An Ordinal, Concept-driven Approach to Measurement: The Lexical Scale

The theoretical burden of social science is carried by highly abstract concepts such as democracy, state capacity, inequality, and rule of law (Calhoun 2002; Kuper and Kuper 1996). We require such concepts in order to articulate high-order theories. Yet, they are difficult to measure, even when agreement can be reached on a general definition. Everyone agrees that democracy means "rule by the people," but there is acute disagreement over how this general principle should be operationalized.

A key obstacle is aggregation. Faced with a number of indicators that seem relevant to a concept the researcher must decide how to combine them into a single index. Where indicators tap into common latent traits in theoretically meaningful ways the problem of aggregation may be solved by applying a data-informed (and in this sense "inductive") measurement model – of which factor analysis, structural equation models, and item response theory (IRT) are the most commonly practiced genres (DeVellis 2011). But where they do not, as is generally the case with multivalent concepts, researchers are at pains to solve the aggregation problem in a more deductive fashion.

A classic example of this problem is presented by the concept of *democracy*.[1] Some years ago Robert Dahl (1971) pointed out the complex relationship between contestation (aka competition) and participation (aka inclusion), two fundamental aspects of democracy that scarcely correlate with one another (Coppedge, Alvarez and Maldonado 2008). Polities with high levels of contestation may exhibit low levels of participation (e.g., nineteenth-century Britain) and polities with high levels of participation may exhibit low levels of contestation (e.g., present-day Cuba). Additional complexity

[1] Democracy has been the touchstone for a good deal of discussion over conceptualization and measurement in recent years (e.g., Collier and Levitsky 1997; Collier and Adcock 1999; Goertz 2006), and thus provides a fitting centerpiece for this paper. We trust that this concept is representative of problems encountered by other abstract concepts of interest to social scientists – some of which are included in the following discussion.

stems from the fact that these two components interact with each other. Specifically, the meaning of participation depends on the level of competition obtaining in a country. Voting means something quite different in contemporary Cuba and nineteenth-century Britain, for example.

Anyone who wishes to construct a composite index of democracy must grapple with this fundamental problem. Presumably, the resolution will depend upon a theoretical argument about how the components (contestation and participation) interact to produce the higher-level concept of interest (democracy). It will not depend on patterns of covariance found in the universe of nation-states. Whether lots of polities look like nineteenth-century Britain, or only a few, does not help in determining the relationship between contestation and participation in a composite index. It is a pre-empirical question.

Indeed, the most widely used indices of democracy are *concept-driven* ("deductive") insofar as aggregation schemes are derived primarily from theoretical priors that do not presume patterns of covariance across observable indicators. This includes the Polity2 index from the Polity IV database (Marshall et al. 2016), the Political rights index from Freedom House (2015), the Democracy-Dictatorship ("DD") index (Alvarez et al. 1996; Cheibub et al. 2010), and the Democracy Barometer (Bühlmann et al. 2011). For example, the DD index operationalizes democracy as a series of necessary conditions, all of which must be satisfied in order to warrant a "positive" coding (i.e., democratic) on this binary scale. Ordinal indices are more complex but in principle no less concept-driven insofar as theoretical imperatives rather than the empirical relationships among chosen indicators guide aggregation rules.

The distinction between concept-driven (deductive) and data-informed (inductive) indices is of course a matter of degrees. All indices are concept-driven to a certain extent through the selection of indicators, and "inductive" approaches are often confirmatory, using patterns in the data to verify theoretical suppositions about the relationships between latent traits and observables. After all, one

must appeal to the meaning of a concept in order to identify potential indicators – what belongs in an index and what does not. Measurement models may also incorporate theoretical priors, e.g., ways in which elements of a concept interact. In practice, however, this is rarely done. Even if strong priors are integrated into the model, indices based on some version of factor analysis, structural equation modelling, or IRT rely on patterns of covariance in the data to determine key elements of the model (e.g., factor loadings, coefficients, or discrimination parameters). This is what differentiates data-driven democracy indices (e.g., Coppedge et al. 2008; Kaufmann et al. 2010; Miller 2015; Pemstein et al. 2010) from concept-driven indices, including those listed above.[2] Such approaches are especially valuable when theory implies a relationship between the latent concept of interest and observables that is amenable to statistical analysis, because they allow researchers both to easily construct aggregate measures and test key assumptions of their theoretical models. In particular, all of these approaches assume a generative process whereby latent characteristics generate patterns of observables, at least probabilistically. When theory is not consistent with such a process, researchers must rely on purely deductive approaches to aggregation.

While a vast literature focuses on data-driven approaches to measurement, the literature on concept-driven approaches is less developed. Work in this tradition generally focuses on concept formation (e.g., Collier and Gerring 2009; Gerring 1999; Goertz 2006; Sartori 1984), typologies and taxonomies (Bailey 1994; Collier et al. 2012; Elman 2005; Lazarsfeld 1937; Lazarsfeld and Barton 1951), or on the general problem of concept or construct validity (Adcock and Collier 2001; Goertz 2006; Saylor 2013; Seawright and Collier 2014). Little attention has been paid to how one might construct a scale based on the properties of a concept. Consequently, scholars setting out to operationalize multivalent concepts do not have well-developed procedures to choose from. This is evident in the informal manner by which democracy indices have been developed over the past

[2] Note, however, that "inductive" indices of democracy generally build upon binary and ordinal indices that are composed in a highly deductive manner, so the distinction in this instance is not so clear as it might appear.

several decades. It is also evident in complaints about the ad hoc, arbitrary quality of concept-driven democracy indices (Coppedge and Gerring et al. 2011; Goertz 2006; Hadenius and Teorell 2005; Munck 2009).

This paper introduces a novel approach to scale construction that builds on the properties of concepts to solve the aggregation problem. This is accomplished by treating conceptual attributes as necessary-and-sufficient conditions arrayed in an ordinal scale. We refer to this as a *lexical* scale. Using a cumulative logic to aggregate attributes according to their logical entailments, functional dependence, and conceptual centrality, lexical scales perform both a discriminating function (offering more ordered categories than a dichotomy) and a classificatory function (as each level identifies a unique phenomenon that is qualitatively different from other levels). While different sorts of scales are useful for different purposes, we argue that lexical scales are often superior for research questions where it is relevant to combine the differentiation of an ordinal scale with the distinct, meaningful categories of a typology.

We begin by laying out the core properties of a lexical scale. Next, we offer several examples of how lexical scaling might be applied to social science concepts. The third section situates the lexical scale among other approaches to scaling, with special reference to the literature on concept formation (e.g., minimal and maximal definitions, typologies and taxonomies, and family-resemblance concepts) and the literature on measurement (e.g., Guttman scales and IRT models). The paper concludes with general observations about the strengths and weaknesses of the lexical scale relative to other approaches to scale construction.

A few notes on terminology will be helpful before we begin. *Concept formation* refers to the construction of a concept, including the choice of terms, defining characteristics (attributes), and referents. *Defining* properties of a concept refer to attributes that provide its formal definition. *Associated* properties are thought to be associated with the defining properties, but are not

definitional. *Measurement* refers to concept operationalization, i.e., the instructions or instruments required to identify membership, or degrees of membership, in the extension of a concept. This involves the construction of an *indicator* or *index* (a group of indicators combined in some fashion). *Quantitative* refers to indicators or attributes that are understood as matters of degree, while *qualitative* implies a categorical difference (of types). *Aggregation* refers to the process of combining indicators into an index. A *scale* is a generic type of indicator or index.

I. Concept-Driven Approaches

Many social science indices are concept-driven in the loose sense defined above. Typically, they involve a series of ad hoc judgments that might make sense in a particular context but do not contain rules that are generalizable to other contexts.

Only one concept-driven approach to measurement (that we are aware of) is governed by a determinate set of rules that might be regarded as a recognizable scale type. This is the *binary* scale, where a crisp-set concept is operationalized as a matter of membership (in/out) and membership criteria consist of one or more necessary conditions (jointly understood as necessary-and-sufficient) or, occasionally, a series of sufficient conditions (Goertz 2006). In either case, binary measures are generally constructed with a view to represent ordinary meanings and/or important theoretical properties of a concept. This is central to the "classical" tradition of concept formation (Collier and Gerring 2009; Sartori 1984) and to set-theoretic approaches to social science (Goertz 2006; Goertz and Mahoney 2012; Schneider and Wagemann 2012). It is also implicit in experimental and quasi-experimental studies, where treatments are usually understood in a binary fashion and are derived from a priori research hypotheses (Shadish et al. 2002).

