Concepts, Theories, and Numbers: A Checklist for Constructing, Evaluating, and Using Concepts or Quantitative Measures

Gary Goertz

The Oxford Handbook of Political Methodology
Edited by Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier

Print Publication Date: Aug 2008
Online Publication Date: Sep 2009
Subject: Political Science, Political Methodology, Comparative Politics
DOI: 10.1093/oxfordhb/9780199286546.003.0005

[–] Abstract and Keywords

This article presents guidance on how to think about the concepts. It proposes to explore some issues in the assessment of concepts and quantitative measures. It presents the basic problem in its general outlines. It then offers a very short example, typically using published research. The structuring and aggregating concepts and measures indicates that one must first consider the theory embodied in the concept. Then one should survey plausible aggregation and structural relationships that could be applied in a quantitative measure. It is noted that one needs to ask about the existence or not of zero points. The gray zone needs to be explored independently of the two extremes. Homogeneity is another aspect of comparing within and between various concepts and measures of the same phenomenon. This article generally highlights that it is the lack of integration of theory and methodology which proves problematic.

Keywords: structuring concepts, aggregating concepts, quantitative measures, aggregation procedures, homogeneity, zero points, gray zone

1 Introduction

IN this chapter I propose to examine some issues in the evaluation of concepts and quantitative measures. These issues constitute a checklist of considerations when evaluating or constructing concepts and quantitative measures. They are important (p. 98) questions that the user of concepts and measures should ask when she is planning to construct, evaluate, or use them.[1]

The issues I cover can be grouped into three large categories. The first is that all complex concepts and measures use aggregation procedures. The mathematical operations used in quantitative measures need to represent theoretical considerations on the concept side, what I call the structure of the concept. Rarely do textbooks provide a list of structural or aggregation alternatives. Yet concept and measure validity depends on why and how the dimensions or indicators are aggregated.

The second set of themes deals with important points or zones along the concept or measure scale. Frequently, zero and extreme points play a crucial role in concept and measure construction. Often certain scale points are the focus of the theory to be tested. Similarly, the gray zone in the middle is a site of contention between measures and a place of important choices when dichotomizing.

The third group of considerations deals with the question of equivalence or homogeneity within or between concepts/measures. To code two observations as the "same" reflects decisions about aggregation, zero, and extreme points (among others). Yet rarely do questions about homogeneity of measurement arise. Often one asks if two measures agree on a given observation, but rarely does one ask if one measure is appropriately coding two

observations as the same.

For each issue I introduce the basic problem in its general outlines. I then provide a very short example, typically using published research. The end result (see the Checklist at the end for a summary) is a list of considerations that I think should be automatic and standard when using, constructing, and evaluating concepts and quantitative measures.[2]

2 Structure and Aggregation in Concepts and Measures

One of the most fundamental operations when constructing concepts and measures (by "measures" I mean henceforth quantitative measures or variables, including dichotomous ones) is that of *structure* or *aggregation*. I prefer the term structure because the concept or measure may not really be an "aggregation," but I will use both terms more or less interchangeably, typically using aggregation when the concept or **(p. 99)** measure involves individuals as parts. On the measure side one typically has to aggregate indicators. On the concept side one needs to structure defining characteristics. Hence a central question when evaluating or constructing a concept/measure is why and how this is done.

The qualitative literature on concepts and the quantitative literature on measures differ radically on the default approach to structure and aggregation. These differences reflect the origin of these literatures and where political scientists have borrowed ideas. The quantitative work on measurement—what I would call the Lazarsfeld–Blalock school—borrowed heavily and explicitly from psychology and educational statistics (see Lazarsfeld 1966 for a history). For example, current work on ideal point estimation (e.g. Bafumi et al. 2005) continues this tradition of borrowing from educational testing. The qualitative literature got its ideas from philosophical logic. For example, Sartori's classic 1970 article drew its basic idea of conceptual stretching directly from the classic Cohen and Nagel book (1934) on philosophical logic.

Perhaps the most fundamental difference between these two traditions is the standard way to structure or aggregate a measure or concept. Drawing on philosophical logic (going back to Aristotle) the qualitative literature has structured concepts in terms of necessary and sufficient conditions: Each part is necessary and all the parts together are jointly sufficient. Operationally this means taking the minimum (necessity) or the maximum (sufficiency) of the parts.[3] Quantitative approaches to aggregation most commonly use some additive procedure, either the sum or the mean. When presented with a bunch of indicators of a concept the natural first move is to add them up or take their mean.[4] The key point is that these qualitative and quantitative traditions provide different options on aggregation. Hence when considering a concept or measure one needs to ask about the aggregation technique and whether it is better and more appropriate than other alternatives.

One way to start to bridge the gulf between the qualitative and quantitative schools is to go borrowing from somewhere else. I suggest in this section that a good place to go when thinking about structure and aggregation is the literature on individual or social welfare, well-being, or happiness. This includes a wide range of theoretical and empirical studies from economics, development, psychology, and philosophy. The concepts of individual well-being and social welfare fundamentally deal with aggregation. Social welfare involves by definition aggregating, somehow or another, the welfare of individuals. Individual well-being involves aggregating the various domains of life such as health, family, work, and liberty that constitute individual well-being.

One of the first advantages of using the literature on well-being (individual or social) is that one moves away from the variable–indicator language typical of **(p. 100)** discussions of measurement. For example, social welfare is *constituted* by the well- being of individuals in the society. The well-being of individuals is not an indicator, but a constitutive part of social welfare.

Most quantitative scholars are deeply suspicious of language involving words like "constitutive." This is seen as typical of unclear social constructivist thinking. However, the social welfare example illustrates that such language is quite natural and reasonable. For example, Amartya Sen, a prominent player in the economics, philosophy, and development literatures on individual well-being and social welfare, frequently uses this sort of language to discuss the concept of well-being:

> The well-being of a person can be seen in terms of the quality (the "well-ness," as it were) of the person's

> being. Living may be seen as consisting of a set of interrelated "functionings," consisting of beings and doings. A person's achievement in this respect can be seen as the vector of his or her functionings. The relevant functionings can vary from such elementary things as being adequately nourished, being in good health, avoiding escapable morbidity and premature mortality, etc., to more complex achievements such as being happy, having self-respect, taking part in the life of the community, and so on. The claim is that functionings are *constitutive* of a person's being, and an evaluation of well-being has to take the form of these constitutive elements. (Sen 1992, 39; emphasis in the original).

With such a concept of individual well-being, one must aggregate in some manner or other the various functionings into a global measure.

The literature on international conflict faces the same aggregation problem as the social welfare literature, but on a much reduced scale. Instead of the aggregation of millions of individuals into a society, we have the aggregation of two countries in a dyad. In the one case we have "social" welfare, in the other we have "dyadic" concepts of democracy, trade dependence, and the like. In the former case it is, for example, the problem of aggregating individual utilities into social ones; in the latter, it is aggregating individual levels of, say, democracy, into a dyadic concept.

