

**The Vodka is Potent, but the Meat is Rotten!  
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

August 2008

**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.

Prepared for presentation at the 2008 Annual Meeting of the American Political Science Association, Boston, MA. We thank Fritz Drasgow, Chris Frederico, Gerry Munck, and Jonathan Wand for their comments and suggestions on an earlier draft, Sarah Gleisner for research assistance, and Eunice Chang, Svitlana Chernykh, Aya Kachi, Ika Putri, Zehra Toprak, and Sergio Wals for assistance with translation.

---

<sup>1</sup> Russian translation of the phrase, “The spirit is strong but the flesh is weak” (Smith 2003).

Until now, I'd assumed that there was a universal lexicon of meat terms (after all, 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]

Scholars of comparative politics continue to venture into more and more jurisdictions across longer stretches of time. For example, 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. Other cross-national datasets contain the universe of independent states since 1800. 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. 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, the same in Sri Lanka as it is in Costa Rica? Does a measure of democracy mean the same thing in 1850 as it does in 2000? 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.<sup>2</sup>

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 Almsy 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, Brady's (1985, 1989) work on interpersonal incomparability, Herrera and Kapur's (2007) work on improving data quality, King and co-authors' study of anchoring vignettes (King et al. 2004; King and Wand 2007), and attention to cross-cultural differences in the measures of constructs such as democracy (Bollen 1993; Treier and Jackman 2007), economic development (Sen 1999), ethnic diversity (Fearon 2003), federalism (Rodden 2004), income levels and inequality (Smeeding, O'Higgins, and Rainwater 1990), national identity (Parker 2007), political efficacy (King et al. 2004), postmaterialism (MacIntosh 1998), 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). 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. 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).

---

<sup>2</sup> 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.

Sound advice for addressing non-equivalence is scattered across social scientific treatments of conceptualization and measurement. For example, Adcock and Collier (2001) offer a sensible set of solutions to the problem (e.g., the use of context-specific indicators and adjusted common indicators) and work on anchoring vignettes suggests a clever means of improving comparability for survey items. But several other, and arguably prior, questions have received less attention. First, how do we know if an indicator manifests non-equivalence, and how do we evaluate the magnitude of any non-equivalence? Second, do indicators in major cross-national surveys and other cross-national datasets manifest non-equivalence? Scholars have not developed a standard protocol for assessing equivalence. How prevalent and harmful is it, really?

In particular, 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 conceptualize the forms that non-equivalence takes, develop theory about the sources of such non-equivalence, and demonstrate several tests of such theory. In doing so, we establish some benchmarks for equivalence of measures of two important concepts in comparative politics and summarize results from other measures that we have tested.

### **Conceptualizing Equivalence**

Non-equivalence can derive from multiple sources and can take multiple forms. If we leave aside the sources for now, we can conceptualize the forms of non-equivalence in terms of the parameters of the traditional measurement model. Suppose that a unit  $i$  (in this example, an individual) has an observed value ( $x_i$ ) for some indicator, which in turn is linearly related to the underlying, unobserved (latent) attribute ( $\xi$ ). The strength of the relationship between the latent attribute and the observed indicator is captured by a parameter ( $\lambda$ , the factor loading). The observed value of  $x$  also depends on an additional “uniqueness” parameter ( $\delta$ , i.e., the residual). An intercept term  $\mu$ , 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 Country 1 and another for Country 2:

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

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

Equations (1)-(2) are notated following convention for factor analytic measurement models (Bollen 1989: 17).

An important goal of cross-national research is to use the values of  $x$  in Countries 1 and 2 to make inferences about how much the latent attribute varies across countries. The parallel goal in historical research is to make inferences about variation across time. Either task depends crucially on equivalence in measurement.

Equivalence is a matter of degree, and it is useful to conceive of it 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  $\xi_1$  and  $\xi_2$ , we might imagine that the attributes are really  $\phi_1$  and  $\xi_2$ —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, measuring the strength of party identification in settings in which parties are banned, or measuring attitudes towards organized religion in societies without religious institutions. 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, political organizations instead of parties, or spiritual leaders instead of churches. 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.<sup>3</sup> That is, the set of observed indicators that measures the concept in one context overlaps 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 that a concept is measured with a five-indicator scale administered in Countries 1 and 2. In equations (1) and (2)  $x_i$  is replaced with a  $5 \times 1$  vector of indicators ( $\mathbf{X}$ ) that is related to the same latent attribute ( $\xi$ ) by a vector of loadings ( $\mathbf{\Lambda}$ ) and a vector of uniqueness parameters ( $\delta$ ):

$$\mathbf{X}_1 = \mu_1 + \mathbf{\Lambda}_1 \xi_1 + \delta_1 \quad (3)$$

$$\mathbf{X}_2 = \mu_2 + \mathbf{\Lambda}_2 \xi_2 + \delta_2 \quad (4)$$

Assume that for Country 1, each of the indicators in  $\mathbf{X}_1$  is a valid indicator of the latent attribute  $\xi_1$ . Structural equivalence would not obtain if, for example, any of the indicators in  $\mathbf{X}_2$  is not in fact an indicator of the latent attribute  $\xi_2$ . 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  $\xi_2$  is multi-dimensional in Country 2 but not in Country 1. 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

---

<sup>3</sup> According to Van de Vijver (2003: 154), “An instrument administered 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.”

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.<sup>4</sup>

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 ( $r = 0.29$ ) but not so in Costa Rica ( $r = 0.05$ ). 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  $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2$ , that is, that the factor loadings are equivalent.<sup>5</sup> 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 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 context of cross-national measurement, metric non-equivalence or DIF could mean, for instance, that one of the observed

---

<sup>4</sup> 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 cultures or time periods but that the “structural” relationship between that construct and a set of indicators differs across cultures or time periods.

<sup>5</sup> Cheung and Rensvold (2000) refer to this property as “factorial invariance.”

indicators is strongly associated with the latent attribute in Country 1, but only weakly associated with the attribute in Country 2. 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 Country 1 than Country 2. For example, if in Country 1, respondents 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 Country 2.

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 usually assumes all of the above criteria and 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 prevalence of acquiescence response style differed systematically across contexts, such as countries (Cheung and Resnvold 2000) or groups (Bachman and O’Malley 1984). If the denizens of Country 1 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 Country 1 will appear

“higher” on a given indicator relative to Country 2. 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 have a tendency to assign higher, or “hotter,” scores regardless of the group.