As an example, let us give further consideration to the DD index. According to Przeworski and colleagues (Alvarez et al. 1996; Cheibub et al. 2010: 69), a regime is a democracy if political

leaders are selected through contested elections. To operationalize this conception of democracy, they identify four criteria:

1. The chief executive must be chosen by popular election or by a body that was itself popularly elected.
2. The legislature must be popularly elected.
3. There must be more than one party competing in the elections.
4. An alternation in power under electoral rules identical to the ones that brought the incumbent to office must have taken place.

Like many binary scales, the DD index adopts a minimal definition of democracy and operationalizes it with a series of necessary conditions (in this case, four), all of which must be satisfied in order to receive a score of 1 (=democracy).

For many concepts of importance there is a binary version that relies on necessary conditions to define membership in the positive category. This is, as we have observed, a well-established method of scale construction and deserves to be recognized as a "type" even though the method itself is fairly commonsensical and only one step removed from the classical tradition of concept formation.

That said, we must bear in mind a common complaint about binary measures when imposed on complex concepts – namely, that they reduce all aspects of that concept to two categories, converting a plethora of information into a series of 0's and 1's (Collier and Adcock 1999; Elkins 2000). This serves as a fitting segue to our proposal, which may be regarded as an extension of the principles of binary scaling.

II. A Lexical Scale

We propose to preserve the virtues of a conceptually driven approach to measurement while honoring the need for greater differentiation than is provided by binary measures. This, in brief, is the strategy of the *lexical* scale, which incorporates necessary-and-sufficient conditions as distinct

levels of an ordinal scale. While binary measures treat all conditions as necessary for establishing membership in a concept's extension, the lexical scale enlists conditions to establish levels of membership in that concept.

This is inspired by the procedure by which John Rawls (1971) orders his three principles of justice: (*A*) the Liberty principle, (*B*) the Fair Equality of Opportunity principle, and (*C*) the Difference principle. These are arranged in order of lexical (short for *lexicographical*) priority. That is, one should not consider *B* or *C* until *A* has been satisfied, nor *C* until both *A* and *B* are satisfied. Thus, each principle serves as a necessary condition of the next, creating a lexical scale with four levels. The force of this argument is conceptual, i.e., that the core meaning of justice is reflected in this particular ordering of attributes, with category *A* understood as the most basic or essential (Moldau 1992).

Of course, Rawls was interested in defining the terms by which institutions within a society could be established and justified. He was not interested in measuring the presence/absence or degrees of justice in a society, and it is not clear whether he would have applied the same rules to such a measurement instrument. Even so, his approach is remarkably similar to what we envision for empirical concepts in the social sciences.

Protocol

A concept-led approach to measurement must take seriously the task of definition, for it is this task that sets the framework for scale construction. In order to make sure that the relevant attributes for Concept *X* are considered, and none arbitrarily excluded, one is well-advised to survey proposed definitions and usage patterns of a concept in ordinary and academic language. [3] This will reveal which attributes are commonly regarded as defining, and which are less common, and perhaps

[3] Examples of this sort of semantic surveying can be found in Collier and Gerring (2009) and Sartori (1984).

idiosyncratic. It will also offer hints as to the suitability of the concept for lexical ordering. If varying definitions of a concept suggest movement up and down a ladder of abstraction – where the addition of attributes to the intension entails a narrowing of the phenomenal extension (Sartori 1970) – there is promise for constructing a lexical scale.

In constructing a lexical scale one must consider "a *theory* of the *ontology* of the phenomenon under consideration" (Goertz 2006: 27). One must decide which attributes are most important (i.e., signifying the intrinsic nature of a phenomenon) and why. This also means that one has to move beyond a purely semantic approach.

Specificially, one must order the attributes identified by the survey of concept definitions and usages so that each attribute serves as a necessary condition for achieving a given (or higher) level on the scale and ordered sets of necessary conditions are jointly sufficient to obtain a particular level. That is, each successive level is comprised of an additional condition, which defines the scale in a cumulative fashion. Condition *A* is necessary and sufficient for *L1*; conditions *A*&*B* are each necessary and jointly sufficient for *L2*; and so forth, as illustrated in Table 1. If there are six levels to an index, five necessary conditions must be satisfied in order to justify a score of 5. This means that each level in a lexical scale is defined by a set of conditions that are both necessary and jointly sufficient. Note that the structure of a lexical scale presupposes that there is a true zero, representing phenomena that do not meet the first condition (~*A*).

Table 1: **Generic Lexical scale**

0.	~*A*				
1.	*A*	~*B*			
2.	*A*	*B*	~*C*		
3.	*A*	*B*	*C*	~*D*	
4.	*A*	*B*	*C*	*D*	~*E*
5.	*A*	*B*	*C*	*D*	*E*

0-5 = levels (*L*) of an ordinal scale. ***A-E*** = conditions that are satisfied. ***~A-E*** = conditions that are not satisfied. Relationships are deterministic except where cells are undefined (empty).

In achieving these desiderata three criteria must be satisfied: (1) binary values for each condition, (2) qualitative differences between levels, and (3) entailment, dependence, or centrality governing the ordering of the levels. We review each criterion in turn.

First, each level in the scale must be measurable in a binary fashion without recourse to arbitrary distinctions. It is either satisfied or it is not. Note that the construction of that binary condition may be the product of a set of necessary and/or sufficient conditions; collectively, however, these conditions must be necessary and sufficient.

Second, each level must demarcate a distinct step or threshold in a concept, not simply a matter of degrees. Levels in a lexical concept identify qualitative differences. *L3* is not simply a way-station between *L2* and *L4*. Indeed, each level may be viewed as a subtype of the larger concept. Note that these subtypes are defined by cumulative combinations of the attributes possessed by the full concept – *A*, *A*&*B*, *A*&*B*&*C*, and so forth – fulfilling the criterion of a classical concept. Note, furthermore, that each of the types associated with the intermediate levels of a lexical scale can be understood as a diminished subtype (Collier and Mahon 1993; Collier and Levitksy 1997) of the type associated with subsequent level. From Table 1 we see that a case scoring a *L2* is not only

characterized by *A*&*B* but also the *absence* of a particular attribute, *C*. In this fashion, the lexical scale is tied to classical concepts and to radial types.

The most challenging aspect of lexical scale construction is the ordering of attributes, which rests on the three considerations.

Logical entailment means that one attribute is logically required for another. For example, elections (*A*) must exist in order for clean elections (*B*) to exist. Thus, one would place Elections prior to Clean elections in an index of electoral democracy (*A*>*B*).

Functional dependence means that the function of one attribute (understood relative to the concept of theoretical interest) is dependent upon another attribute. For example, if the extension of suffrage (*D*) to a social group is to enhance democracy, one might argue that contested elections (*C*) must be in place. Otherwise, suffrage extensions will not enhance the degree of democracy existing in that polity. The social group admitted to suffrage will have no more power in a full-suffrage one-party state than in a partial-suffrage one-party state. (It is irrelevant if women can vote in North Korea.) Thus, one would place Contested elections prior to Suffrage in an index of electoral democracy (*C*>*D*).

Centrality means that one attribute is more central to the core meaning of a concept than another.[4] The centrality criterion thus relates to the theoretical importance of attributes, which can often be ordered in meaningful way. For example, one would place elections (*E*) prior to civil liberties (*F*) in an index of electoral democracy because the former is more central to the concept of democracy (*E*>*F*).

The three criteria used to order attributes in a lexical scale are usually in harmony with one another. That is, (i) logical entailments, (ii) functional dependence, and (iii) conceptual centrality will either suggest the same ordering of attributes, or only one will apply to a given situation. If,

[4] This follows a constitutive approach to measurement, where attributes are the defining elements of a concept (Goertz 2006: 15).

however, there is found to be a potential conflict among these criteria there is an implicit dominance relationship in which (i)>(ii)>(iii). This means that consideration of these three criteria will always result in a unique solution for ordering attributes. Of course, one may disagree over judgments, especially those relating to centrality, as discussed below. But disagreements usually concern the *application* of individual criteria, not conflicts across criteria.

Potential Disagreements

The proposed lexical scale depends upon reaching a determination about (a) the defining attributes of a concept and (b) their lexical ordering. Insofar as this solution is persuasive, the scale will be useful. Insofar as it strains the meaning of a concept or theory it will seem arbitrary and forced, and is on that account unlikely to perform any useful function in social science. A lexical scale must resonate with everyday usage of a word as well as with considered judgments about what a concept should mean in a given theoretical context. We shall not consider disagreements over definition, as conceptual disagreement affects all scaling procedures. Remaining disagreements may be placed into three categories.