Table 5.1 gives a brief survey of some common variables in the literature on international militarized conflict. Many or most of these usual suspects will appear in a large-*N* study of international conflict. The first question of importance when looking at dyadic concepts in this theoretical and empirical context is whether there is aggregation at all. In Table 5.1, I have marked those variables that are inherently dyadic as "relational." Some tangos require two, such as military alliance. These are not an aggregation of country-level variables. If the list in the table is representative, then about half of commonly used variables are not aggregations.[5]

The democracy variable illustrates some of the important issues linking concept theory to quantitative measures. First, it is of note that none of the aggregation measures—including the democracy variable—uses the sum or the average. Given individual democracy levels (on a scale from −10 to 10), why not do the obvious thing (p. 101)

Table 5.1. Dyadic concepts and the study of international conflict

Dyadic concept	Sample citation	Structural relationship	Dominant structure
Democracy	Dixon (1993)	aggregation	weakest link
Trade	Gleditsch (2002)	aggregation	weakest link
Major/minor power	Mousseau (2000)	aggregation	none
Level of development	Hegre (2000)	aggregation	weakest link
Arms race	Sample (2002)	aggregation	none
Alliance	Gibler and Vasquez (1998)	relational	n.a.
Contiguity	Bremer (1992)	relational	n.a.
Power	Organski and Kugler (1980)	relational	n.a.
IGO	Oneal and Russett (1999)	relational	n.a.
Issue, territory	Senese and Vasquez (2003)	relational	n.a.

n.a.—not applicable.

Trade—level of trade dependence.

Level of development—e.g. GNP/capita.

Contiguity—geographical contiguity.

Power—military capabilities.

IGO—memberships in intergovernmental organizations.

Territory—conflict is over territory.

Source: Goertz (2006, 133).

and take the average? Some early work did in fact use some variation on the mean.[6] However, Dixon (1993) made a strong theoretical case that it was the least democratic of the dyad that determined the impact of democracy in the dyad as a whole. The "weakest-link" approach quickly became the standard used in the vast majority of studies on the liberal peace. Others, notably Russett and Oneal (2001), have extended this logic to the trade dependency variable, and Hegre (2000) has used it for the level of development variable.

The democracy variable illustrates that in good research there is a strong theory of the dyadic concept (e.g. dyadic democracy) which is used to the structure of the quantitative measure. One can contrast the strong theory of the democracy variable with another usual suspect, major power status. This variable is my candidate for most popular and least theorized of the common international conflict variables. It seems that about half of the time this is coded as "at least one major power" (i.e. maximum) and about half the time as "both major powers" (i.e. minimum). If one is constantly asking the question "what structure" and "why" then it is less likely that scholars will automatically include such undertheorized variables.

The trade dependency variable is a good example where different structures are used, but these are based on

good theoretical positions (which may or may not be (p. 102) born up in empirical analyses). For example, Barbieri (2002) has made a strong case for using the geometric mean as a measure of the salience of trade relationships. Here we have a case where differences between quantitative measures reflect real theoretical differences.

Returning to the literature on individual and social welfare, we can see that the structure question is very much about the weighting of the individual parts. Just as the weakest-link measure of dyadic democracy gives determining weight to the least democratic country, so do various theories of justice give differing weights to individuals in society. For example, theories of (social) justice have very large and direction implications for the measurement of social welfare. A Rawlsian theory puts tremendous weight on the individuals who are least well off in aggregating to the social level. A utilitarian theory in contrast gives every individual equal weight in determining social welfare. As with the dyadic democracy variable, it is a theory (in this case a normative one) that determines the weighting of the individual parts. Often we have weak theory and that results in the equal weighting of the sum or average. However, when we have stronger theory that can often lead to unequal weighting.[7] It is the philosophy of justice and welfare that determines the weighting used in any eventual quantitative measure. A wide variety of aggregation techniques have been used to implement a theory of social welfare, e.g. sum maximization (Harsanyi 1955), lexicographic priorities and maximin (Rawls 1971; Sen 1977), equality (Foley 1967; Nozick 1974; Dworkin 1981), or one of various other combining rules (Varian 1975; Suzumura 1983; Wriglesworth 1985; Baumol 1986; Riley 1987). It is because of the variety of aggregation procedures used that I have suggested the well-being and social welfare literature as a source of inspiration for thinking about how the theory embodied in concepts can be implemented in various quantitative measures.

One concept and aggregation problem Paul Diehl and I have wrestled with over the last ten years is that of the severity of a militarized interstate rivalry (Diehl and Goertz 2000). Here we see the problem of aggregation over time since a rivalry by definition is characterized by a series of militarized interactions. One question is how to aggregate those interactions into a measure of rivalry severity at any given time. One obvious option would be a weighted average of all the previous actions, with each observation exponentially discounted by its elapsed time to the present (basically this is the Crescenzi and Enterline 2001 proposal). I have recently been intrigued by prominent findings in the psychological literature on happiness. Rivalry deals with emotions and feelings of hatred, while happiness deals with the opposite, but both face the same aggregation problem. A prominent finding due to Kahneman and his colleagues (e.g. Kahneman et al. 1993; Kahneman 1999; Oliver 2004) is that current happiness follows a "peak-end" aggregation rule. Basically, current happiness is the average of the happiness at $t-1$ (i.e. "end") and the maximum happiness (i.e. "peak") over the relevant time period.

(p. 103) This is an interesting hybrid structure for a concept/measure: It uses both the average and the maximum. It means that most past periods receive no weight at all, which is the impact of the maximum. It implies that exponential memory models are dramatically off since the peak experience remains very important and shows little decay. I have no idea whether this would make sense for dyadic relationships between states, but it is an interesting aggregation option that I have permanently added to my tool kit.

This brief section on structuring and aggregating concepts and measures suggests that one must first consider the theory embodied in the concept. Then one should survey plausible aggregation and structural relationships that could be applied in a quantitative measure. A key issue throughout is the nature of the weighting scheme implied by the theory and implemented by the measure.

3 Zero Points

The zero point often plays an important role in theoretical and methodological research programs. As prospect theory and our checkbooks show, there is a major difference between positive and negative. Methodologically the existence of zero points has many important implications. A long article could easily be written on zero points; I would like to discuss an example that illustrates some key issues that users of concepts and measures should be asking about when constructing and evaluating measures.