A final form of equivalence involves the variance of the residuals in each of the measurement equations. Thus, from equations (3) and (4),  $\text{var}(\delta_1) = \text{var}(\delta_2)$ . 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 ( $\Lambda$ ). An example concerns the passage from the Gospel of Matthew in this paper’s title. Imagine that Americans were asked something (admittedly in florid terms) about 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.<sup>6</sup>

---

<sup>6</sup> However, measurement error non-equivalence—as well as that of other parameters in the measurement model, such as the means of the latent variables and covariances among latent variables—are 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.”

The various kinds of equivalence draw our attention to the complexities of measurement in cross-national context 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,” 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 our estimates of the relationship between observed indicators. For example, if a researcher wanted to understand the relationship between religiosity and conservatism within each of several countries, non-equivalence is just as important as if this researcher wanted to compare mean levels of religiosity and conservatism across countries. Thus, equivalence in measurement is critical for both the assessment of mean differences across units and causal inference in a structural model that relates latent constructs.

### **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 is 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 then became one of the most forwarded stories in the newspaper’s online history and no doubt provoked many conversations around water coolers and in bar rooms nationwide. The question has no straightforward answer because Woods and Jordan played different sports, and the talents needed to excel in one sport are not as 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).<sup>7</sup> Finally,

---

<sup>7</sup> 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; Dragow and

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 of this is differences in reference points across contexts, as discussed earlier. Another 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). However, 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 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.

---

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).

One example is Parker (2007). He argues that the black Americans have distinct sources of political information and ideas that render black political ideology a distinct creed (see Dawson 2001). Thus, traditional measures of values and issue positions work quite differently among black Americans as opposed to white Americans. 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.

A final source of cross-section non-equivalence is the individual—a source that is most relevant in the study of mass behavior. Individuals, like contexts, have distinct values, orientations, and reference points that may affect response to measurement instruments, such as surveys (see Brady 1985). Non-equivalence may also arise because of variation in the survey interview, such that an individual's responses depend on the interaction with the interviewer. Those interactions vary considerably in the degree of rapport between the interviewer and respondents, in how well they communicate with each other, and thus in the quality and content of the response provided. Interviewers are, of course, trained to as to minimize idiosyncrasies by adhering to particular protocols. And yet rigid adherence to a protocol could itself create non-equivalence if additional explanation or information from the interviewer would actually 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. Is it better to have an interview protocol that calls for interviewers to do precisely the same things (i.e., behave "identically"), no matter what issues may arise in an interview? Or is it better to allow interviewers to 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 basic nature of an 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 baseball 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.

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 a result, 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

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?<sup>8</sup>

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 Country A 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).<sup>9</sup> 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

---

<sup>8</sup> 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).

<sup>9</sup> 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 various literatures—on cross-cultural measurement, interpersonal incomparability, differential item functioning—there are a number of methods devoted to diagnosing non-equivalence and to correcting it.<sup>10</sup> The extant literature has focused on methods within two data-analytic frameworks, structural equations modeling (SEM) and item response theory (IRT).<sup>11</sup> 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 some detail in Appendix 1, in the interest of clearing the underbrush for others distracted by these seemingly different approaches to measurement. Our conclusion is that the differences between the two approaches are largely terminological and notational, with one unit-of-measurement assumption separating the two. Our own empirical analysis draws on SEM.<sup>12</sup>

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; see also Cheung and Rensvold 2000). In our case, we propose and investigate 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

---

<sup>10</sup> A very productive line of inquiry rightly focuses on improving the ways in which data are gathered, e.g., via the inclusion of anchoring vignettes in survey data (King et al. 2004). Our focus here is on diagnostic techniques that can be applied to extant cross-national data, such as surveys that do not contain anchoring vignettes. Because most surveys do not contain vignettes but nevertheless provide valuable data, it is important to develop a means to evaluate the equivalence of the measures they contain.

<sup>11</sup> SEM is sometimes referred to as “covariance structure modeling” or as “mean and covariance structure modeling” (MACS).

<sup>12</sup> Both IRT and SEM are employed when there are multiple indicators of the construct in question. When there are not, these techniques are less helpful. Thus, our account does not speak to measurement challenges involve constructing or improving a single measure (e.g., the poverty rate or the level of ethnic diversity).

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 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 to the model.

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.<sup>13</sup> 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.<sup>14</sup> 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. 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

---

<sup>13</sup> 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).

<sup>14</sup> 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 against 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.

### **Two Evaluations of Equivalence: Political Involvement and Democracy**

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. Ideally, we can

accomplish four tasks: (1) identifying items that “function” differently across different groups or countries; (2) understanding the reasons why these items function differently; (3) determining the practical implications of including and removing these items (see Chan 2000); and (4) evaluating the reliability and sensitivity of the equivalence tests themselves. We test the effects of sources of non-equivalence across two constructs central to comparative politics, one measured at the individual level (political involvement) and the other at the country level (democracy). Within both 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.

### *Political Involvement*

We start with the concept of political involvement, a concept central to cross-national studies of political behavior (e.g., Almond and Verba 1963). 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 harbor several suspicions regarding equivalence. To begin with, we recall Przeworski and Teune’s admonitions regarding the items in their model of “political activity.” To summarize, bumper stickers are meaningful measures of political activity in the United States but not Poland. Consider the two activity items: signing petitions and party membership. While these activities are not quite as mundane and, presumably, context-specific as bumper-stickering, they likely 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 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.

Consider another item in the model, attention to political matters 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 we will not be able as easily to tell the 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, 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, the translator used “hablar,” which is closer to the meaning in the source question. The error was not independent since several countries used the Mexican “discutir” translation as their baseline. Thus, respondents in Mexico, Chile, Venezuela, and Colombia were asked, essentially, how often they *argued* about politics with their friends, while those in Argentina, Spain, Peru, El Salvador, and the Dominican Republic were asked something closer to the original version.<sup>15</sup> We suggest two hypotheses about how this translation discrepancy might affect equivalence. 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, since it is likely that people of equal levels of political involvement would be less likely to report “arguing” about politics than “discussing” politics.<sup>16</sup>

We are left, then, with several sources of non-equivalence that merit evaluation. Here we test the effect of two sources: (1) the discutir/hablar difference, a fairly subtle source of non-equivalence, and (2) the potentially severe differences relating to the election calendar.

