, NJ: Wiley.

Sniderman, Paul M., and Philip E. Tetlock. 1986. "Symbolic Racism: Problems of Motive Attribution in Political Analysis." *Journal of Social Issues* 42: 129-50.

Stark, Stephen, Oleksandr S. Chernyshenko, and Fritz Dragow. 2006. "Detecting Differential Item Functioning with Confirmatory Factor Analysis and Item Response Theory: Toward a Unified Strategy." *Journal of Applied Psychology* 91: 1292-1306.

Suchman, Lucy, and Brigitte Jordan. 1990. "Interactional Troubles in Face-to-Face Survey Interviews." *Journal of the American Statistical Association* 85(409): 232-241.

Sullivan, John L., James Piereson, and George E. Marcus. 1982. *Political Tolerance and American Democracy*. Chicago: University of Chicago Press.

Takane, Yoshio and Jan de Leeuw. 1987. "On the Relationship between Item Response Theory and Factor Analysis of Discretized Variables." *Psychometrika* 52(3): 393-408.

Triandis, Harry C. 1995. *Individualism and Collectivism*. Boulder, CO: Westview Press.

Treier, Shawn, and Simon Jackman. 2007. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1): 201-217.

Triest, Robert K. 1998. "Has Poverty Gotten Worse?" *Journal of Economic Perspectives* 12 (1): 97-114.

Van de Vijver, Fons J.R. 2003a. "Bias and Equivalence: Cross-Cultural Perspectives." In Janet A. Harkness, Fons J.R. Van de Vijver, and Peter Ph. Mohler (eds.), *Cross-Cultural Survey Methods*. Hoboken, NJ: Wiley. pp. 143-155.

--------. 2003b. "Bias and Substantive Analysis." In Janet A. Harkness, Fons J.R. Van de Vijver, and Peter Ph. Mohler (eds.), *Cross-Cultural Survey Methods*. Hoboken, NJ: Wiley. pp. 69-91.

Van de Vijver, Fons J.R., and Kwok Leung. 1997. *Methods and Data Analysis for Cross-Cultural Research*. Thousand Oaks, CA: Sage.

Vandenberg, Robert J. 2002. "Toward a Further Understanding of and Improvement In Measurement Invariance Methods and Procedures." *Organizational Research Methods* 5: 139-158.

Vanhanen, Tatu. 2000. "A New Dataset for Measuring Democracy: 1810-1998." *Journal of Peace Research* 37 (2): 51-265.

van Herk, Hester, Ype H. Poortinga, and Theo M.M. Verhallen. 2004. "Response Styles in Rating Scales: Evidence of Method Bias in Data from Six EU Countries." *Journal of Cross-Cultural Psychology* 35(3): 346-60.

Wilcox, Clyde, Lee Sigleman, and Elizabeth Cook. 1989. "Some Like It Hot: Individual Differences in Responses to Group Feeling Thermometers." *Public Opinion Quarterly* 53: 246-257.