Let me start with a personal anecdote. The zero point plays a large role in some expected utility theories of international conflict. For example, Bueno de Mesquita's (1981) main hypothesis was that a *negative* expected utility was a necessary condition for war initiation. As a result he needed a measure of preferences and utilities that

had a zero point. He developed what is known as the τ_b measure of preferences (because it is uses the τ_b statistical measure of association). When Joe Hewitt and I were looking for a measure of "willingness" to initiate a militarized conflict we immediately thought of the τ_b measure. A negative τ_b would be a signal of hostile relationships and hence a willingness to initiate militarized conflict (other factors such as weakness might prevent a country from acting on this willingness). Operationally, willingness was then a negative τ_b score for a dyad.[8]

We first presented this paper at a Peace Science Conference and Bueno de Mesquita was in the audience. In the question period he remarked that we misused his τ_b measure. The reason was that the "nominal" zero in the data (e.g. produced by the EUgene software) was not the "true" zero. The true zero point varies with system size and corresponds to a negative nominal value. As system size goes to infinity the **(p. 104)** nominal zero approaches the true zero at zero. The story ends happily with Bueno de Mesquita working with us to develop the appropriate modifications (Goertz 2006, ch. 8).

This anecdote has a number of important lessons.

The first lesson is to ask whether the theory in question does in fact need a zero point. In most uses of τ_b (or its competition *S*: Signorino and Ritter 1999; see Sweeney and Keshk 2005 for a bibliography of uses of *S* and τ_b) these measures are treated as interval ones.[9] The zero point plays no role since the hypothesis is usually of the form, the less similar the preferences the more likely war or military conflict. This correlational hypothesis does not require a zero since it only proposes that increasing probability of war with decreasing preference similarity. In this sense Bueno de Mesquita's (e.g. 1981) and our explicit use of the zero point is relatively rare. The moral is that one needs to ask whether zero plays a role in the theory and hence matters in the measure.

The second lesson is that one should ask whether the measure in fact has a zero point. The main alternative to τ_b is the *S* measure (Signorino and Ritter 1999): Does it have a zero point? If you examine the data as generated by EUgene you would say yes, because the data range from −1 to 1. However, if you look at how the data are generated, the answer is not so obviously yes. Here is a simplified version of the *S* measure (see Signorino and Ritter 1999 and Sweeney and Keshk 2005 for more details):

$$S_{ij} = 1 - 2\left(\frac{\sum_{i=1}^{N} |fpp_i - fpp_j|^k}{N}\right). \qquad (1)$$

The last step in the measure-generating process consists of 1–2(·) which standardizes the measure into the [−1,1] interval.[10] This is an arbitrary scale transformation so the resulting zero is not a real one. As one can easily see, the range of the substantive part of the measure is [0,1]. Instead of zero being a middle point it is in fact an extreme point. For example, Gibler and Rider (2004) use [0,1] *S* data, which implies that they do not see a zero point in the middle. The second lesson is thus that just because the scale of the measure has zero values does not mean it is a real zero.[11]

This leads to the third lesson: What is the measurement theory that determines the zero point? Recall that Bueno de Mesquita told us that the nominal zero was not the true zero. He must therefore have had a measurement theory about alliance **(p. 105)** configurations that he used to determine the true zero point. So one needs always to ask about the theory that determines how to measure the zero point.[12]

Braumoeller (2004) and Brambor, Clark, and Golder (2006) have brought the attention of the political science public to the fact that there are many easy-to- fall-into pitfalls in the use of interaction terms. One important implication of the presence or absence of a zero point is exactly the role ratio variables play in interaction terms.

One issue in interaction term analysis lies in the interpretation of the individual terms of the interaction term, e.g. $\beta_1 X_1$ and $\beta_2 X_2$. Typically, the interpretation is that β_1 is the impact of X_1 when $X_2 = 0$. This then assumes obviously that $X_2 = 0$ really means something. If X_2 is an interval variable then $X_2 = 0$ is completely arbitrary (see Friedrich 1982 and Allison 1977). For example, Gibler and Rider (2004) use *S* in interaction with level of threat to study alliance reliability. Since level of threat is always greater than zero, it could make a significant difference if *S* is seen to have a true zero.

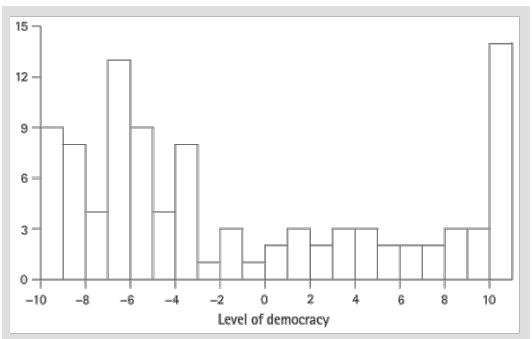
In a related manner, standardization of variables with mean zero is common. For example, Beck, King, and Zeng

(2004) do this for all the variables in their neural net analysis. These standardized variables are then used in a large variety of interaction terms.

In summary, one needs to ask about the existence or not of zero points. Does the theory need them? Does the use of the variable in interaction terms and the like imply that there is a true zero point?

4 Extreme Points and Ideal Types

The ideal-type way to construct concepts has a long and distinguished history. In the social sciences it is Max Weber (1949) who made a prominent case for this procedure (e.g. see Burger 1987 for a discussion). While scholars often use ideal types to construct concepts (e.g. Gunther and Diamond 2003), treatment of the methodology of ideal types is almost completely absent from textbooks. We lack analyses of how to construct an ideal type, or what constitutes a good ideal type. In spite of this, one can discern two distinctive characteristics of ideal types as they appear in nature: (1) the ideal type is an extreme point on the continuum, and (2) actual cases of that extreme are rare or nonexistent.



Click to view larger

Fig. 5.1. Distribution at extreme points: polity democracy measure

I have argued elsewhere (Goertz 2006, ch. 3) that ideal-type concepts are not really useful once one has a coherent system for constructing concepts. However, the idea of an ideal type does raise an important theoretical and methodological question **(p. 106)** that must be attended to when evaluating and constructing concepts: What is the distribution of cases at the ideal point extreme? Ideal-type concepts are characterized by zero cases at the extreme: Is that a good, bad, or indifferent characteristic? One can ask the contrasting question: Is it good, bad, or indifferent if there are a lot of cases at the extreme?

Figure 5.1 shows the distribution of polity democracy scores (Jaggers and Gurr 1995) for all countries 1816–1999. You will see a high spike at the democracy extreme. When I see a histogram like this my first reaction is to think that the "true" scale really extends further. Because the measure stops too soon we get a piling up cases at the barrier (Gould 1996).[13]

Looking at the polity scores for the United States might confirm the feeling that the scale stops too soon. Beginning in 1870 the United States always receives the maximum score of 10. However, the fact that large parts of the population—e.g. blacks, hispanics, Indians—in some regions, notably the South and Southwest, were either *de jure* or de facto prevented from voting after 1870 suggests that a country could be more democratic than the United States.

The moral here is that one needs to examine the distribution of cases at the extremes. "Ideal typish" concepts and measures with few cases at the extreme might often be a good goal. If our temperature scale maxed out at 100 degrees we would be mismeasuring a lot of temperatures as 100. While not necessarily conclusive evidence against a measure, large concentrations at either extreme need to be consciously justified, not accepted as "that is just what happens when you code the data."