*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 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 provides an initial examination of whether the translation discrepancy

---

<sup>15</sup> 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.”

<sup>16</sup> 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.”

induced a significantly different loading or intercept. The top panel of Figure 1 suggests 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 (not shown), constraining all of the factor loadings to be equal, the model fit, as measured by the CFI, TLI, and RMSEA is virtually unchanged.<sup>17</sup> Clearly, any distortions in translation were not enough to disturb metric equivalence. It is not clear why the news item would be non-equivalent across these groups, unless the hablar/discutir grouping is correlated with some other difference in measurement between the groups, something we can investigate more thoroughly in the subsequent set of analyses.

[insert Figure 1 about here]

The bottom panel of Figure 1 presents the intercepts for the two groups. Here, there are several statistically significant differences across the groups. As we hypothesized above, the intercept for “discuss politics” is lower in the “discutir” countries than in the “hablar” countries. It is possible, then, that the mistranslation 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. The fit of the model is still reasonable, suggesting that any difference in intercepts is minimal.

Overall, then, 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 other items also exhibited lower intercepts. In general, however, the prevalence of scalar non-equivalence across these items necessitates further investigation into its implications for the estimation of country means and causal inference.

---

<sup>17</sup> 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.

*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.<sup>18</sup>

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 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.<sup>19</sup>

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-

---

<sup>18</sup> Data on the timing of elections and the WVS country surveys are available from the authors.

<sup>19</sup> 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.

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 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))? 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 standard, the Polity participation scores before the women and blacks earned the (*de jure* and *de facto*) right to vote must be inflated and, it seems, incomparable to contemporary scores (Johnson 1999). In this case, the issue is probably *not* one of structural non-equivalence. The basic components of the Polity scale—participation, the election of the executive, and real constraints on the executive—arguably form a core set of indicators across the last two hundred years. Of course, one could argue that this set of measures underrepresents the concept of democracy to a certain extent. For example, outside of political participation, there

is no attention in the Polity measure 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. If, as we assume, the latent score for democracy is lower in 1875 than it is in 2000, and Polity has scored the two cases equivalently on its participation item, then either the slope or the intercept of that item must be lower in 1875 than it is in 2000. Of course, if one intends to construct a measure that is adjusted for the era's cultural norms (which is not Polity's intent, at least explicitly), then the conclusions will be reversed. In that scenario, one could assume that the latent scores for democracy are indeed equivalent, but that the slopes and/or intercepts for a non-adjusted participation measure like voter turnout will be non-equivalent across eras.

We can evaluate these forms of non-equivalence empirically. We start by building a measurement model with data that ranges across waves. At least two data sources, Polity and Vanhanen (2000), 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 comprise the Vanhanen scale (participation and competition), we can construct a measurement model that ranges back to 1816. Figure 3 plots the standardized factor loading for each indicator in the model, when estimated yearly. The data suggest significant metric non-equivalence across waves. The Vanhanen measures, which are measured objectively and, thus, are perhaps less "contextualized" (for better or worse) than a subjective measure might be, manifest dramatic differences in the loadings across waves. Competition loads around 0.5 throughout most of the first wave and around 0.9 through most of the second. The loading for participation varies even more, with loadings around 0.3 after the traumatic events of 1848, rising to 0.8 in the early part of the 20th century, and hovering at 0.4 through the second wave.

[insert Figure 3 here]

We can take a closer, and more evaluative, look at these differences by comparing two representative snapshots. Figure 4a plots the factor loadings and the intercepts for measurement models from 1875, the middle of the first wave of democracy, and 2000, twenty-five years into the third wave. All five of the

loadings show some degree of non-equivalence, but especially the Vanhanen measures and the Polity participation measure. While these look to be sizable substantive differences, only one of these three indicators, Vanhanen's competition variable, shows any difference between the two groups beyond the wide confidence intervals of the estimates. Nevertheless, these differences in loadings conform well to our expectations. Indicators of participation and electoral competition are not as closely related to the concept of democracy in the first wave as they are in the third. The intercepts (Figure 4b), meanwhile, show no signs of such non-equivalence. Nevertheless, the damage is already done, since non-equivalence of loadings represents a graver concern than would that of the intercepts, as we describe earlier.

[insert Figure 4 here]

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 abrupt breakdown of the Soviet Union and the cascade of 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.<sup>20</sup> 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

---

<sup>20</sup> 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.

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 is 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, as any finding of construct non-equivalence will certainly imply non-equivalence at the lower orders of non-equivalence. Such an exploration makes sense here, since Bunce is skeptical, but not wholly opposed to, measuring democracy in the east.

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 the transitologists had not considered before scoring countries in the region on the concept. For example, she suggests that analysts take into account the presence of members of the *ancien régime*: “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). That is, 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 5 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 5a and b plot the standardized factor loadings and intercepts, respectively, for separate models for the South and East in the year 1995. Immediately, we notice that the confidence intervals are substantially larger than they were in earlier comparisons, as we would expect with the small sample (25 and 20 for the South and East, respectively). Nevertheless, the loadings all 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. Nevertheless, it is doubtful that a reliance on turnout data could have been 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.

## **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 of some forms, depending upon the construct. Measures of democracy across time seem

particularly vulnerable to acute problems of non-equivalence. Other measurement models, however, were fairly resilient to what seemed to be significant contextual differences.

Accordingly, an initial lesson of this paper is that the mere existence of non-equivalence does not necessarily signal doom for cross-national research projects. Often, methodological and statistical problems are framed almost as diseases, with the implication that once you “catch” them, all is lost: “Oh, you have multicollinearity? I’m so sorry to hear that.” However, 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 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.

A second lesson directly concerns the ability of researchers to locate non-equivalence and interpret its meaning and consequences. In short, there is a decided need for better theorizing about the sources of non-equivalence. The common practice of exploratory equivalence testing is unsatisfying on a number of 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, there will likely exist multiple sources of non-equivalence, 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.

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. What, then, should scholars do, if anything, about potential

non-equivalence? Our prescriptions assume that analysts are secondary users of data, rather than collectors of such data—a description that we think applies to the large majority of researchers.