First, there are potential disagreements over which attributes associated with a concept's definition should be included or excluded in a lexical index. While this sort of disagreement is potentially damaging we note that disagreements over attribute inclusion are more likely to affect positions on the periphery of the scale (attributes that are sometimes associated with Concept A) than at the core of a scale (attributes that are almost always associated with Concept A). As such, this sort of disagreement is likely to affect index values at the high end of the scale. For example, if two 7-point scales for the same concept differ in the chosen attributes, these differences are most likely to be located at levels 6 and 7 and least likely to be located at levels 1 and 2. As such, only cases with the highest scores, i.e., those whose score is affected by the 6^{th} or 7^{th} conditions, will be

affected. In this sense, measurement error stemming from aggregation is minimized and resulting scores for entities will be fairly similar.

A second sort of disagreement concerns the *number of levels* assigned to a lexical scale. In principle, there is no limit to the number of levels in a lexical scale. In practice, we anticipate that the number of levels is not likely to be very numerous. This is because few concepts have a great many attributes that are truly distinct, independently measurable, and satisfy the criteria required for a lexical scale, as specified above.

In any case, scholars working on the same concept may produce scales of differing lengths. Indeed, the decision about when to aggregate and when to disaggregate attributes (to form conditions) is somewhat arbitrary, hinging on contextual matters such as the sort of data that is at-hand and the use envisioned for the scale. However, disagreements over scale length are not critical, as differently-sized scales will co-vary so long as the identifiable elements are ordered in the same fashion. Compare two hypothetical lexical scales: *I* (*A-B-C*) and *II* (A_1-A_2-B_1-B_2-C_1-C_2), where the latter disaggregates each element of the former into two components. These alternate scales for the same concept are different insofar as one has more levels than the other; but they are not in conflict with each other and they are highly correlated.

A third, more problematic, sort of disagreement concerns the *lexical priority* of different elements. If researchers cannot agree on how to prioritize the attributes of a concept there is little hope of arriving at a useful lexical scale – or, to put the matter differently, each scale will be useful only in a very narrow context and may seem idiosyncratic. This sort of basic-level disagreement is encountered whenever the attributes of a concept bear no logical, functional, or semantic relationship to each other.

While there is no simple solution to this situation, one strategy is to redefine the boundaries of the concept in a narrower fashion so as to exclude elements that cannot easily be integrated. This

may be understood as a shift from a background concept to a *systematized* concept (Adcock and Collier 2001), and causes no damage so long as the re-definition is plausible since it does not strain the meaning of the core concept. Such a redefinition can be communicated by a compound noun that makes clear how the narrower concept relates to the parent concept. Accordingly, we eschew *democracy* in favor of a diminished subtype, *electoral democracy*, in one of the examples discussed below.

III. Examples

As with most methods, it is easier to grasp the approach when specific examples are brought into view. In this section we provide a cursory exploration of possible lexical indices for four well-traveled concepts: *electoral democracy*, *civil liberty*, *party strength*, and *rule of law*.[5] Each section begins with a brief definition, clarifying how we understand the concept. This is followed by a proposed index, followed by a discussion of the principles at work in that index. While the discussion is terse we hope that it fulfills a heuristic function, i.e., showing how the lexical approach to measurement might be applied to a variety of key social science concepts.

Electoral democracy

Electoral democracy refers to the idea that democracy is achieved through competition among leadership groups that vie for the electorate's approval during periodic elections before a broad electorate. This narrow – but nonetheless extremely influential – conception may also be referred to as competitive, elite, minimal, realist, or Schumpeterian (Alvarez et al. 1996; Dahl 1971; Schumpeter

[5] For further examples of concept operationalizations that seem to follow the logic of a lexical scale one might consider *constitutionalism* (Nino 1998: 3-4), *human security* (Tadjbakhsh and Chenoy 2007: Ch.2), *peasants* (Kurtz 2000: 96), and *liberal democracy* (Howard and Roessler 2006; Møller and Skaaning 2013).

1950). To operationalize this concept, we propose the following lexical index (note: elections and institutions refer to *national* elections and institutions):[6]

0. *No elections*. Elections are not held for any policymaking offices. This includes situations in which elections are postponed indefinitely or the constitutional timing of elections is violated in a more than marginal fashion.
1. *Elections with no parties or only one party*. There are regular elections but they are non-partisan or only a single party or party grouping is allowed to participate.
2. *Multi-party elections for legislature*. Opposition parties are allowed to participate in legislative elections and to take office.
3. *Multi-party elections for executive*. The executive is chosen directly or indirectly (by an elected legislature) through elections.
4. *Minimally competitive elections for both executive and legislature*. The chief executive offices and the seats in the effective legislative body are – directly or indirectly – filled by elections characterized by uncertainty, meaning that the elections are, in principle, sufficiently free to enable the opposition to win government power.
5. *Male or female suffrage*. Virtually all adult male *or* female citizens are allowed to vote in elections.
6. *Universal suffrage*. Virtually all adult citizens are allowed to vote in elections.

Recall that in order to achieve a score above 0 on a lexical scale all prior criteria must be met. Thus, in order to achieve a score of 6 on this proposed index of electoral democracy criteria 1-5 must also be satisfied.

This suggested ordering of attributes is sensitive to relationships of logical entailment. Evidently, the existence of elections (*L1*) is a pre-condition for attributes that describe the quality and purview of elections. Likewise, the existence of multi-party elections is a pre-condition for elections that are minimally competitive. In some respects, the ordering of attributes rests on considerations of centrality. We regard legislative elections as more central to the concept of electoral democracy than executive elections. (Accordingly, a country that has elections for the executive only would be regarded as less democratic than a country with elections for the legislature only.) And in some respects, the ordering of attributes rests on considerations of functional

[6] For a more extended treatment of the proposed lexical index of electoral democracy, including various applications, the reader may refer to OMITTED.

dependence. In order for universal suffrage to translate into effective rule by the people multi-party elections must be in place, as discussed. (Accordingly, we regard a country with competitive elections and a narrow suffrage such as Britain in the nineteenth century as more democratic than a country with universal suffrage and non-competitive elections such as North Korea.)

Civil liberty

Civil liberty is a human right, as well as a key component of democracy. Here, we shall understand civil liberty as a property of a government (including parties, civil society groups, and paramilitary groups that are closely associated with that government). Thus, we intend to measure the extent to which governments respect civil liberties *not* the extent to which civil liberties exist in a society (which may depend on things other than government action).[7] Another important caveat is that we are concerned with the actions of government relative to the citizens of a polity. Its actions towards non-citizens (foreign nationals) lie outside the boundaries of our concept. With these qualifications, the following lexical index may be considered:

0. *Political and extra-judicial murder.*
1. *No political and extra-judicial murder.* The government does not organize or condone arbitrary killings or the killing of dissidents or of citizens based on their ascriptive characteristics (e.g., ethnic minorities).
2. *No torture.* The government does not organize or condone the torture of dissidents or of citizens based on their ascriptive characteristics (e.g., ethnic minorities). (Torture is understood as extreme suffering that does not result in death.)
3. *Due process.* The government does not arbitrarily arrest, imprison, or harass its citizens.
4. *Free movement.* The government does not restrict movement and residence within the polity.
5. *Free discussion.* The government does not restrict discussion in private arenas (among family, friends).
6. *Free public speech.* The government does not restrict speech in public arenas including the media.
7. *Free association.* The government does not restrict association, including political parties, labor unions, religious organizations, and other civil society organizations.

[7] Of course, one might argue that it is the responsibility of governments to protect civil liberties, regardless of who is infringing upon them. Nonetheless, it seems important to distinguish between the actions of governmental and non-governmental actors.

The ordering of attributes follows the logic of centrality. Specifically, we suppose that freedom from politically motivated murder is more central to the concept of civil liberty than freedom from torture, freedom from torture is more central than due process, and so forth. Insofar as civil liberty guarantees basic liberties, those liberties that are more fundamental should be granted priority over liberties that are less fundamental. In specific terms, we are claiming that a polity that allows political and extra-judicial murder is in more flagrant violation of the ideal of civil liberty than a polity that allows torture (short of death).

Party strength

Political parties may be defined minimally as organizations that nominate officials for public office, a key function in most theories of representative democracy (Schumpeter 1950). Within this context, the relative strength of these organizations may be regarded as an important component of democracy and good governance (Hicken 2009; Mainwaring and Scully 1995; Ranney 1962; Schattschneider 1942). (The strength of parties within an authoritarian context may also be important, e.g., for regime stability. However, measuring party strength in this context would require a different sort of scale since parties function quite differently in authoritarian contexts.) Party strength is understood here as the mean strength of all political parties that gain entrance into the legislature, and is thus differentiated from party *system* strength (the durability of a set of parties within a polity). With these clarifications, we propose the following lexical index:[8]

0. *Not allowed.* Parties are not allowed to organize.
1. *Allowed.* Parties are allowed to organize. If the system is minimally democratic, the state may restrict entry to small parties judged to be hostile to democratic principles.
2. *Independence.* Parties are independent of the state (e.g., the bureaucracy, the military) and independent of each other (though naturally members of a coalition will be to some extent constrained by coalition agreements).