(p. 107)

Table 5.2. Disagreement in the gray zone: level of democracy in Costa Rica, 1901–10

Year	Polity IV	Vanhanen	Gasiorowski	BLM
1901	100	0	0	0
1902	100	0	0	50
1903	100	0	0	50
1904	100	0	0	50
1905	100	0	0	0
1906	100	1	0	0
1907	100	1	0	50
1908	100	1	0	50
1909	100	1	50	50
1910	100	1	50	50

All measures have been rescaled onto the [0,100] interval.

Source: Bowman, Lehoucq, and Mahoney (2005).

5 The Gray Zone

When comparing various concepts and measures one usually finds that correlation coefficients are used to assess similarity. This procedure often dramatically underestimates the dissimilarity of measures. One reason for this is that observations at the ends of the spectrum usually have more weight (in statistical terms, more leverage; Belsley, Kuh, and Welsh 1980) than those in the middle. It is often the case that concepts and measures agree on the extreme cases since they are clear-cut and easy to code, while at the same time disagreeing frequently on cases in the middle. Points in the middle often have a "half fish, half fowl" character that makes them hard to categorize and classify. I call this area the gray zone, because values in it are neither black nor white.

Democracy is a concept where the gray zone often plays a large role in various theoretical contexts ranging from the war-proneness of transitional democracies (e.g. Mansfield and Synder 2002) to successful democratic transitions (e.g. Linz and Stepan 1996). Costa Rica has long been seen as one of the most democratic countries in Latin America. As Table 5.2 illustrates, prominent measures differ significantly on how they code Costa Rica in the crucial first decade of the twentieth century.

When there is a significant number of cases in the gray zone using a correlation coefficient as a measure of similarity can wildly underestimate discrepancies between measures. For example, take the democracy data in Figure 5.1. If one takes the cases at extreme values (i.e. −10 and 10) as given which consists of 23 percent of the data, and then replaces all the observations in between with *independent, random*, and (p. 108)

Table 5.3. Systematic disagreement in the gray zone

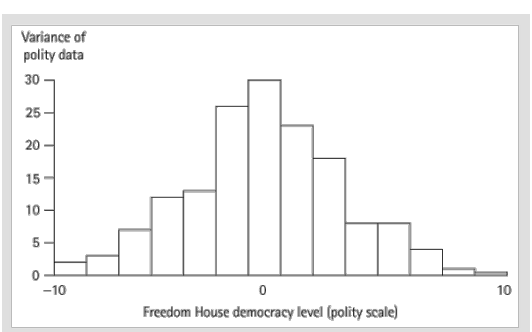
X_1	X_2					
	0	**1**	**2**	**3**	**4**	**5**
0	50	10	0	0	0	0
1	0	50	40	40	0	0
2	0	0	50	50	40	0
3	0	0	0	50	40	0
4	0	0	0	0	50	10
5	0	0	0	0	0	50

uniform data one still gets a correlation coefficient of almost .5. In short, there can exist extensive disagreement between measures in the gray zone and one can still get quite respectable correlation coefficients.

Suppose that the relationship between the two measures is like that of Table 5.3 (see Goertz 2006, ch. 3 for an example with real data). There is excellent agreement on the extremes but substantial disagreement in the middle. Yet a high correlation of .87 masks differences between the two. Notably measure X_1 is always less than measure X_2 (these kinds of triangular data patterns are not uncommon in comparative research; see also Bennett 2005, figure 1 for a triangular relationship between two dyadic democracy variables). But because a large percentage of observations do lie on the diagonal one will get substantial correlations. This example suggests that there may not only be disagreement on the middle zone, but there is a pattern to that disagreement.

Patterns of disagreement like those of Table 5.3 suggest that the *variance* between two measures changes systematically as one moves away from the extremes and toward the middle. The change in variance is driven once again by agreement at the ends and disagreement in the middle.

Figure 5.2 charts the changes in variance when comparing the polity concept and measure of democracy (Jaggers and Gurr 1995) with Freedom House's concept and measure (Karantycky 2000). To do this I added the scores of the Freedom House variables “political rights” and “civil liberties” which each range from 1 to 7. I then converted them to a -10 to 10 scale which then matches the polity scale. Figure 5.2 gives the variance of the polity scores for all cases where the Freedom House codes a nation-year at a certain level.



Click to view larger

Fig. 5.2. Variance and disagreement in the gray zone

We see then at the extremes of autocracy and democracy (i.e. -10 and 10) there is very little variance in polity codings when the Freedom House sees an extreme autocracy or democracy. For example, on the X-axis we see that there is almost no variance in the polity measure cases when the Freedom House codes a maximal democracy

(i.e. 10). As we move toward the gray zone in the middle we see that the variation in how polity codes a given nation-year increases significantly: As we move from 10 to 0 the variance increases 1,000-fold from .025 to 22.6. The same (p. 109) sort of thing happens from the autocracy side, though the increase is "only" by a factor of 10.[14]

A lesson here is that one needs to use multiple criteria to evaluate concepts and measures. In particular, the gray zone needs to be examined independently of the two extremes. Table 5.3 and Figure 5.2 illustrate two patterns that might be quite common. Table 5.3 shows a triangular relationship between measures, while Figure 5.2 shows increasing variance as one moves toward the gray zone.

We need a greater variety of techniques for evaluating concepts and measures. In particular one needs to look closely at particular parts of the concept continuum. This will depend on the theory and hypotheses being tested, but in general the extreme points and the middle always deserve special attention.

6 Homogeneity between and within Concepts and Measures

A key issue in the analysis of individual concepts as well as the comparison of two or more concepts or measures, is what Przeworski and Teune (1970) called "functional equivalence" or what I prefer to call "concept homogeneity" (Gerring and Thomas 2005 talk about "comparability"). Within a concept or measure one assigns the same value to a potentially large number of observations. The concept (p. 110) homogeneity question is whether all these observations are really instances of the same thing. For example, is the United States receiving a polity score of 10 in 1950 homogeneous or equivalent to its receiving a value of 10 in 2000? The key question in terms of constructing and evaluating concepts and measures then is the degree to which codings within a measure or between measures agree on coding observations as the same.

The homogeneity issue arises as a direct consequence of aggregation. In short, aggregation procedures produce homogeneity claims. For example, in the polity democracy measure there are a variety of ways to get, say, 5. The homogeneity claim is that all these ways are *substitutable* or *equivalent* in terms of causal analyses.

Table 5.3 illustrates the problem with concepts and measures of democracy for Costa Rica. All measures are homogeneous for the years 1909 and 1910. They see the level of democracy being the same for those two years. This is homogeneity *between* concepts, or "relative homogeneity." Posner (2004, 851) remarks that a problem with the Herfindahl index (used to study the impact of ethnic fractionalization) is that it gives quite different fractionalizations the same value. This is homogeneity *within* a measure. These are both important criteria for evaluating concepts and measures.