A first imperative is, on its face, obvious, but we suspect that it is rarely a top priority: evaluate the appropriateness of the cases (e.g., countries) in the data and the items under consideration. Are there particular cases where the phenomena of interest may be substantively very different? Straightforward ways to evaluate this question are to distinguish cases in terms of attributes known to produce differences in norms and institutions, which may in turn complicate the measure of mass behavior. Examples of such attributes include the level of economic development, the level of democracy, the length of democratic rule, the presence or absence of communist or authoritarian rule, and so forth. Simply because survey data have been gathered in a wide array of cases does not mean that researchers must analyze all of these countries at once.

Having evaluated the cases, researchers can then turn to the items themselves. A first task would be to investigate the procedures that the data-gathering organization(s) employed to ensure a degree of equivalence—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? A second task is to interrogate the indicators themselves. Are they applicable and relevant to each country? Again, cross-national differences norms and institutions will be relevant here, as will be the timing of the data collection vis-à-vis events such as elections.

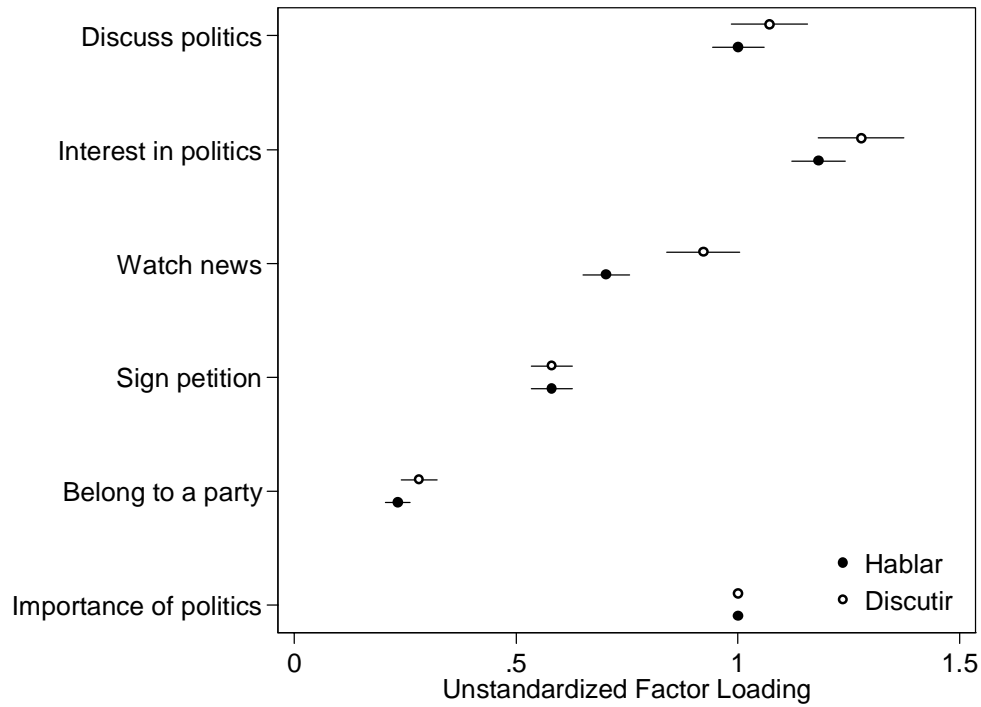
A third imperative is to conduct measurement diagnostics. There are simple methods available in basic statistical software. Such methods might include a global test of the equivalence of covariance matrices and visual inspection of factor loadings and intercepts across group-specific models. Ideally researchers would go further and employ a more formal tests using SEM or IRT. Such analysis can identify the items that are outside the bounds of randomly occurring non-equivalence.

Finally, what to do once one finds a non-equivalent item? A first task is to evaluate the substantive significance of the non-equivalence. As is true of substantive significance in other contexts—e.g., the interpretation of regression coefficients—there is no simple formula for evaluation and we cannot make any universal prescription, other than to urge researchers to marshal arguments in favor of their decision. A second solution is a familiar one: evaluate the robustness of the results under different measurement approaches. Researchers can exclude non-equivalent item(s) entirely and then reproduce the results with one or more alternative measures. Researchers might also choose to leave in the offending items but allow the parameters in question to vary across contexts. Such “partially equivalent” measures might involve country-specific parameters for some items. Ultimately, just as researchers often report alternative model specifications to demonstrate the robustness of their results, they can do the same but with alternative measurement models.

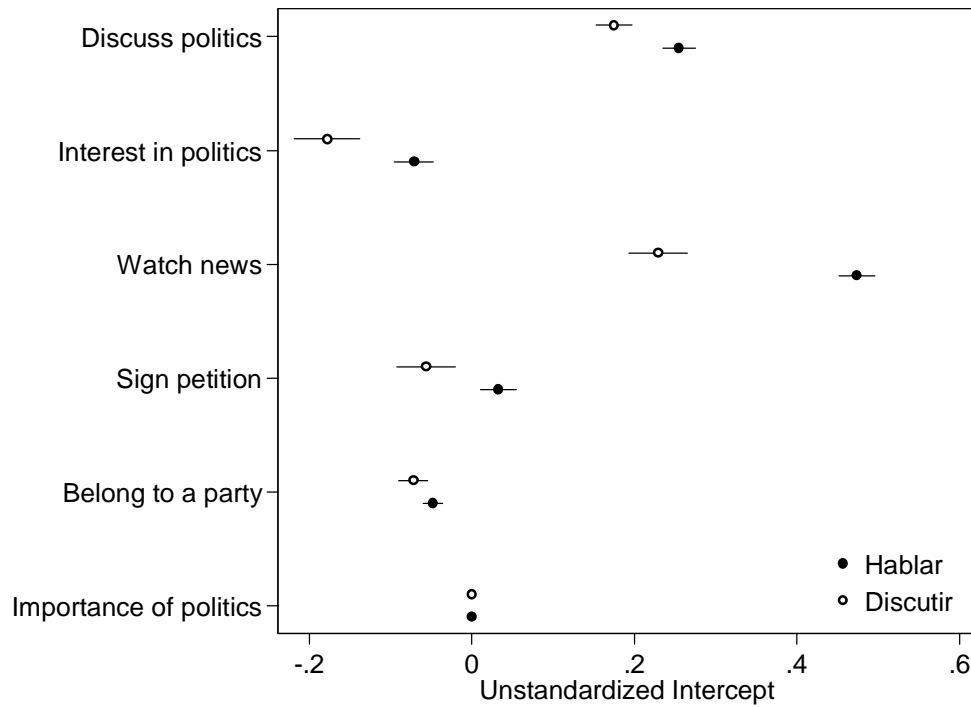
Ultimately, our hope is that with more concerted attention to measurement, knowledge within political science would begin to accumulate. We would know more about which sources of non-equivalence are especially troublesome. We would have diagnostic reports about commonly used batteries of items, and how well they travel in particular countries or regression of the world. The result would be better mid-range theories about which constructs and items function differently in different contexts, and why they do so. This would enable researchers in a variety of fields to have some *a priori* sense of whether certain kinds of indicators are likely to be equivalent. Furthermore, we are confident that attention to measurement will produce important substantive 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*