[8] The index does *not* include a consideration of party nationalization. In our view, parties may be strong while also being rooted in particular regions, as in the United Kingdom.

3. *Defections rare*. Party officials rarely leave their party voluntarily (to join another party or to continue their political career as an independent). Expulsions and retirements are not counted as defections.
4. *Legislative cohesion*. Members of a party usually vote together in the legislature.
5. *Centralization*. Parties do not have strong factions or regional strongholds with distinct organizational structures; important decisions over policy and candidate selection are made at the center, or can be overturned by central party leadership.
6. *Programmatic*. Parties publically embrace policies and ideologies that are relatively distinct.

In prioritizing these attributes we rely, first of all, on logical entailments. The character of political parties cannot be considered unless and until parties are allowed to exist. Second, we enlist functional dependence to prioritize L2 over L3. That is, party defection is an indication of party strength only where parties enjoy independence from government control. If the state controls political parties (generally, it is just one political party) then one cannot defect from one party to another; one can only defect from the regime, which is something else entirely. We rely, third, on considerations of centrality. For example, we regard a party's independence from the state as more central to the concept of party strength than its level of centralization or programmatic orientation. Likewise, defections (members who leave one party and join another) from a party are judged more consequential than cohesion (voting behavior). The programmatic nature of a party is regarded as the least central element of our proposed index. This is consistent with the view that parties need not be ideological in order to be regarded as strong, and ideological parties are not necessarily strong. The concepts of party strength and party ideology should not be conflated. That said, a party possessing all other attributes of party strength that also has a clearly differentiated ideology should be considered stronger than a party whose philosophy and issue-positions are indistinguishable from others in the same party system.

Rule of law

The rule of law is a virtually universal political ideal which has in recent decades been identified as crucial for economic and human development (Tamanaha 2004: 1-4). Among the varying definitions of this concept, most of the attributes may be understood along a continuum of "thin/thick" conceptions (Bedner 2010; Møller and Skaaning 2012; Tamanaha 2004; Trebilcock and Daniels 2008: 12-13). Incorporating these various properties into a single graded scale, we suggest a lexical index as follows:

0. *No rule by law.*
1. *Rule by law***.** Law is used as instrument for government action.
2. *Formal legality***.** Laws are general, clear, prospective, certain, and consistently applied.
3. *Institutional checks***.** An institutionalized system of government characterized by constraints on the executive, including an independent judiciary and penalties for misconduct.
4. *Civil liberties***.** Liberal (negative) rights in the form of physical integrity rights and First Amendment-type rights are safeguarded.
5. *Democratic consent determines laws***.** The citizens, through their elected representatives, are the ultimate source of laws.

In arriving at an ordering of these attributes we are once again cognizant of logical entailments. Attributes 2-4 are impossible to implement if law is not a principal instrument of governmental action. Institutional checks are generally inoperable without a system of formal legality. Civil liberties cannot be instituted unless there is a system of formal legality and institutional checks, including an independent judiciary. The final condition, democratic consent, follows the criterion of centrality. That is, the fullest, most complete realization of rule of law cannot be achieved if the citizens of a state are not sovereign (Habermas 1996). Any time an unelected individual or group of individuals are capable of altering the law in fundamental ways without recourse to democratic approbation the rule-of-law ideal is violated. Likewise, in a situation where this is possible the structure of law is, by definition, ad hoc and unpredictable since it is privy to the whims of whoever happens to be serving as head of state. However, democratic consent is judged

less central to the concept than other elements of our proposed index, which accounts for its position at the outer periphery.

Summary

In this brief and necessarily schematic discussion we hope to have demonstrated the potential applicability of a lexical approach to measurement for a broad range of social science concepts. Granted, the lexical approach may not be practicable for all concepts. And where it applies, the lexical scale constitutes just one of many possible approaches to measurement. Choices among available options are likely to hinge on the uses to which the concept is being put in a particular research context, as discussed below. We turn now to a discussion of how the lexical scale compares and contrasts with other measurement strategies.

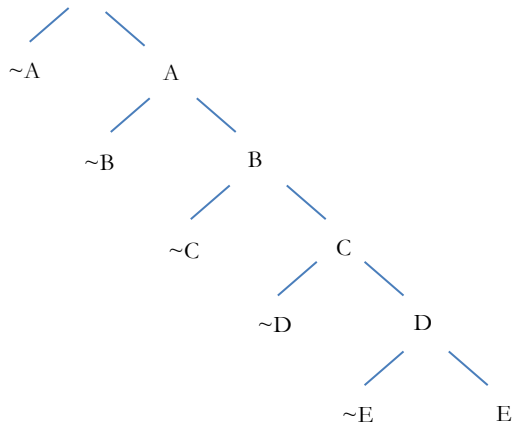
IV. Situating the Lexical Scale

Within the literature on concept formation, the lexical scale may be viewed as an attempt to reconcile *minimal* (thin) and *maximal* (thick, ideal-type) strategies (Coppedge 1999; Gerring 2012: Ch. 5). Note that the first condition (or first several conditions) in a lexical index establishes a minimal core of a concept while the last condition in the scale completes what might be viewed as an ideal-type concept. Granted, ideal-type definitions are often more expansive than those envisioned by the lexical scale, primarily because the requirements of an ideal-type are less restrictive (anything that coheres with the concept is admissible). Even so, the lexical scale serves as a bridge between minimal and maximal concepts.

The lexical scale is also closely linked to a long intellectual tradition focused on typologies and taxonomies referred to above. Note that each level of a lexical scale corresponds to a distinctive category or type and is defined by all attributes contained in the superordinate category *plus one*.

Accordingly, the levels of a lexical scale may be represented in a tree diagram, as depicted in Figure 1. By contrast, a lexical approach to measurement is *not* suitable if the attributes of a concept follow a family-resemblance logic, where no features are necessary to its definition (Wittgenstein 1953; Goertz 2006).

Figure 1: **Lexical Scale in Tree-Diagram (Taxonomic) Format**



Within the literature on measurement, the lexical scale is similar in structure to Guttman scaling (Coppedge and Reinicke 1990; Guttman 1950). Guttman scales are based on a series of observable binary attributes—typically survey items—that, for each case, are rank-ordered with respect to some unobservable latent trait. For example, one might measure the latent trait of mathematical ability with a series of test questions of increasing difficulty, observing binary correct/incorrect responses. In its idealized form, observing a positive value for any item on a Guttman scale implies that one must also observe positive values for all lower ranked, or less "difficult," items. For example, one might assume that students learn to count before they tackle

addition, and learn to add before they master multiplication. Thus, one could consider a test containing three questions—testing counting, addition, and multiplication, respectively—as a Guttman scale for a subset of mathematical ability (Abdi 2010). Students who correctly answer the multiplication question would, by assumption, also provide correct answers to the other two questions. The original Guttman model posits a deterministic—rather than probabilistic—relationship between latent traits and observables, and idealized Guttman scales are built using purely deductive reasoning. In the preceding example, one deduces that multiplication is more difficult than addition, which is more difficult than counting. Moreover, one assumes a systematic relationship between (in)correct question answers and latent mathematical ability. The lexical scale is also deterministic and concept-driven, and both scales are cumulative and unidimensional. Nonetheless, the cumulative relationships modeled by lexical and Guttman scales differ fundamentally.

To clarify this distinction, return to Table 1 and consider a case for which conditions *A*, *C*, *D*, and *E*—but not *B*—are met. Such a case would score as a "1" on the generic lexical scale depicted in Table 1 because *B* is a necessary condition for classification at ordinal level 2 and above. But knowing that a case scores 1 on the generic lexical scale in Table 1 tells us only that *A* is satisfied and *B* is not, providing no clue about the values of conditions *C*-*E*. By contrast, when working with a deterministic Guttman scale, knowing the Guttman score for a case provides perfect information about which observable conditions that case satisfies. A Guttman scale for this example would posit a strict ordering of conditions A-E, as displayed in Table 2. Thus, observing E, would tell us that A-D must be satisfied, under the strict Guttman assumption. Indeed, it should be impossible to observe the above-described case if the latent concept in question actually follows a deterministic Guttman scale.