It is important to note that concept homogeneity is different than examining the extent to which measures or concepts agree on a given observation. For the years 1909–10 all the measures are homogeneous but they disagree radically on the level of democracy. While the degree of agreement on level is certainly correlated with the degree of homogeneity, they are conceptually separate criteria of evaluation.

Figure 5.2 directly assesses the degree of relative homogeneity of the polity and Freedom House measures of democracy. For each level of Freedom House democracy we can determine how homogeneous the polity measure is relative to Freedom House. If the polity measure and data coded democracy homogeneously with regard to Freedom House then the variance of the polity scores would be zero: In other words, polity would code the same value and the variance would be zero. Notice here we are looking at the variation of the scores, not their level. It is possible—if very unlikely— that the level is not the same. In Figure 5.2 we see that when Freedom House codes observations as completely democratic then it is almost certain that polity codes them at the same level. However, once we move into the gray zone the degree of relative homogeneity declines precipitously.

In short, homogeneity is another aspect of comparing within and between various concepts and measures of the same phenomenon. As the comparison of polity with Freedom House illustrates, the degree of relative homogeneity between measures can vary significantly along the continuum from the negative pole to the positive. Looking at the polity scores for the United States over time might suggest that there are homogeneity concerns within the polity measure. Homogeneity comparisons between and within concepts and measures should become standard practice when evaluating different concepts and measures.

(p. 111) **7 Homogeneity of Negative or Zero Cases**

We have seen that the zero point can play a key role in constructing and evaluating concepts. The zero *category* can be problematic from a homogeneity perspective, especially for dichotomous variables. Frequently the zero category is a catch-all for all observations that are "not 1." For example, Mahoney and I (2004) have analyzed this problem in the context of choosing the population of "negative" cases, which typically receive zero in a dichotomous coding, e.g. nonsocial revolutions. Sweeney and Keshk (2005) have discussed the same problem in the context of the *S* measure. In one application of *S* they use militarized dispute data coded dichotomously. They wonder about the very many zeros (i.e. no dispute) in the data since "the large number of zeros in the MID data may be due to the fact that countries did not have anything to fight about or because they chose to settle any possible conflicts in nonmilitarized ways (expressions of foreign policy preferences), or the large number of zeros may be due to the fact that countries could not engage in MIDs because they were too far apart and did not interact in any way that would give rise to the possibility of a MID (most assuredly not a foreign policy preference revelation)" (Sweeney and Keskh 2005, 174). Similarly, Goertz, Jones, and Diehl (2005) have argued that periods of zero militarized conflict after the end of a rivalry are not homogeneous as they are typically considered in "repeated conflict" studies (e.g. Werner 1999). The first fifteen years or so after the last militarized conflict are different because the rivalry is ending and there is still a possibility of further conflict. However, after those fifteen years the rivalry is over and the dyad drops out of the data-set. In repeated conflict studies the dyad remains in until the end of the period, typically 2001. Hence, Goertz et al. see heterogeneity in the zeros of repeated conflict studies. Thus in a variety of settings, the homogeneity of the "no dispute/war" observations can be called into question.[15]

The Przeworski et al. (2000) analysis of the causes and consequences of democracy illustrates the nature of the problem. Their dichotomous concept of democracy uses the necessary condition aggregation procedure on four dichotomous components. Their concept of democracy states that if a country has a zero value (dichotomously) on any one of the four components, then the country is coded as a nondemocracy. Democracy can be achieved in only one way (i.e. a one on all four components), whereas nondemocracy can occur in fifteen different ways (i.e. $2^4 - 1 = 15$).

The homogeneity hypothesis then becomes the question whether these fifteen different ways of being a nondemocracy have the same consequences for causal inference when introduced into analysis. For example, when assessing the consequences of (p. 112) nondemocracy on fertility rates, as Przeworski et al. (2000) do, can we assume that a country that has zero value on only one of the components is causally equivalent to a country that has a zero value on all four components?

Przeworski et al.'s first analysis of the relationship (2000, 81) between the level of economic development and democracy is a probit analysis with a variety of independent variables which are prominent in the literature. As an exercise, we can examine the homogeneity of the nondemocracy codings and its impact on causal inference using Przeworski et al.'s data and methods.

Given the necessary condition aggregation procedure used, we can easily rank in the zeros in terms of the number—1–4—of components that are equal to zero. One can then empirically evaluate whether the assumption of the conceptual homogeneity of zeros seems confirmed in causal analysis. Since I am also interested in comparing measures, it is useful to take a democracy measure with a structure analogous to Przeworski et al.'s for this exercise.[16]

The "modified polity" measure is one with three dimensions, "Competitiveness of Participation," "Executive Recruitment," and "Constraints on Executive" (see Goertz 2006, ch. 4 for details). The first two dimensions correspond to the two higher-level dimensions of the Przeworski et al. view of democracy which are "Contestation" and "Offices;" the former refers to multiple parties and executive turnover and the latter refers to executive and legislative offices being filled by contested elections.[17] As I have reformulated the polity measure we have three dichotomous dimensions and I require that all three be present for a country to be coded as a democracy. So structurally we have the same basic logic for the Przeworski et al. measure and the modified polity. We also have the same potential problem with the homogeneity of the nondemocracy cases, which can be zero on 1, 2, or 3 dimensions.

As is commonly reported, the correlation between the modified polity and Przeworski et al. measure of democracy

is high at .87. Przeworski et al. (2000, 56–7) say that the standard polity measure predicts 91 percent of Przeworski et al. values. If it were not for the above sections, I might claim that since correlations are high the measures are basically the same. Table 5.4 shows that in spite of a .87 correlation when using the modified polity data in Przeworski et al.'s analysis of the causes of democracy, some important differences appear. The first column of Table 5.4 replicates the probit analysis discussed in Przeworski et al. (p. 81).[18] Some variables, notably the key level of development variable, are very similar with both measures of democracy. However, about half of the variables differ significantly in sign or significance level, i.e. Stratification, Catholic, Moslem, and Ethnic Fraction(alization); consistent results show up for Development, New (p. 113)

Table 5.4. Causal homogeneity of nondemocracy: *democracy and development*

Variable	Przeworski	Polity	Modified Polity Measure		
			One zero	Two zeros	Three zeros
Intercept	−2.7976	−2.0734	.1729	−2.0839	−12.6123
($Pr > X^2$)	.0001	.0001	.6817	.0001	.0001
Development	.0003	.0003	.0002	.0004	.0018
($Pr > X^2$)	.0001	.0001	.0001	.0001	.0001
New Colony	−.8490	-1.2740	-3.7547	−1.1456	−11.4318
($Pr > X^2$)	.0001	.0001	.0001	.0001	.9998
British Colony	1.0167	1.2703	3.4428	1.4706	10.2029
($Pr > X^2$)	.0001	.0001	.0001	.0001	.9998
Stratification	−.0000	−.1420	−.2386	−.1372	∞[a]
($Pr > X^2$)	.9996	.0018	.0004	.0112	—
Catholic	.0038	−.0004	−.0058	.0000	−.0366
($Pr > X^2$)	.0005	.7336	.0206	.9951	.0103
Protestant	.0025	.0043	−.0049	.0070	.4853
($Pr > X^2$)	1028	.0131	.0707	.0010	.0001
Moslem	−.0038	−.0013	.0003	−.0005	.0225
($Pr > X^2$)	.0030	.3448	.8879	.7571	.0001
Ethnic Fraction	.0163	.070 9	−.7415	−.0517	8.4373
($Pr > X^2$)	.3242	.0472	.0001	.2843	.0001
Global Democracy	4.0812	1.9914	1.1357	1.8348	14.7266
($Pr > X^2$)	.0001	.0003	.1587	.0031	.0001
N of nondemocracy	2120	1738	346	1258	134

(a) Results are basically the same when removing the Stratification variable except for the Catholic variable, which changes signs.