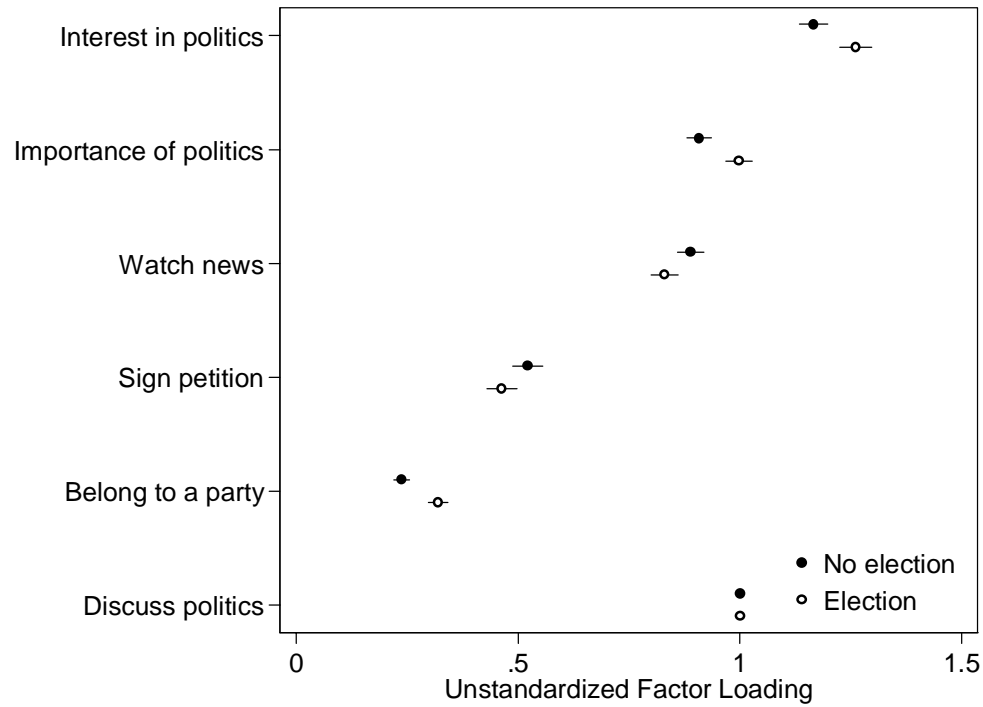
*B. Intercepts*



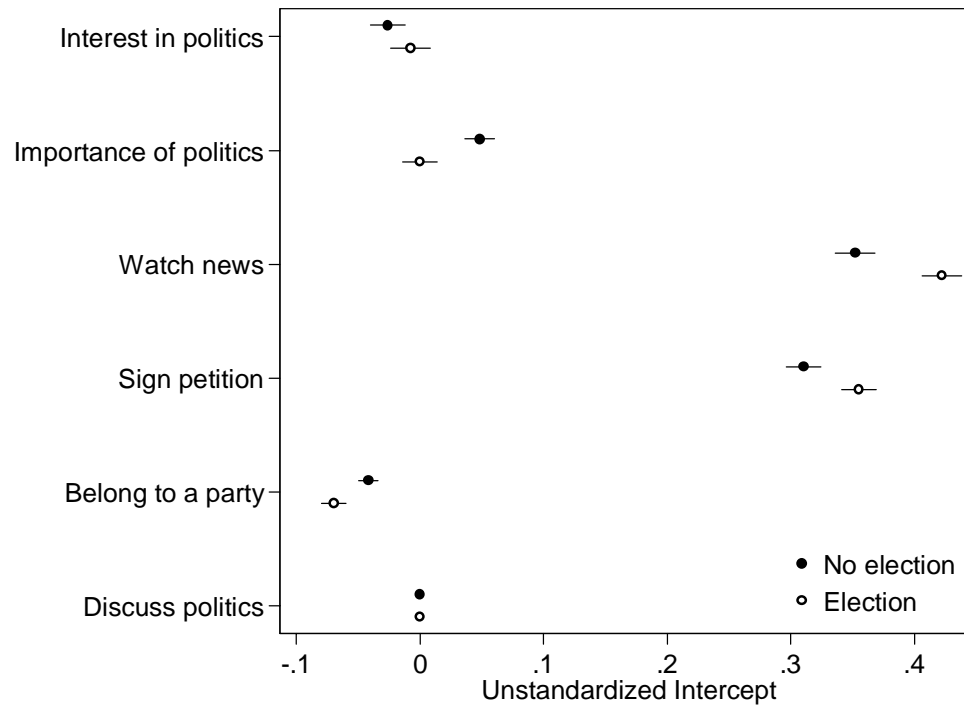
Plots present estimated loadings or intercepts, with 95% confidence intervals. Fit statistics for the model: CFI=.989; TLI=.982; RMSEA=.034. Source: 1999-2001 World Values Survey.

