An Ordinal, Concept-driven Approach to Measurement:

The Lexical Scale

John Gerring
(corresponding author)
Boston University
jgerring@bu.edu

Daniel Pemstein
North Dakota State University

Svend-Erik Skaaning
Aarhus University

Draft: 6 August 2015
Please do not cite without permission
Comments welcome!
ABSTRACT

A key obstacle to measurement in the social sciences is the aggregation problem, how to combine numerous potential indicators 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 (“inductive”) measurement model, of which factor analysis, structural equation models, and item response theory are the most commonly practiced genres. Where they do not, researchers solve the aggregation problem by appeal to concept-driven (“deductive”) criteria, i.e., aggregation schemes derived from theoretical priors rather than from patterns of covariance across the data.

While a vast literature focuses on data-driven approaches to measurement, the literature on concept-driven approaches is less developed. 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 regarding conceptual attributes as necessary-and-sufficient conditions arrayed in an ordinal scale. Following Rawls (1971), 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 categories than a dichotomy) and a classificatory function (as each level identifies a unique phenomenon). While different sorts of scales are useful for different purposes, we argue that lexical scales are in many cases superior for research questions where it is relevant to combine the differentiation of an ordinal scale with the distinct, meaningful categories of a typology.
The theoretical burden of social science is carried by highly abstract concepts such as democracy, state capacity, inequality, and rule of law. We require such concepts in order to articulate high-order theories. Yet, they are difficult to operationalize, even when agreement can be reached on a general definition. For example, everyone agrees that democracy means “rule by the people,” but there is considerable disagreement over how this general principle should be measured.

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 non-arbitrary fashion.

A classic example of this problem is presented by the concept of democracy. 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 do not correlate highly with one another (Coppedge, Alvarez & 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 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. 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. 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.

The most widely used indices of democracy are concept-driven (“deductive”) insofar as aggregation schemes are derived primarily from theoretical priors rather than from patterns of covariance across the data. This includes the Political Rights, Civil Liberty, Nations in Transit, and Countries at the Crossroads indices, all sponsored by Freedom House (2015); the Polity2 index from the Polity IV database (Marshall et al. 2014); the Democracy-Dictatorship (“DD”) index (Alvarez et al. 1996; Cheibub et al. 2010), the “BMR” index developed by Boix, Miller & Rosato (2013), the “BNR” index developed by Bernhard, Nordstrom & Reenock (2001); the Bertelsmann Transformation Index funded by the Bertelsmann Foundation (2015); 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 proposed by Freedom House and Polity are more complex but in principle no less concept-driven insofar as theoretical imperatives (rather than the empirical relationships between their indicators) guide their 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. After all, one must appeal to the meaning of a concept in order to identify potential indicators – what belongs in an

---

1 Democracy has been the touchstone for a good deal of discussion over conceptualization and measurement in recent years (e.g., Collier & Levitsky 1997; Collier & 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.
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 such as those listed above.

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 & Gerring 2009; Gerring 1999; Goertz 2006; Sartori 1984), typologies and taxonomies (Bailey 1994; Collier et al. 2012; Elman 2005; Lazarsfeld 1937; Lazarsfeld & Barton 1951), or on the general problem of concept or construct validity (Adcock & Collier 2001; Goertz 2006; Saylor 2013; Seawright & 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 protocols to choose from. This is evident in the informal manner by which democracy indices have been developed over the past several decades. It is also evident in complaints about the ad hoc, arbitrary quality of concept-driven democracy indices (Coppedge & Gerring et al. 2011; Goertz 2006; Hadenius & Teorell 2005; Munck & Verkuilen 2002).

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 regarding conceptual attributes as necessary-and-sufficient conditions arrayed in an ordinal scale. Following Rawls (1971), 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 categories than a dichotomy) and a classificatory function (as each level identifies a unique phenomenon). While different sorts of scales are useful for different purposes, we argue that lexical scales are in many cases 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 in the measurement literature, with special reference to Guttman scales and IRT. 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, and referents. Defining properties of a concept refer to attributes that provide its formal definition. Associated properties are attributes that 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). 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

Lots of social science indices are concept-driven in the loose sense defined above. Most 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 approach (that we are aware of) is governed by a determinate set of rules such that it 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 & Gerring 2009; Sartori 1984) and to set-theoretic approaches to social science (Goertz 2006; Goertz & Mahoney 2012; Schneider & 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 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 one or more necessary conditions (in this case, four), all of which must be satisfied in order to receive a score of 1 (=democracy).

Other binary indices have been proposed to measure democracy (e.g., BMR, BNR), and for most 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 (Elkins 2000). This serves as a fitting segue to our subject, which may be regarded as an extension of the principles of the binary scale.

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 follows closely the procedure by which John Rawls (1971) orders the three core principles that establish his theory 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 hinges, in large part, on a conceptual argument – 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 that which we envision for empirical concepts in the social sciences.

In this section, we lay out a protocol for the construction of a lexical scale, followed by a discussion of disagreements that may arise in this scaling procedure

Protocol

Prior to aggregation it is important to conduct a thorough survey of potential attributes attached to the concept of theoretical interest. 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 all possible attributes for Concept X are considered, and none arbitrarily excluded, one is well-advised to survey definitions and usage patterns of a concept in ordinary language and in whatever specialized language region may be relevant to the research. The initial culling of attributes should be as comprehensive as possible, excluding only idiosyncratic features.2

Next, one must order these attributes so that each 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 five 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 (which, incidentally, fulfils a goal of the classical concept). 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 about here]

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. (The construction of a binary condition may be the product of a set of necessary and/or sufficient conditions. Collectively, however, these conditions must be regarded as necessary and sufficient.)