Colony, British Colony, and Protestant. Here is then yet another example of how high correlations can mask

significant differences, this in the estimation of causal impacts.

The rest of the columns of Table 5.4 examine the impact of homogeneity assumptions of the negative cases. Each of these columns uses a different population of negative cases; for example, "one zero" means that for the negative cases one of the modified polity dimensions is zero but the other two are one. Hence, these negative cases are closer to democracy than the negative cases used in "three zeros" which have zero on all three polity dimensions. The cases of one on the dependent variable remain the same in all of these analyses but the number of zeros varies from column to column (they are given at the bottom of each column).

The probit results in the "one zero" column represent what might be called the "most similar" analysis. These are the negative cases most similar to the positive ones because they are missing only one dimension of democracy. Space constraints prohibit an extensive comparison, but one can look at three things when comparing across columns: (1) sign changes, (2) significance level changes, (3) trends, increasing or decreasing, in parameter estimates. Comparing the "polity" to the "one (p. 114) zero" columns we see that the central economic development variable is consistent. However, the Catholic variable which was insignificant in the polity column is now significantly negative. Overall, four variables vary in important ways between the two columns: Catholic, Protestant, Ethnic Fractionalization, and Global Democracy (ODWP in the Przeworski et al. naming scheme).

When moving further away from democracy by examining the population with two zeros constituting the negative population, we can see a pattern forming that some variables are robust while others are not. Once again the economic Development is very important along with the New Colony, British Colony, and Stratification variables. Again, the religion variables—i.e. Catholic, Protestant, and Moslem, and ethnicity—move a lot.

Moving to the least similar countries—i.e. those with zero on all three dimensions—we see very clear-cut results. All the variables are very important. In fact Stratification is a perfect predictor.[19] All the religion variables are now significant. Hence when we choose the most contrasting set of negative cases we clearly see the impact of variables which are sometimes ambiguous in other comparisons.

Of course, the numbers in Table 5.4 only provide a quick first look at the question of concept homogeneity in a causal setting. A variety of other analyses would be useful in an extended analysis. For example, one might want to run a Poisson or negative binomial regression on the number of zeros for the nondemocracy cases. This would give some idea of the extent to which the independent variables can distinguish between various kinds of nondemocracies. One would want to think about how dramatic and clear the findings tend to be when only using complete nondemocracies; the stratification variable in the "three zeros" column in Table 5.4 perfectly predicts the outcome, though here the small *N* of nondemocracies may be part of the story[20]

8 Checklist

When structuring and aggregating concepts and measures there are three related sets of items on a checklist for constructing or evaluating concepts and measures.

- What is the theory embodied in the concept?
- How is that theory translated into a quantitative measure?
- What are the plausible options for aggregation? In particular, what is the weighting scheme to be used?

(p. 115) In addition to overall evaluations of various concepts and measures, one needs to investigate individual parts or points of the scale or concept continuum.

- Are there big spikes at either extreme? Does that suggest extending the scale?
- Is there a zero point? Does the theory under examination need a zero point?
- Does the zero point or lack thereof play a role in the creation or interpretation of interaction terms?
- What is the theory that determines the zero point?
- What is going on in the gray zone? Is that zone crucial for theory testing?

All concepts and quantitative measures imply homogeneity claims. These need to be investigated.

- When comparing measures are there zones where homogeneity is low (e.g. gray zone)?
- Does homogeneity vary in a systematic manner across the continuum?
- If the measure or concept is dichotomous are there significant concerns about the homogeneity of the negative or zero cases? Should some zeros be removed from the data-set?
- Do concept homogeneity concerns appear in causal analyses? Are some variables more robust in the face of heterogeneity than others?

Of course this checklist is not exhaustive. It is a list of concerns which rarely make it into methodology and research design textbooks and courses. I have tried to illustrate briefly how these issues can arise in common data-sets and concepts. Of course, a lot will depend on the specific theory and hypothesis under investigation. This chapter stresses that it is the lack of integration of theory and methodology which proves problematic. In particular this is true of aggregation and structure problems. Typically they arise because numeric measures are not closely enough tied to the theories they are supposed to embody. The same is true of many of the issues surrounding zero points. In short, one needs continually to ask whether the numeric measures are really doing what the concepts and theories prescribe.

Bibliography

References

ALLISON, P. 1977. Testing for interaction in multiple regression. *American Journal of Sociology*, 83: 144–53.

BAFUMI, J. et al. 2005. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, 13: 171–87.

BARBIERI, K. 2002. *Liberal Illusion: Does Trade Promote Peace?* Ann Arbor: University of Michigan Press.

BAUMOL, W. 1986. *Superfairness*. Cambridge, Mass.: MIT Press.

BECK, N. **KING**, G. and **ZENG**, L. 2004. Theory and evidence in international conflict: a response to de Marchi, Gelpi, and Grynaviski. *American Political Science Review*, 98: 379–89.

(p. 116) **BELSLEY**, D. **KUH**, E. and **WELSH**, R. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley and Sons.

BENNETT, D. 2005. Towards a continuous specification of the democracy—autocracy connection. *International Studies Quarterly*, 50: 513–37.

BOWMAN, K. **LEHOUCQ**, F., and **MAHONEY**, J. 2005. Measuring political democracy: case expertise, data adequacy, and Central America. *Comparative Political Studies*, 38: 939–70.

BRAMBOR, T. **CLARK**, W., and **GOLDER**, M. 2006. Understanding interaction models: improving empirical analyses. *Political Analysis*, 14: 63–82.

BRAUMOELLER, B. 2004. Hypothesis testing and multiplicative interaction terms. *International Organization*, 58: 807–20.

BREMER, S. 1992. Dangerous dyads: interstate war, 1816–1965. *Journal of Conflict Resolution*, 36:309–41.

BUENO DE MESQUITA, B. 1981. *The War Trap*. New Haven, Conn.: Yale University Press.

BURGER, T. 1987. *Max Weber's Theory of Concept Formation: History, Laws, and Ideal Types*. Durham, NC: Duke University Press.