**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**  
Standardized Factor loadings (1830-2000)

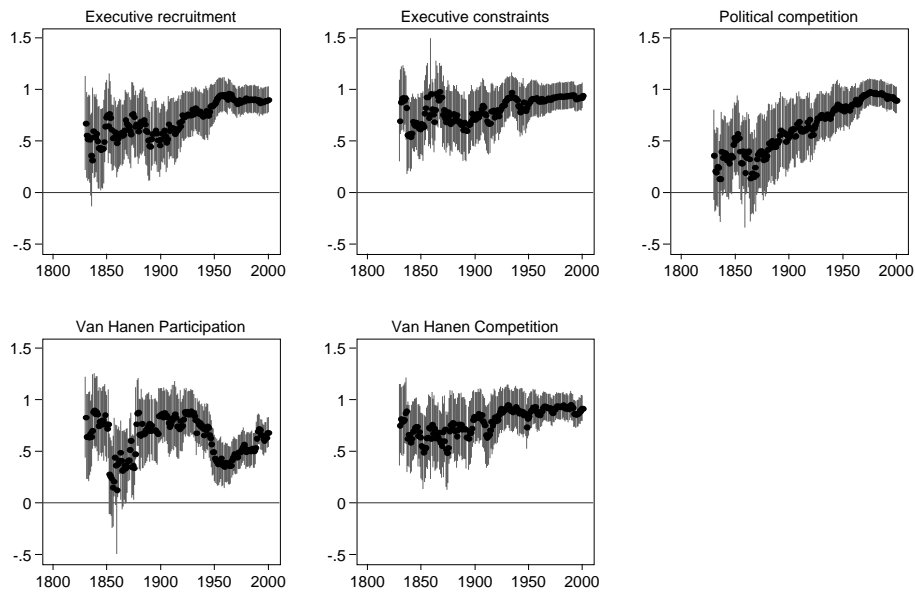
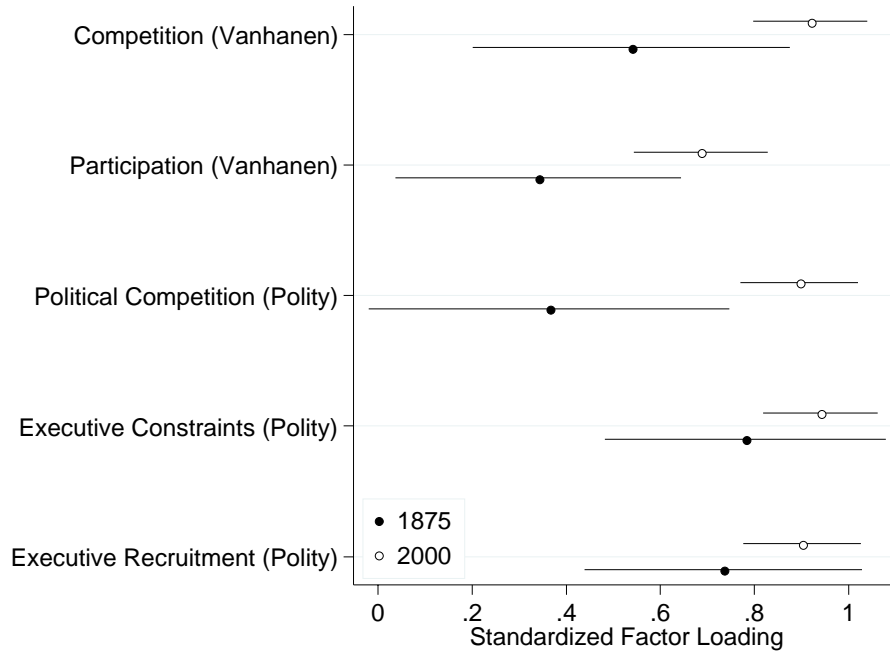