To clarify this point, Table 2 displays a hypothetical Guttman scale that mirrors Table 1 in terms of conditions (items) and scale levels. Note that, in contrast to Table 1, Table 2 has no missing cells, or undefined relationships between scale level and observable traits. Thus, a case scoring 1 on this generic scale must satisfy condition *A*, but not conditions *B*-*E*, a case scoring 2 would satisfy *A* and *B* but not *C*-*E*, and so on. Similarly, knowing that a case meets a particular condition on a Guttman scale allows one to infer that the case likewise meets all of the "easier" conditions belonging to the scale. For example, if we know that a case meets condition *D* on the generic scale in Table 2, we know that it must also meet conditions *A*-*C*. Guttman scales are a form of generative model that predicts observable characteristics based on an unobservable trait. For instance, a person's unobservable level of mathematical reasoning ability predicts their success on an exam. The assumptions inherent in a Guttman scale are testable. In its idealized, deterministic, form, one can falsify a Guttman scale simply by finding a contradiction. For example, referring again to Table 2, a case sporting condition *D* but not *B* would contradict the assumptions in our generic Guttman scale.

Table 2: **Generic Guttman Scale**

0.	~*A*	~*B*	~*C*	~*D*	~*E*
1.	*A*	~*B*	~*C*	~*D*	~*E*
2.	*A*	*B*	~*C*	~*D*	~*E*
3.	*A*	*B*	*C*	~*D*	~*E*
4.	*A*	*B*	*C*	*D*	~*E*
5.	*A*	*B*	*C*	*D*	*E*

0-5 = ordinal scale. ***A-E*** = conditions that are satisfied. ***~A-E*** = conditions that are not satisfied. Relationships are probabilistic.

Because real-world conditions rarely match the strict conditions of an idealized Guttman scale, applied work typically adapts Guttman scaling in a probabilistic manner. Modern applications of Guttman scaling thus fall under the broad umbrella of IRT (see Fox 2010 for an overview). These

techniques relax the strict determinism inherent in idealized Guttman scales, allowing researchers to estimate latent traits based on a probabilistic understanding of the relationships between latent variables and observed characteristics. In terms of the generic scale in Table 2, for example, knowing that a case meets condition *D* might tell us that it is likely—but not certain—to meet conditions *A*-*C*. In general, whenever one assumes that the Guttman assumptions hold in a probabilistic fashion, one can derive probability distributions for observable attributes as a function of latent traits. Given that certain assumptions are met, one can use these distributions to estimate latent traits from observable attributes, even when the empirical distribution of those attributes does not perfectly conform to the ordering of a pure Guttman scale.

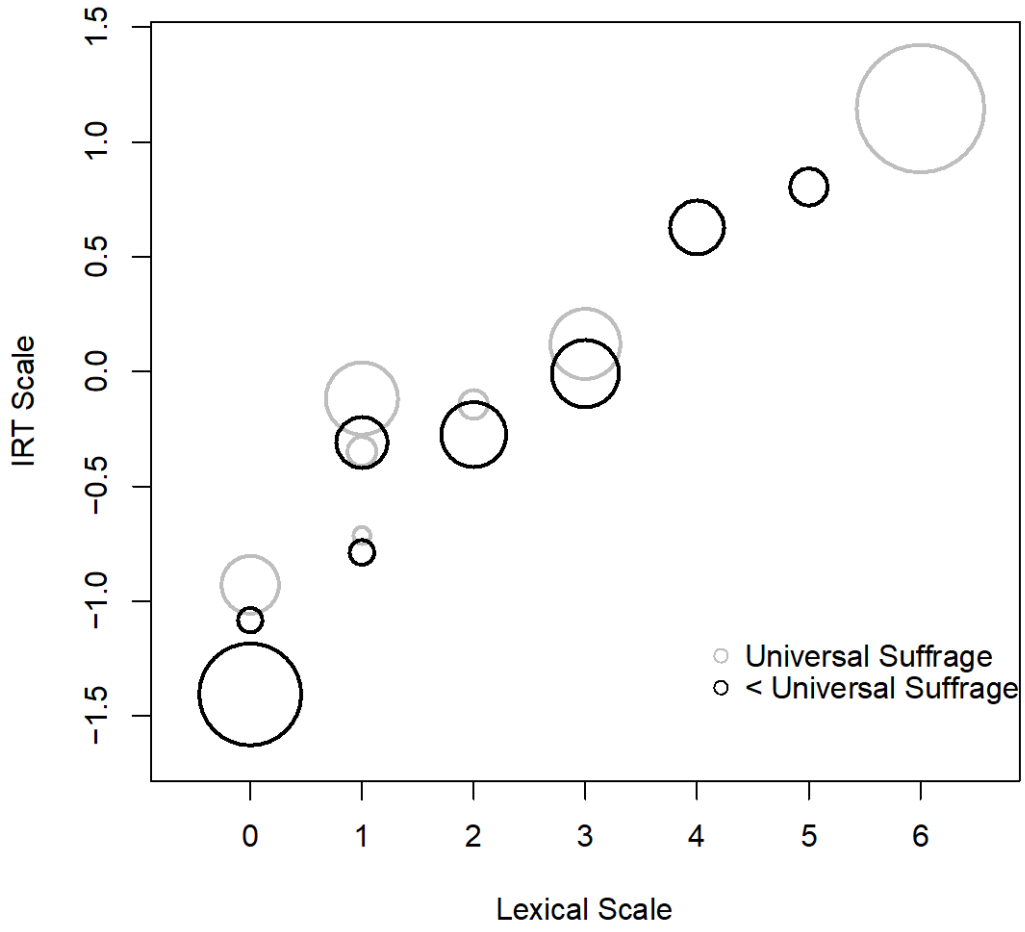
When the relationship between a concept and observable attributes matches Guttman assumptions, at least probabilistically, it makes sense to model that concept with a (probabilistic) Guttman scale. Indeed, because Guttman/IRT models generate predictions about patterns of covariation between observable attributes, such models allow one to test the plausibility of core scaling assumptions, providing a powerful tool for modeling cumulative concepts and justifying scaling decisions. Nonetheless, many concepts fall outside the scope of Guttman models.

While all Guttman scales are lexical scales, not all lexical scales satisfy Guttman assumptions. None of our examples in section III are Guttman scales, although civil liberty would nominally constitute a Guttman scale if we excluded the sixth level, because levels 1-5 are ranked purely based on logical entailments, rather than functional dependence or centrality. Indeed, lexical scales are also Guttman scales only when functional dependence and centrality play no role in scale construction. Perhaps more fundamentally, lexical scales are not generative, and posit no relationship between an unobserved latent trait and a set of observable variables. Rather, they are deductive classification rules for observable traits. Thus, lexical scales are appropriate for a class of concepts that do not fit into the Guttman framework.

First, rankings in lexical scales are based not just on logical ordering requirements but also considerations of functional dependence and conceptual centrality. Returning to our first exemplar, electoral democracy, let us consider the items "Are there competitive elections?" and "Is suffrage universal?" It is not obvious how one would rank these two traits in terms of logical requirements or "difficulty." One can have universal suffrage without competitive elections, and vice versa, thus these two conditions cannot both appear on a single Guttman scale. While one might consider competitive elections more *central* to the concept of democracy than universal suffrage, or argue that the meaning and importance of universal suffrage *functionally depends* on the existence of competitive elections, these a priori theoretical classifications do not necessarily correspond to empirical patterns in a way that would fit a (probabilistic) Guttman model. In particular, observing that a country has universal suffrage tells us little about whether or not it has competitive elections.

Second, Guttman-based approaches are only appropriate when it is possible to identify "outcome" indicators that measure a latent concept of interest. For example, subjects' performance on a test offers a good indication of their aptitude – if not of their overall intelligence at least of their knowledge in a subject area. However, with many concepts an outcome-based approach is difficult to apply. Often, we think of observable case characteristics not as the empirical implications of a continuous latent trait—such as intelligence—but as *constitutive* attributes, ones that define the concept of interest. Here, lexical scales allow one to describe concepts that are not a good fit for latent variable modeling techniques, including those built upon Guttman assumptions.

Figure 2: **Lexical vs IRT Scales of Electoral Democracy**



The sample represents country-years, covering most countries in the world from 1800 onward. See [citation removed] for a full-fledged description of the lexical scale of electoral democracy; we use version 3 of the dataset. The MCMCirt1d function in the MCMC pack R library (Martin, Quinn and Park 2011) is used to fit the model. Observations are grouped along the IRT (vertical) axis by identifying gaps between clusters of observations. Each group is represented by a bubble centered at the group mean, with bubble sizes proportional to the number of cases within the group. We chose a gap size of roughly 0.1 to cleanly separate clusters of observations. Sets of cases with universal suffrage are depicted in gray; cases with less than universal suffrage in black.