CHIN, H. and **QUDDUS**, M. 2003. Modeling count data with excess zeroes: an empirical application to traffic accidents. *Sociological Methods and Research*, 32: 90–116.

COHEN, M. and **NAGEL**, E. 1934. *An Introduction to Logic and Scientific Method*. New York: Harcourt, Brace.

CRESCENZI, M. and **ENTERLINE**, A. 2001. Time remembered: a dynamic model of interstate interaction. *International Studies Quarterly*, 45: 409–32.

DAVIS, J. 2005. *Terms of Inquiry: On the Theory and Practice of Political Science*. Baltimore: Johns Hopkins University Press.

DIEHL, P. and **GOERTZ**, G. 2000. *War and Peace in International Rivalry*. Ann Arbor: University of Michigan Press.

DIXON, W. 1993. Democracy and the management of international conflict. *Journal of Conflict Resolution*, 37: 42–68.

DWORKIN, R. 1981. What is equality? Part 1: equality of welfare, Part 2: equality of resources. *Philosophy and Public Affairs*, 10: 185–246, 283–345.

FOLEY, J. 1967. Resource allocation in the public sector. *Yale Economic Essays*, 7: 73–6.

FORTNA, V. 2003. Inside and out: peacekeeping and the duration of peace after civil and interstate wars. *International Studies Review*, 5: 97–114.

FRIEDRICH, R. 1982. In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science*, 26: 797–833.

GARTZKE, E. 1998. Kant we all just get along? Opportunity, willingness, and the origins of the democratic peace. *American Journal of Political Science*, 42: 1–27.

GERRING, J., and **THOMAS**, C. 2005. Comparability: a key issue in research design. Presented at the annual meetings of the American Political Science Association.

GIBLER, D. and **VASQUEZ**, J. 1998. Uncovering the dangerous alliances, 1495–1980. *International Studies Quarterly*, 42: 785–807.

—— and **RIDER**, T. 2004. Prior commitments: compatible interests versus capabilities in alliance behavior. *International Interactions*, 30: 309–29.

GLEDITSCH, K. 2002. Expanded trade and GDP data. *Journal of Conflict Resolution*, 46: 712–24.

GOERTZ, G. 2006. *Social Science Concepts: A User's Guide*. Princeton, NJ: Princeton University Press.

—— **JONES**, B. and **DIEHL**, P. 2005. Maintenance processes in international rivalries. *Journal of Conflict Resolution*, 49: 742–69.

GOULD, S. J. 1996. *Full House: The Spread of Excellence from Plato to Darwin*. New York: Three Rivers Press.

(p. 117) **GUNTHER**, R. and**DIAMOND**, L. 2003. Species of political parties: a new typology. *Party Politics*, 9: 167–99.

HARSANYI, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63: 309–21.

HEGRE, H. 2000. Development and the liberal peace: what does it take to be a trading state? *Journal of Peace Research*, 37: 5–30.

JAGGERS, K. and T. **GURR**. 1995. Tracking democracy's third wave with the Polity III data. *Journal of Peace Research*, 32: 469–82.

KAHNEMAN, D. 1999. Objective happiness. In *Well-Being: The Foundations of Hedonic Pyschol-ogy*, ed. D. Kahneman et al. New York: Russell Sage Foundation.

—— et al. 1993. When more pain is preferred to less: adding a better end. *Psychological Science*, 4: 401–5.

KARATNYCKY, A. (ed.) 2000. *Freedom in the World*, 1999–2000. Washington, DC: Freedom House.

KING, G. and **MURRAY**, C. 2002. Rethinking human security. *Political Science Quarterly*, 116: 585–610.

LAZARSFELD, P. 1966. Concept formation and measurement in the behavioral sciences: some historical observations. In *Concepts, Theory, and Explanation in the Behavioral Sciences*,ed. G. DiRenzo. New York: Random House.

LINZ, J. and **STEPAN**, A. 1996. *Problems of Democratic Transition and Consolidation: Southern Europe, South America, and Post-communist Europe*. Baltimore: Johns Hopkins University Press.

MAHONEY, J. and **GOERTZ**, G. 2004. The Possibility Principle: choosing negative cases in comparative research. *American Political Science Review*, 98: 653–69.

MANSFIELD, E. and **SYNDER**, J. 2002. Democratic transitions, institutional strength and war. *International Organization*, 56: 297–337.

MAOZ, Z. and **RUSSETT**, B. 1993. Normative and structural causes of democratic peace, 1946–1986. *American Political Science Review*, 87: 624–38.

MOUSSEAU, M. 2000. Market prosperity, democratic consolidation, and democratic peace.*Journal of Conflict Resolution*, 44: 472–507.

MUNCK, G. and **VERKUILEN**, J. 2002. Conceptualizing and measuring democracy: evaluating alternative indices. *Comparative Political Studies*, 35: 5–34.

NOZICK, R. 1974. *Anarchy, State and Utopia*. Oxford: Basil Blackwell.

OLIVER, A. 2004. Should we maximise QALYs? A debate with respect to peak-end evaluation. *Applied Health Economics and Health Policy*, 2004: 61–66.

ONEAL, J. and **RUSSETT**, B. 1999. The Kantian peace: the pacific benefits of democracy, interdependence, and international organizations, 1885–1992. *World Politics*, 52: 1–37.

ORGANSKI, A. and **KUGLER**, J. 1980. *The War Ledger*. Chicago: University of Chicago Press.

PECENY, M. and **BEER**, C. 2002. Dictatorial peace? *American Political Science Review*, 96: 15–26.

PEVEHOUSE, J. 2004. Interdependence theory and the measurement of international conflict. *Journal of Politics*, 66: 247–66.

POSNER, D. 2004. Measuring ethnic fractionalization in Africa. *American Journal of Political Science*, 48: 849–63.

PRZEWORSKI, A. et al. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950–1990*. Cambridge: Cambridge University Press.

—— and **TEUNE**, H. 1970. *The Logic of Comparative Social Inquiry*. New York: John Wiley and Sons.

RAWLS, J. 1971. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

(p. 118) **RILEY**, J. 1987. *Liberal Utilitarianism: Social Choice Theory and J.S. Mill's Philosophy*. Cambridge: Cambridge University Press.

SAMBANIS, N. 2004. What is civil war? Conceptual and empirical complexities of an operational definition. *Journal of Conflict Resolution*, 48: 814–58.

SAMPLE, S. 2002. The outcomes of military buildups: minor states vs. major powers. *Journal of Peace Research*, 39: 669–91.

SARTORI, G. 1970. Concept misformation in comparative politics. *American Political Science Review*, 64: 1033–53.

SEN, A. 1977. On weights and measures: informational constraints in social welfare analysis. *Econometrica*, 45: 1539–72.

—— 1992. *Inequality Reexamined*. Cambridge, Mass.: Harvard University Press.

SENESE, P. and **VASQUEZ**, J. 2003. A unified explanation of territorial conflict: testing the impact of sampling bias, 1919–1992. *International Studies Quarterly*, 47: 275–98.