Figure 4 Comparing Models of Democracy Across Waves, 1875 versus 2000

(a) *Standardized Factor Loadings*



(b) *Standardized Intercepts, 1875 versus 2000*

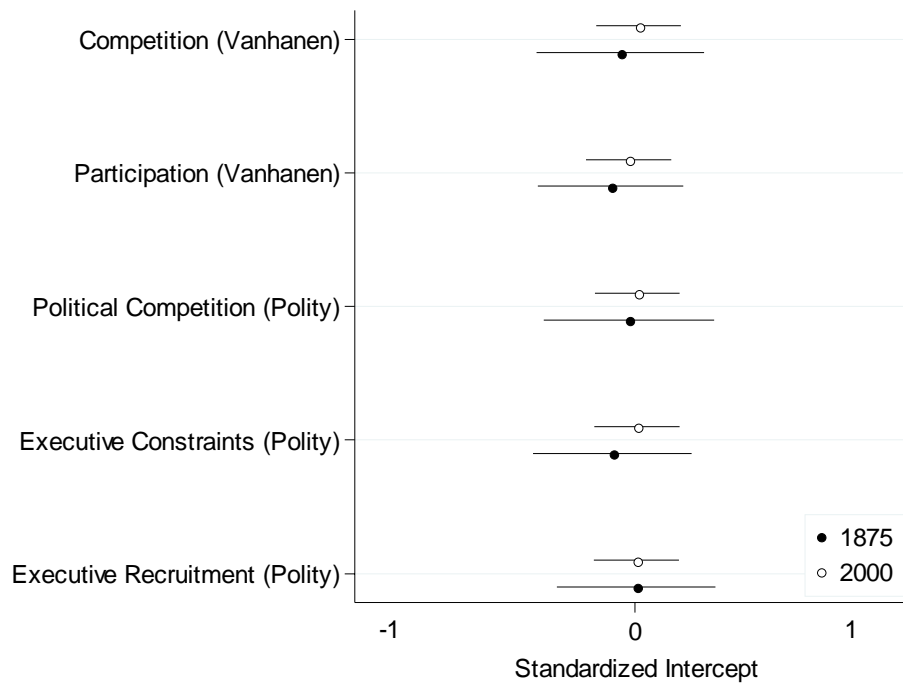
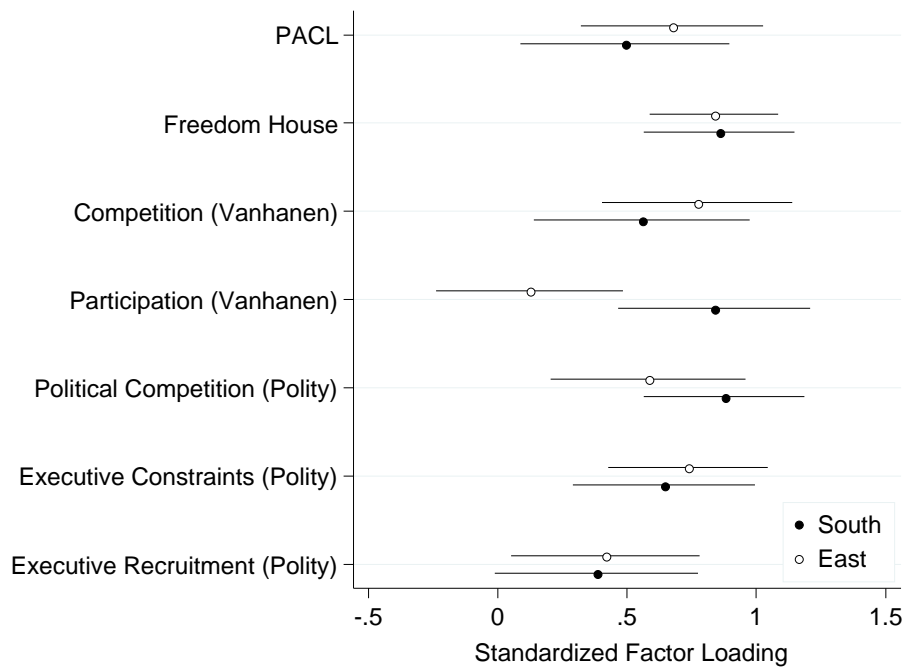
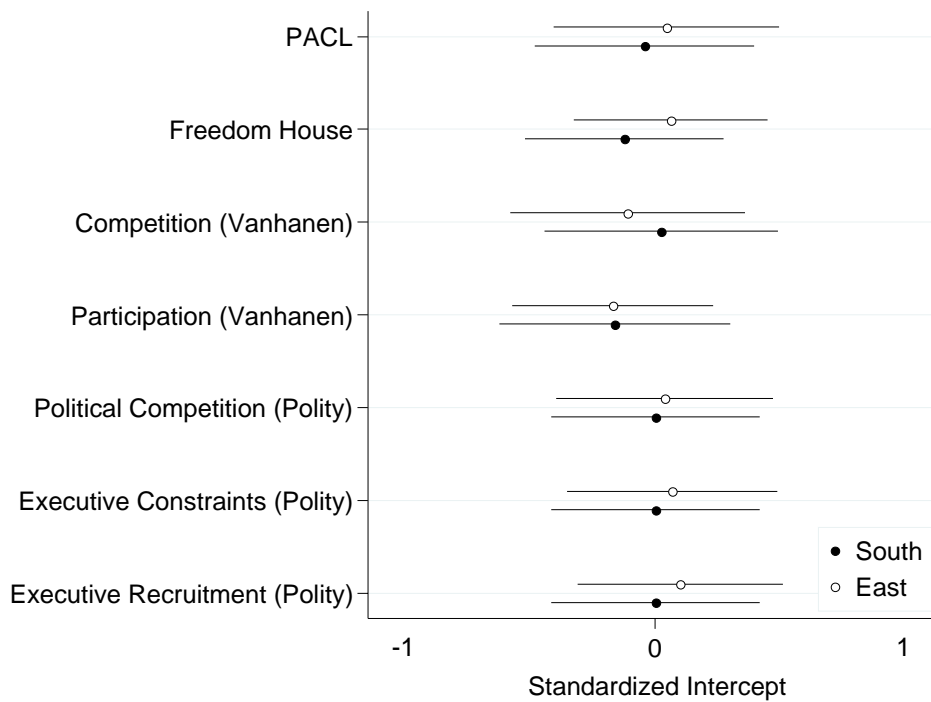


Figure 5. Comparing Models of Democracy between “East” and “South”

(a) Standardized Factor Loadings



(b) Standardized Intercepts



## Appendix 1. Data Analytic Methods of Diagnosing Equivalence: Item Response Theory (IRT) versus Structural Equation Modeling (SEM)

Despite their different disciplinary origins, terminology, and notation, IRT and SEM are virtually equivalent. In equation (4), we repeat the SEM factor-analytic notation of equation (1), with  $i$  indexing items and with no subscripts for individuals or groups (i.e., countries).

$$x_i = \mu_i + \lambda_i \xi_i + \delta_i \quad (4)$$

In equation (5), we present standard IRT notation for a two-parameter logistic model.<sup>21</sup> One can also use a normal functional form—the “normal ogive model”—in lieu of the logistic functional form. Because IRT typically concerns dichotomous test items, the observed data is modeled as the probability of giving a correct answer on item  $i$ , given a score of  $\theta$  on the underlying attribute that the test seeks to measure.

$$P_i(\theta) = \frac{1}{1 + \exp(-Da_i(\theta - b_i))} \quad (5)$$

In this model,  $a_i$  is referred to as the item discrimination parameter, and  $b_i$  is referred to as the item difficulty parameter. ( $D$  is a constant equal to 1.702, which renders the logistic model virtually identical to the normal ogive model.) The higher the value of  $b$ , the more likely an individual with a given  $\theta$  will answer the item correctly—i.e., the less difficult the item is. The higher the value of  $a$ , the more steeply the probability of a correct answer increases with increasing values of  $\theta$ ; items with higher values of  $a$  therefore better “discriminate” among individuals with different levels of  $\theta$ .

Both models share certain similarities. First, in each case, we are using observed data—the  $x_i$  in (4) and the answers to the items 1- $i$  in (5)—to estimate the individual’s “score” on some latent trait ( $\xi_i$  or  $\theta$ ). IRT’s discrimination parameters ( $a_i$ ) are analogous to SEM’s factor loadings ( $\lambda_i$ ), each of which controls the strength of the relationship between the latent trait and the observed score. The difficulty parameters ( $b_i$ ) are

---

<sup>21</sup> One can also estimate one-parameter IRT models, also known as Rasch models, as well as three-parameter models, which include a parameter,  $c$ , for guessing. We present the two-parameter model only for the purposes of illustrating the similarity between it and the factor-analytic model.

analogous to the intercepts ( $\mu_i$ ) in the factor-analytic model.<sup>22</sup> Second, with regard to non-equivalence, both methods are used to examine the same basic quantities: whether two groups with the same score on the latent construct have the same score on the observed indicators (Raju, Laffitte, and Bryne 2002). Third, either can be used fruitfully to identify the extent and source of measurement equivalence.