Figure 2 illustrates the distinction between lexical and (probabilistic) Guttman scales by plotting a lexical scale of electoral democracy against latent trait estimates generated by fitting a Bayesian IRT model to the same data. While the two measures correlate highly (r=0.96), they disagree fundamentally on the rank ordering of numerous cases. In particular, universal suffrage

does not follow the Guttman assumption, occurring in cases lacking all other characteristics of electoral democracy.

The IRT model, following probabilistic Guttman assumptions, treats suffrage as consistently beneficial across the scale (note that cases with universal suffrage are plotted in gray in Figure 2). While this assumption makes sense at the top of the spectrum, one might argue that suffrage is meaningless, both practically and theoretically, towards the bottom of the scale. That is, in the absence of multi-party competition it hardly matters what portion of the electorate is allowed to participate in elections. Because patterns of suffrage fit patterns across other indicators poorly the IRT model considers suffrage a poorly discriminating indicator of electoral democracy, placing less weight on this aspect of electoral democracy. Yet, because the IRT model gives *some* weight to suffrage when estimating the latent trait, we see substantial variation in IRT scores within lexical categories, contra general understandings of democracy. Roughly five per cent of the cases in the dataset receive the lowest possible score on the lexical scale (because national elections do not exist or have been suspended) and a score that falls above the lowest level on the IRT scale (because the institution of universal suffrage is established by statute or by precedents set in previous elections).

For example, the lexical scale accords China a higher score in the republican era (1912-1949) than in the communist era (1949-), while the IRT scale reverses these scores. We propose that the lexical index, derived purely from "deductive" theorizing, offers a more compelling operationalization of the core concept of democracy than the IRT model, in which scaling is dependent upon the distribution of attributes across the cases. China fools the IRT model, but not the lexical index.

We find similar sorts of variation in IRT scores among cases in categories 1-3 of the lexical index which are difficult to justify from a theoretical perspective. Consider the large gray bubble situated at 1 on the lexical index, but near zero on the IRT model in Figure 2. These polities exhibit

universal suffrage but no-party or single-party elections, and include cases such as Cambodia under the Khmer Rouge. The IRT model nonetheless ranks them similarly to cases with multiparty elections for legislative and executive posts, those cases in the large black bubble at the lexical score of three in Figure 2. Note that because the IRT approach weighs suffrage equally regardless of the configuration of other traits it sees little distinction between the level of electoral democracy under the Khmer Rouge and that in Britain before the first Reform Act in 1932.

More broadly, while there is a clear separation on the IRT scale between cases that exhibit minimally competitive elections and those that do not, cases falling between one and three on the lexical scale are generally muddled on the IRT dimension. As we have highlighted, this muddle is largely an artifact of applying the Guttman assumption to suffrage, which does not align properly (i.e., in accordance with theoretical suppositions about democracy) with other indicators of the concept. This exercise demonstrates the difficulty of applying probabilistic Guttman scaling—or any scaling method that relies on empirical covariance between observable characteristics—to concepts like electoral democracy where constituent traits cannot be reasonably modeled as observable implications of latent states.

Finally, it is important to emphasize that the validity of a lexical scale is independent of the empirical distribution of the attributes that make up that scale. This point stems from the fact that lexical scales need not imply any deterministic or probabilistic relationship between observable conditions. This distinction between lexical and Guttman scales has two key implications.

First, one can use Guttman logic to inductively measure concepts. Researchers often build Guttman scales inductively. In particular, after identifying a set of observable indicators relevant to a concept—say correct answers to math questions on a test—one can use statistical methods to inductively learn the ordering of the observed variables with respect to the latent trait, conditional on the assumption that such an ordering follows a (probabilistic) Guttman process. This process

mixes deductive and inductive reasoning. One deducts the relevant observable attributes and assumes that they follow some Guttman ordering, while learning that ordering inductively. This is exactly what we did when we fit the IRT model in the preceeding example. Lexical scales built purely from logical entailments are identical to Guttman scales. But once one one considers functional dependence or conceptual centrality, lexical scales no longer strictly order attributes, and it becomes impossible to observe more "difficult" conditions in the absence of "easier" ones. Such lexical scales imply no specific observed correlation between such conditions, making inductive scale construction impossible.

Second, and relatedly, the strict relationship between observable indicators assumed by the Guttman scale means that such scales are testable. In particular, if one deduces the ordering of a Guttman scale a priori, one can use IRT or similar methods to test the extent to which the data follow the deduced ordering. Since less "difficult" Guttman conditions should (probabilistically) be met before more "difficult" Guttman conditions, one can determine the extent to which real world observations match scale assumptions. The same is not true of the lexical scale, once functional dependence and conceptual centrality enter the picture, although one could test orderings implied by logical entailment.[9] Furthermore, our example shows that IRT and similar tools can be useful for testing the necessity of lexical scaling methods. Suffrage is clearly an important part of the concept of electoral democracy, but does not follow a Guttman data generating process. Lexical scaling provides a reasoned tool for including such relevant concepts into a measure when the simpler Guttman approach is inconsistent with theory.

On one hand, these differences between Guttman and lexical scales may be regarded as weaknesses of the lexical scale, since researchers can neither infer lexical scales from patterns in the data nor justify their many of their conceptual choices by appealing to consistency with the empirical

[9] Logical entailment is deterministic and should always hold in a well-designed lexical scale.

record. On the other hand, our examples demonstrate that inductively constructed measures of concepts may be invalid when the empirical distribution of attributes does not fit intuitive or theoretical meanings of those concepts. In particular, when one has strong theoretical reasons to conceptualize and measure ordinal constructs based on considerations of functional dependence and conceptual centrality it is inappropriate to use Guttman scales, or related tools like IRT, to generate measures of those constructs. Fundamentally, lexical scales are not generative; they do not predict data based on latent traits. Rather, they describe a particular class of cumulative concept that is logically defined by a specific aggregation of sub-components, as set forth in section II.

V. Discussion

We shall now attempt to summarize our wide-ranging discussion pertaining to the strengths and weaknesses of the lexical scale. In principle, the recalcitrant aggregation problem is solved by treating defining attributes as necessary-and-sufficient conditions arrayed in an ordinal fashion.[10] If the scale is true to its objectives, each level in the scale defines a stronger, more complete instantiation of the underlying concept. This is no mean feat, given that composite indices are often plagued by problems of aggregation (Goertz 2006; Munck 2009).

Because conceptualization is integrated into measurement there should, in principle, be less slippage between concept and indicator than is typically encountered with other methods of scale construction.[11] Nonetheless, if the analyst drops attributes of a concept from an index (because they

[10] This presumes, of course, that each condition can be accurately measured in a binary fashion without too much loss of information.

[11] Authors' choices of indicators to include in an index are often somewhat arbitrary (Goertz 2006; Haig and Borsboom 2008; Munck 2009). For example, the Worldwide Governance Indicator for "rule of law" (Kaufmann et al. 2010) primarily measures crime and property rights, downplaying or entirely excluding other attributes of the concept (Skaaning 2010). Likewise, the Freedom House Political Rights and Civil Liberties indices are based on indicators that have changed over time and some of them pertain to corruption, civilian control of the police, the absence of widespread violent crime, willingness to grant political asylum, the right to buy and sell land, and the distribution of state enterprise profits (Freedom House 2015). Some observers might regard these features as elements of political rights and civil liberties; others might not. Since most abstract concepts can be defined in a variety of ways and do not possess

cannot be meaningfully arrayed in an ordinal scale), forces continuous phenomena into an arbitrary binary coding, or prioritizes conditions without some underlying rationale, the resulting index will depart from ordinary meanings implied by the concept. The lexical scale is not immune to problems of concept-measure consistency.

Likewise, the strictures of the lexical scale are not universally applicable. They require that relevant attributes of a concept be coded in a binary fashion without undue distortion and that the chosen attributes be arrayed along a single dimension according to their centrality to the concept or relations of dependence. These are not easy requirements to satisfy.

We have also noted that the deductive properties of a lexical scale require many judgments on the part of the analyst. Accordingly, different analysts may arrive at different scales for the same concept. This, by itself, does not differentiate the lexical scale from other scales, including those constructed in a more inductive fashion. After all, there are many moving parts to any scale, particularly when one is attempting to operationalize a highly abstract concept. One must choose an indicator or a set of indicators to represent a concept and, if more than one indicator is chosen, one must decide upon an aggregation technique(s) that combines those elements into a single scale. Accordingly, it is not the case that lexical scales are more "subjective" than other scales.