SIGNORINO, C. and **RITTER**, J. 1999. Tau-b or not tau-b: measuring the similarity of foreign policy positions. *International Studies Quarterly*, 43: 115–44.

SUZUMURA, K. 1983. *Rational Choice, Collective Decisions and Social Welfare*. Cambridge: Cambridge University Press.

SWEENEY, K. and **KESHK**, O. 2005. The similarity of states: using *S* to compute dyadic interest similarity. *Conflict Management and Peace Science*, 22: 165–87.

VAREY, C. and **KAHNEMAN**, D. 1992. Experiences extended across time: evaluation moments and episodes. *Journal of Behavioral Decision Making*, 5: 169–86.

VARIAN, H. 1975. Distributive justice, welfare economics and the theory of fairness. *Philosophy and Public Affairs*, 4: 223–47.

WEBER, M. 1949. "Objectivity" in social science and social policy. In *The Methodology of the Social Sciences*. New York: Free Press.

WERNER, S. 1999. The precarious nature of peace: resolving the issues, enforcing the terms, and renegotiating the settlement. *American Journal of Political Science*, 43: 912–34.

WRIGLESWORTH, J. 1985. *Libertarian Conflicts in Social Choice*. Cambridge: Cambridge University Press.

Notes:

I would like to thank Bear Braumoeller, Bruce Bueno de Mesquita, David Collier, Brad Jones, Kevin Sweeney, and Chad Westerland for comments on this chapter. I would like to also thank Scott Bennett and Eric Gartzke for responding to queries regarding the *S* measure.

(1) The choice of topics arises from work on my book *Social Science Concepts: A User's Guide* (2006). They represent issues that are almost ignored in that book (e.g. the importance of zero points) or those that deserve much more attention than they were given in the book. That book focused on concept construction and only secondarily on quantitative measures. Here I reverse the balance by tilting more toward issues of constructing quantitative measures. The distinction between the two should not be pushed too far, as we shall see many methodological problems really need to be resolved first on the theoretical and conceptual side.

(2) It should be obvious that the checklist is not exhaustive. Rather, it consists of factors rarely considered but that should be.

(3) Davis (2005) criticizes the necessary and sufficient condition view of concepts from the qualitative perspective, but his proposal to use fuzzy logic remains in the domain of logic, albeit a twentieth-century kind.

(4) The big exception to this rule seems to be concepts that are used to collect populations of data. Here the dominant procedure is an implicit, necessary, and sufficient condition structure. Typically, a potential observation must satisfy all the coding rules (the sufficiency condition) and if it fails on one coding rule it is excluded from the population (i.e. necessity). See Sambanis's (2004) survey of civil war concepts and data-sets for examples of this.

(5) Aggregation issues can arise even in these relational variables. For example, if two countries have multiple alliance commitments then one must aggregate them to form a single dyadic alliance measure. Typically the strongest (i.e. maximum) alliance commitment is the aggregation procedure used in this case.

(6) Maoz and Russett (1993) use the formula $Dem_{ij} = ((Dem_h + Dem_l)/(Dem_h - Dem_l + 1))$ where Dem_h is the maximum democracy score and Dem_l is the minimum. This is interesting because it is basically a measure of how spread apart the two regime types are. This suggests one potential aggregation category based on the idea of variance; measures of inequality would fall into this category. See Bennett (2005) for another measure of spread

between regime types.

(7) Sometimes scholars think that by using necessary condition aggregation that no weighting is used. This is clearly incorrect; for an example of this confusion see King and Murray's (2002) measure of "human security." This measure is closely related to work on social welfare.

(8) Most often $\tau_b$or S is used as a control variable and hence there are no real theoretical claims regarding it, e.g. Fortna (2003) or Pevehouse (2004).

(9) An interesting question is the extent to which this is an issue for Gartzke's (1998) measure of "affinity" which uses Spearman's rank order correlation. Like τ_b this ranges from -1 to 1.

(10) In equation (1) *fpp* is the "foreign policy preference," k is a standardization parameter which makes the absolute difference in foreign policies range from zero to one. The N in the denominator then makes this the average difference in foreign policies.

(11) For example, many people rescale the polity measure of democracy (Jaggers and Gurr 1995) from its original $[-10,10]$ to $[0,20]$. As an exercise for the reader, I ask whether the zero in either of these ranges could be considered a true zero? A true zero can of course be the lowest or the highest point on a scale. See Bennett (2005) for a variety of examples where the scaling of the polity measure is important. See Beck et al. (2004, 382) who treat the polity scale as ratio.

(12) For example, Sweeney and Keshk (2005) note that if the number of categories used in constructing S increases, the measure moves toward 1. The same is true as the system size increases. Hence, there may be other comparability concerns beyond the existence or not of a zero point.

(13) See the histograms in Sweeney and Keshk (2005, e.g. figures 3 and 4) for other examples of large spikes at one extreme for the S measure.

(14) I leave it as an exercise to re-evaluate Przeworski et al.'s (2000, 58–9) argument that their dichotomous coding of democracy produces less error than a continuous measure if error follows the variance as illustrated in Figure 5.2 and the cut point between democracy and autocracy is zero.

(15) A potentially useful statistical technique for dealing with the heterogeneity of zeros is Zero-Inflated Poisson (ZIP) regression (e.g., Chin and Quddus 2003). Zeros are modeled to arrive through a "zero- event" state, i.e. where the event basically cannot happen, or through a state where $n > 0$ events can occur.

(16) The standard polity measure is a weighted average of the five indicators, hence I have preferred to use a modified polity measure with the same logical structure as the Przeworski et al. one.

(17) The polity measure is unique in its incorporation of constraints on the executive as a core part of the democracy concept. In fact, it is the most heavily weighted of the five indicators used; see Munck and Verkuilen (2002) for a discussion.

(18) The variable RELDIF—religious fractionalization—is not in the data-set for the book so it does not appear.

(19) Some software, e.g. Stata, automatically removes these very important variables because of technical problems in statistical estimation. I prefer to include them and indicate their importance with parameter estimates of "∞."

(20) It is striking how the stratification variable was not significant when using the Przeworski et al. democracy variable but was consistently important using the modified polity measure.

Gary Goertz

Gary Goertz is professor of political science at the University of Arizona. He is the author or editor of nine books and over forty articles on issues of methodology, international institutions, and conflict studies, including *Necessary Conditions: Theory, Methodology, and Applications* (Rowman & Littlefield, 2003), *Social Science Concepts: A User's Guide* (Princeton University Press, 2006), *Explaining War and Peace: Case Studies and Necessary Condition Counterfactuals* (Routledge, 2007), *Politics, Gender, and Concepts: Theory and Methodology* (Cambridge University Press, 2008), and *A Tale of Two Cultures: Contrasting Qualitative and Quantitative Paradigms* (Princeton University Press, 2012).