What are the main differences between these models? First, factor-analytic models are mathematically simpler because they assume linear relationships—that is, the observed indicators are linear combinations of the latent variables. By contrast, IRT assumes non-linear relationships. Either could be appropriate, depending on the circumstance. Second, IRT models typically assume uni-dimensionality. Naturally, this assumption can produce incorrect inferences about item bias when the latent space has two or more dimensions (Ackerman 1992). There have been advances in multi-dimensional IRT, but this area of research “is still in its infancy” (Reckase 1997: 33). Thus, the ability of IRT models to examine the dimensionality of attitudes or traits—and how this dimensionality may vary across countries—is not as well-developed as that of SEM. Third, IRT models are designed for use with dichotomous indicators and applications to polychotomous data, while possible, may not be appropriate (Raju, Laffitte, and Byrne 2002: 523-524).<sup>23</sup> Moreover, IRT models are more taxing to estimate as they include separate “boundary” parameters for the different response options (which are loosely analogous to the cutpoints estimated in a standard ordinal probit model).<sup>24</sup> Data that combine both features—multi-dimensionality and polychotomous variables—would present further challenges. Given that attitude measures on surveys often have these properties, SEM may be more appropriate. Finally, as Chan (2000) and Reise, Widaman, and Pugh (1993) note, SEM has a larger array of indices with which to assess model fit than does IRT, which generally relies on likelihood ratio chi-squared tests that are highly sensitive to sample size. Chi-squared tests in SEM

---

<sup>22</sup> Moreover, as Lord and Novick (1968) demonstrate, the factor-analytic parameters are mathematically related to parameters from the normal ogive IRT model, though in a non-linear fashion. See Takane and de Leeuw (1987) for extensions of this result.

<sup>23</sup> Drasgow and Hulin (1990) discuss several such models, their assumptions, and their applicability to different kinds of data.

<sup>24</sup> However, with polychotomous data, IRT models also provide the probability of selecting each response option, while SEM cannot do so (Raju, Laffitte, and Bryne 2002: 524).

have the same sensitivity, which is why researchers have developed these other indices (see Browne and Cudeck 1993).

The most thorough evaluation of SEM and IRT as methods for diagnosing non-equivalence is Stark, Chernyshenko, and Drasgow (2006). They conduct an extensive series of Monte Carlo tests varying numerous aspects of the simulated data: the amount of non-equivalence; whether it was located in the intercepts, loadings, or both; whether there was actual difference between the two groups in the underlying trait; whether the observed indicators were dichotomous or polychotomous; and the size of the sample. (In all simulations, however, the data were assumed unidimensional.) Their first finding concerns the choice of baseline model in the sequence of models used to test invariance. (As we discuss below, both IRT and SEM involve the comparison of nested models which assume equivalence in one set of parameters or the other.) Stark, Chernyshenko, and Drasgow (2006) demonstrate that researchers should begin by estimating an unconstrained baseline model and then estimate a series of constrained models, e.g., by constraining an item's intercept and loading to be equal across groups, examining each model's fit to see if it worsens significantly.<sup>25</sup> This sequence for evaluating model constraints seems familiar to most political scientists. However, in IRT equivalence testing, the opposite strategy has traditionally prevailed, i.e., beginning with a constrained model and then freeing parameters. Stark, Chernyshenko, and Drasgow find that this strategy led to a large Type I error rate in their simulations—that is, incorrectly signaling that an item was non-equivalent—and suggest that this is because the constrained model fails to fit the data correctly. Thus, the sequence of models is crucial and the unconstrained baseline approach is superior.<sup>26</sup>

Across the other conditions in their simulations, IRT performed slightly better than SEM when the data were dichotomous. SEM performed better than IRT when the data were polychotomous, and especially when the sample size was small ( $N=500$ ). But in general, they found that both SEM and IRT were

---

<sup>25</sup> Before doing so, one group must be selected as the reference group, and its mean on the underlying trait set to 0. Second, one indicator must be chosen as the referent. In the SEM framework, its loading is set to 1 (as is common in such models) and its intercept constrained to be equal across groups. In the IRT framework, that item's difficulty and discrimination parameters are constrained to be equal across groups. Of course, it is crucial that the referent item is itself equivalent. See Stark, Chernyshenko, and Drasgow (2006: 1304) for strategies on choosing an equivalent item.

<sup>26</sup> By contrast, Meade and Lautenschlager's (2004) comparison of these methods uses the unconstrained baseline for SEM and the constrained baseline for IRT, complicating any inferences about the utility of these methods.

“remarkably similar in their DIF detection efficacy when implemented using comparable procedures” (1303), i.e., the unconstrained baseline approach. Ultimately, then, neither approach dominates the other. The specific choice researchers make should depend on which model is most appropriate for their data.

## Bibliography

- Achen, Christopher H. 1975. "Mass Political Attitude and the Survey Response." *American Political Science Review* 69: 1218-1231.
- 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.
- 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.
- 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.
- 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.
- Canache, Damaris, 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.
- 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 Leetta 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.
- Fearon, James D. 2003. "Ethnic and Cultural Diversity by Country." *Journal of Economic Growth* 8 (2): 195-222.
- 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.

- 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.
- Johnson, Ollie. 1999. "Pluralist Authoritarianism in Comparative Perspective: White Supremacy, Male Supremacy, and Regime Classification." *National Political Science Review* 7: 116-36.
- Jorgenson, Dale W. 1998. "Did We Lose the War on Poverty?" *Journal of Economic Perspectives* 12 (1): 79-96.
- 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.
- 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 Jagers, 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 Mitchell S. Sanders. 2003. "Tolerance and Intolerance, 1976-1998." *American Journal of Political Science* 47(3): 492-502.
- 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.
- Parker, Christopher. 2007. "On the Meaning and Measurement of National Identity." Unpublished manuscript.
- Paxton, Pamela. 1999. "Is Social Capital Declining in the United States: A Multiple Indicator Assessment." *American Journal of Sociology* 105(1): 88-127.
- Przeworski, Adam, and Henry Teune. 1966-67. "Equivalence in Cross-National Research." *Public Opinion Quarterly* 30 (4): 551-568.
- 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.
- 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.
- Reckase, Mark D. 1997. "The Past and Future of Multidimensional Item Response Theory." *Applied Psychological Measurement* 21(1): 25-36.
- 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.

- 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 Dalmás 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.
- 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.
- 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.