Arguably, the assumptions employed in the construction of a lexical scale are more transparent than the assumptions used to construct many other composite indices, especially when a number of aggregation principles are embedded in a complex statistical model.[12] Then again, because of the set-theoretic nature of the lexical scale it seems likely that alternative lexical scales for the same concept will be less highly correlated than varying inductive scales for the same concept. A

sharp boundaries, it is no surprise to discover that one analyst's bundle of indicators may be quite different from another's, even when they purport to operationalize the same term.

[12] Of course, the use of "inductive" approaches need not imply that the analyst is doing theory-free measurement. Indeed, the best applications of these approaches deploy models explicitly because theory implies that these models are appropriate, or because these models allow the analyst to test theoretical assumptions using the empirical record. Nonetheless, in practice, many indices aggregated through latent variable modeling approaches are under-theorized.

small change in an ordinal scale generally has greater consequences than a small change in an interval scale.

Lexical scale construction is a highly deductive enterprise insofar as the resulting index is constructed to suit a priori requirements drawn from the concept rather than from the empirical distribution of the data. Yet, many concepts do not provide clear guidance with respect to the relative priority of their defining conditions, a limiting condition on the applicability of lexical scaling.

Where applicable, however, the deductive properties of the scaling procedure offer certain advantages. Note that insofar as the distribution of data is allowed to influence the construction of an index, the resulting variable is sample-dependent. If key properties of a sample change (e.g., when drawn from different populations or when drawn non-randomly from a single population), so does the resulting scale. In most circumstances (and especially where the population extends into the future), it is not possible to determine what the shape of a larger population looks like. In these situations, indices are biased – or, alternatively stated, they lack generalizability because they are sample-dependent. An IRT-based index of electoral democracy, for example, is likely to vary across sample periods for the very reason that the composition of regimes around the world (the basis for the index) has varied enormously over the past two centuries. An IRT-based index constructed with data for the nineteenth century will be different from an IRT-based index constructed on the basis of data for the twentieth century. Indeed, every time a new decade of data is added to a sample (assuming the sample is updated regularly), the resulting index of democracy could change. This sort of instability can be problematic.

Relatedly, a basic (and nearly universal) operating assumption of inductive index construction is that one can combine information from observed variables by paying attention to their commonalities and discarding their differences as error. This is a reasonable set of assumptions

in many circumstances, especially when the commonalities are great and the remaining differences do not seem to represent anything of substantive significance, i.e., they do not compose an identifiable dimension. However, it involves a considerable simplification of reality, especially when co-variation is modest. In such circumstances, the lexical scale offers a viable alternative.

With respect to discrimination, the lexical scale may be counted as modestly successful. It provides much more information than the classical concept, understood as a binary scale. It is on par with many ordinal indices, which generally incorporate a handful of levels. It is also on par with indices that purport to be interval scales but, in reality, are probably better understood as ordinal such as the Polity and Freedom House indices of democracy (Armstrong 2011; Cheibub, Gandhi and Vreeland 2010; Pemstein, Meserve and Melton 2010; Treier and Jackman 2008). To be sure, a lexical scale will discriminate less successfully than a scale whose construction is geared to detect small differences (e.g., IRT models).

While sensitivity to small differences is valuable, it is not the only factor of importance in constructing a scale. Note that some concepts in the social science universe are probably lumpy rather than continuous. This appears to be the case with electoral democracy. One is at pains to describe the difference between a regime with popular elections and one without (the first condition of our proposed Lexical index) as a matter of degrees. The same point might be made with reference to the other examples discussed above.

Likewise, where a concept is being formulated as a right-side variable in a causal model it may be helpful to recognize distinct treatments, understood as a cumulative series of compound treatments – *A*, *A*&*B*, *A*&*B*&*C*, et al. These can be tested with (a) pairwise comparisons and matching algorithms, (b) dummy variables in a regression model, (c) generalized additive models (Beck and Jackman 1998), or (d) Bayesian shrinkage models (Alvarez et al. 2011). If used to achieve

covariate balance in a matching analysis a categorical variable is generally more tractable than a continuous variable. In these respects, lexical scales are well-suited for causal inference.

By contrast, inductively derived indices often function awkwardly on the right side of a causal model. A useful treatment is uniform, imposing the same condition on all those within the treatment group. However, indices usually include heterogeneous elements – a little bit of this and little bit of that, in portions that are difficult to account for. Typically, there are many ways to obtain a score of "3" along a continuous scale.[13] Consequently, it is difficult to say what the treatment consists of, what causal mechanisms might be at work, and whether the resulting relationship should be interpreted as causal.

Composite scales generally indicate differences of degree, but not of kind. A "4" on the Polity2 scale indicates that a regime is more democratic than a country receiving a "2." But it offers no additional information about the qualities of these regimes. In this respect, the information contained in a standard composite index is "quantitative" (more/less) rather than "qualitative" (differences of type). Accordingly, a point on a composite index rarely has an obvious interpretation or meaning except in terms of standard deviations from the mean, and thresholds used to convert a continuous scale into a nominal or ordinal scale are apt to be highly arbitrary. This makes it difficult to evaluate concept validity, even if aggregation rules are perfectly transparent. And it makes it difficult to apply concepts to real-world situations, detracting from social science's relevance to politics and policy.

[13] With respect to the Freedom House indices, Cheibub, Gandhi and Vreeland (2010: 75) note: "for each of the ten categories in the political rights checklist and the 15 categories of the civil liberties checklist, coders assign ratings from zero to four and the points are added so that a country can obtain a maximum score of 40 in political rights and 60 in civil rights. With five alternatives for each of ten and 15 categories, there are 510 = 9,765,625 possible ways to obtain a sum of scores between zero and 40 in political rights, and 515 = 30,517,578,125 possible ways to obtain a sum of scores between zero and 60 in civil liberties. All of these possible combinations are then distilled into the two seven-point scales of political rights and civil liberties."

By contrast, a lexical scale is relatively transparent. Researchers and reviewers know exactly what a shift from "2" to "3" or "3" to "4" means because each level in the scale is achieved by only one additional criterion. This eases the burden of ex ante coding and ex post interpretation. Likewise, insofar as levels correspond to distinctive types, membership in each category of an ordinal scale is meaningful. Units coded as "3" share various characteristics, which may signal important theoretical properties (e.g., as inputs or outputs of a causal model). Qualitative differences are sometimes more informative than quantitative differences.

We are not proposing that Lexical indices are superior to other sorts of indices, each of which represent certain aspects of reality and each of which has its uses. Sometimes, relationships are continuous (and hence best measured with an interval scale) and sometimes they have only one threshold (and hence best measured with a binary scale). By the same token, sometimes causal relationships are ordinal in character, or they require an ordinal scale to test various threshold possibilities. In these settings, which surely apply to many theories, a lexical scale – where ordinal levels represent qualitatively different categories – may be appropriate. In this fashion, we propose to add another tool to the measurer's toolkit.

VI. References

Abdi, Herve. 2010. "Guttman Scaling." Pp. 558-60 in Encyclopedia of Research Design, edited by N. J. Salkind, D. M. Dougherty, and B. Frey. Thousand Oaks, CA: Sage.

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

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

Alvarez, R. Michael, Delia Bailey, and Jonathan N. Katz. 2011. "An Empirical Bayes Approach to Estimating Ordinal Treatment Effects." *Political Analysis* 19:1, 20-31.

Armstrong, David A. II. 2011. "Stability and Change in the Freedom House Political Rights and Civil Liberties Measures." *Journal of Peace Research* 48(5): 653–62.

Bailey, Kenneth D. 1994. *Typologies and Taxonomies: An Introduction to Classification Techniques*. Thousand Oaks: Sage.

Beck, Nathaniel, Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42(2):596–627.

Bedner, Adriaane. 2010. "An Elementary Approach to the Rule of Law." *Hauge Journal on the Rule of Law* 2(1): 48-74.

Brady, Henry E., David Collier (eds). 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards, 2d ed.* Lanham: Rowman & Littlefield.

Bühlmann, Marc, Wolfgang Merkel, Lisa Müller, Bernhard Weßels. 2011. "The Democracy Barometer: A New Instrument to Measure the Quality of Democracy and Its Potential for Comparative Research." *European Political Science* 11(4): 519-36.

Calhoun, Craig. 2002. *Dictionary of the Social Sciences*. Oxford: Oxford University Press.

Cheibub, Jose Antonio, Jennifer Gandhi, and James Raymond Vreeland. 2010. "Democracy and Dictatorship Revisited." *Public Choice* 143(1–2): 67-101.

Cingranelli, David L. and David L. Richards. 1999. "Measuring the Level, Pattern, and Sequence of Government Respect for Physical Integrity Rights." *International Studies Quarterly* 43(2): 407-17.

Collier, David and James Mahon. 1993. "Conceptual 'Stretching' Revisited: Adapting Categories in Comparative Analysis." *American Political Science Review* 87(4): 845-65.

Collier, David, and Steven Levitsky. 1997. "Democracy with Adjectives: Conceptual Innovation in Comparative Research." *World Politics* 49(3): 430-51.

Collier, David, Robert Adcock. 1999. "Democracy and Dichotomies: A Pragmatic Approach to Choices about Concepts." *Annual Review of Political Science* 2: 537-65.

Collier, David, Jody LaPorte, and Jason Seawright. 2012. "Putting Typologies to Work: Levels of Measurement, Concept Formation, and Analytic Rigor." *Political Research Quarterly* 65(2): 217-32.

Collier, David, Fernando Daniel Hidalgo, and Andra Olivia Maciuceanu. 2006. "Essentially Contested Concepts: Debates and Applications." *Journal of Political Ideologies* 11(3): 211-46.

Collier, David and John Gerring (eds). 2009. *Concepts and Method in Social Science: The Tradition of Giovanni Sartori*. London: Routledge.

Coppedge, Michael. 1999. "Thickening Thin Concepts and Theories: Combining Large N and Small in Comparative Politics." *Comparative Politics* 31(4): 465-76.

Coppedge, Michael., Angel Alvarez, and Claudia Maldonado. 2008. "Two Persistent Dimensions of Democracy: Contestation and Inclusiveness." *Journal of Politics* 70(3): 335–50.

Coppedge, Michael and Wolfgang H. Reinicke. 1990. "Measuring Polyarchy." *Studies in Comparative International Development* 25(1): 51-72.

Dahl, Robert A. 1971. *Polyarchy: Participation and Opposition*. New Haven: Yale University Press.

DeVellis, Robert F. 2011. *Scale Development: Theory and Applications*. Thousand Oaks, CA: Sage.

Elkins, Zachary. 2000. "Gradations of Democracy? Empirical Tests of Alternative Conceptualizations." *American Journal of Political Science* 44(2): 287-94.

Elman, Colin. 2005. "Explanatory Typologies in Qualitative Studies of International Politics." *International Organization* 59(2): 293-326.

Fox, Jean-Paul. 2010. *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer.

Freedom House. 2015. http://www.freedomhouse.org, accessed April 5, 2018.

Gerring, John. 1999. "What Makes a Concept Good? An Integrated Framework for Understanding Concept Formation in the Social Sciences," *Polity* 31(3): 357-93.

Gerring, John. 2012. *Social Science Methodology: A Unified Framework, 2d ed.* New York: Cambridge University Press.

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

Goertz, Gary and James Mahoney. 2012. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton: Princeton University Press.

Guttman, Louis. 1950. "The Basis for Scalogram Analysis." In *Measurement and Prediction: Studies in Social Psychology in World War II, Vol. 4*, edited by Samuel A. Stouffer, Louis Guttman, Edward Suchman, Paul Lazarsfeld, Shirley Star, and John Clausen. Princeton, NJ: Princeton University Press.

Habermas, Jürgen. 1996. *Between Facts and Norms: Contributions to a Discourse Theory on Law and Democracy*. Cambridge, MA: MIT Press.

Hadenius, Axel and Jan Teorell. 2005. *Assessing Alternative Indices of Democracy*. Committee on Concepts and Methods Working Paper Series, August.

Haig, Brian D. and Denny Borsboom. 2008. "On the Conceptual Foundations of Psychological Measurement." *Measurement* 6, 1-6.

Hicken, Allen D. 2009. *Building Party Systems in Developing Democracies*. New York: Cambridge University Press.

Howard, Marc Morjé and Philip G. Roessler. 2006. "Liberalizing Electoral Outcomes in Competitive Authoritarian Regimes." *American Journal of Political Science* 50(2): 362-78.

Kaufmann, Daniel, Aart Kraay, and Massimo Mastruzzi. 2010. *The Worldwide Governance Indicators: Methodology and Analytical Issues*. World Bank Policy Research Working Paper No. 5430.

Kuper, Adam, Jessica Kuper. 1996. *The Social Science Encyclopedia, 2d ed.* New York: Routledge.

Kurtz, Marcus. 2000. "Understanding Peasant Revolution: From Concept to Theory and Case." *Theory and Society* 29(1): 93-124.

Lazarsfeld, Paul F. 1937. "Some Remarks on the Typological Procedures in Social Research." *Zeitschrift für Sozialforschung* 6: 119-39.

Lazarsfeld, Paul F. and Allen H. Barton. 1951. "Qualitative Measurement in the Social Sciences: Classification, Typologies, and Indices." Pp. 155-92 in *The Policy Sciences*, edited by Daniel Lerner and Harold D. Lasswell. Stanford: Stanford University Press.

Lazarsfeld, Paul F. and Neil W. Henry. 1968. *Latent Structure Analysis*. Boston: Houghton Mifflin.

Mainwaring, Scott and Timothy Scully (eds). 1995. *Building Democratic Institutions: Party Systems in Latin America*. Stanford: Stanford University Press.

Marshall, Monty G., Ted Gurr, and Keith Jaggers. 2016. *Polity IV Project: Political Regime Characteristics and Transitions, 1800–2016*. http://www.systemicpeace.org/inscr/p4manualv2016.pdf, accessed April 5, 2018.

Martin, Andrew D., Kevin M. Quinn, and Jong Hee Park. 2011. "MCMCpack: Markov Chain Monte Carlo in R." *Journal of Statistical Software* 42(9).

Miller, Michael. 2015. "Democratic Pieces: Autocratic Elections and Democratic Development since 1815." *British Journal of Political Science* 45(3): 501-30.

Moldau, Juan Hersztajn. 1992. "On the Lexical Ordering of Social States According To Rawls' Principles of Justice." *Economics and Philosophy* 8: 141-48.

Møller, Jørgen and Svend-Erik Skaaning. 2012. "Systematizing Thin and Thick Conceptions of the Rule of Law." *Justice Systems Journal* 33(2): 136-153

Møller, Jørgen and Svend-Erik Skaaning. 2013. "Regime Types and Democratic Sequencing." *Journal of Democrac*, 24(1): 142-156.

Munck, Gerardo L. 2009. *Measuring Democracy: A Bridge between Scholarship and Politics*. Baltimore: John Hopkins University Press.

Nino, Carlos Santiago. 1998. *The Constitution of Deliberative Democracy*. New Haven: Yale University Press.

Pemstein, Daniel, Stephen Meserve, and James Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18(4): 426-49.

Przeworski, Adam, Michael Alvarez, Jose Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Material Well-Being in the World, 1950-1990*. New York: Cambridge University Press.

Ranney, Austin. 1962. *The Doctrine of Responsible Party Government: Its Origins and Present State*. Urbana: University of Illinois Press.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.

Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64(4): 1033-46.

Sartori, Giovanni. 1984. "Guidelines for Concept Analysis." Pp. 15-48 in *Social Science Concepts: A Systematic Analysis*, edited by Giovanni Sartori. Beverly Hills: Sage.

Saylor, Ryan. 2013. "Concepts, Measures, and Measuring Well: An Alternative Outlook." *Sociological Methods Research* 42(3): 354-91.

Schattschneider, E. E. 1942. *Party Government*. New York: Rinehart.

Schneider, Carsten Q. and Claudius Wagemann. 2012. *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. New York: Cambridge University Press.

Schumpeter, Joseph A. 1950. *Capitalism, Socialism and Democracy*. New York: Harper & Bros.

Seawright, Jason, David Collier. 2014. "Rival Strategies of Validation: Tools for Evaluating Measures of Democracy." *Comparative Political Studies* 47(1): 111-38.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Skaaning, Svend-Erik. 2010. "Measuring the Rule of Law." *Political Research Quarterly* 63(2): 449-60.

Tadjbakhsh, Sharbanou and Anuradha Chenoy. 2007. *Human Security: Concepts and Implications*. London: Routledge.

Tamanaha, Brian Z. 2004. *On the Rule of Law: History, Politics, Theory*. Cambridge: Cambridge University Press.

Trebilcock, Michael and Ronald Daniels. 2008. *Rule of Law Reform and Development*. Cheltenham: Edward Elgar.

Treier, Shawn and Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1): 201–17.

Vanhanen, Tatu. 2000. "A New Dataset for Measuring Democracy, 1810-1998." *Journal of Peace Research* 37(2): 251-65.