---

2 Examples of this sort of semantic surveying can be found in Collier & Gerring (2009) and Sartori (1984).
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. A “3” on a lexical scale is not simply a midway station between “2” and “4.” Indeed, each level may be viewed as a subtype of the larger concept. Note that these sub-types 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.

The most challenging aspect of lexical scale construction is the ordering of attributes, which rests on the following considerations: (i) logical entailments, (ii) functional dependence, and (iii) conceptual centrality. Attribute $A$ may be judged prior to attribute $B$ if $A$ is logically required for $B$. For example, “elections” are logically required for “clean elections.” Thus, one would place elections prior to clean elections in an index of electoral democracy. Moreover, attribute $A$ may be judged prior to $B$ if the function of $B$ (understood relative to the definition of the concept) is functionally dependent upon $A$. For example, in order for “suffrage” to enhance democracy (understood as rule by the people) “multi-party elections” must be in place. Otherwise suffrage extensions have no relevance. Thus, one would place multi-party elections prior to suffrage in an index of electoral democracy. Finally, attribute $A$ may be considered prior to $B$ if $A$ is more central to the core meaning of a concept than $B$.3 Thus, one would presumably place “elections” prior to “civil liberty” in an index of electoral democracy.

The three criteria used to order attributes in a lexical scale are usually in harmony with one another. That is, logical entailments, functional dependence, and conceptual centrality will either suggest the same ordering of attributes, or only one will apply to a given situation. If, however, there is found to be a potential conflict among these criteria there is an implicit dominance relationship in which (i) > (ii) > (iii). Therefore, 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 the disagreement concerns the application of individual criteria not conflicts across criteria.

### Potential Disagreements

The proposed lexical scale depends upon reaching a reflective equilibrium (Daniels 1996) with respect to the defining attributes of a concept and 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 6th or 7th conditions, will be

---

3 This follows a constitutive approach to measurement, where attributes are the defining elements of a concept (Goertz 2006: 15).
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 \( II (A_1, B_1, A_2, B_2, C_1, C_2) \), 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 & 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 things, it is easier to grasp the workings of a method 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. 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 range of key social science concepts.

Electoral democracy. Electoral democracy refers to the idea that democracy is achieved through competition among leadership groups which 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 a competitive, elite, minimal, realist, or Schumpeterian model.  

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 & Chenoy 2007: Ch.2), peasants (Kurtz 2000: 96), and liberal democracy (Howard & Roessler 2006; Møller & Skaaning 2013).
of democracy (Alvarez et al. 1996; Schumpeter 1950). To operationalize this concept, we propose
the following lexical index:

0. No elections. Elections are not held for any national-level 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. No-party or one-party elections. There are regular national 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 multi-party elections.

4. Minimally competitive elections. 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
national elections.

6. Universal suffrage. Virtually all adult citizens are allowed to vote in national elections.

This 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 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). 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.

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

6 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.
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.

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 & 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:7

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).

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.

---

7 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.
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 publicly 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 & Skaaning 2012; Tamanaha 2004; Trebilcock & 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 checks and balances, 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.

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 these 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 definition 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 (see sources referenced 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.

[Figure 1 about here]

Within the literature on measurement the lexical scale is similar in structure to Guttman scaling (Coppedge & 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. 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—one each testing counting, addition, and multiplication—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 concept-driven reasoning. 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. On the other hand, knowing the
Guttman score for a case provides perfect information about which observable conditions that case satisfies. Indeed, it should be impossible to observe the above-described case if the latent concept in question actually follows a (deterministic) Guttman scale.

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 \( A-C \). In contrast, a given lexical scale implies no particular empirical pattern of correspondence between attributes. Therefore, when considering attributes that form a lexical scale, one cannot necessarily predict case attributes based on partial observations.8

Table 2 about here

Guttman scales are a form of generative model that predict observable characteristics based on an unobservable trait—for instance, one’s unobservable level of mathematical reasoning ability predicts 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. Because real-world conditions rarely match the strict conditions of an idealized Guttman scale, applied work typically applies Guttman scaling in a probabilistic manner. Generally, modern applications of Guttman scaling 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—rather than 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 theoretically matches Guttman assumptions, at least probabilistically, it makes sense to model that concept with a (probabilistic) Guttman scale. Indeed, because Guttman and subsequent IRT models generate predictions about patterns of covariation between observable attributes, allowing one to test the plausibility of core scaling assumptions, they provide a powerful tool for modeling cumulative concepts and justifying scaling decisions. Yet, many concepts fall outside the scope of Guttman models, as discussed at the outset.

While all Guttman scales are lexical scales, not all lexical scales meet the 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

8 That is, the fact that the attributes make up a lexical scale does not allow for such inferences. Other factors might allow an analyst to draw such conclusions.
Thus, lexical scales are appropriate for a class of concepts that do not fit into the Guttman framework. First, Guttman scales 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. With many concepts the outcome-based approach is often 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 substantive attributes that define the concept of interest. In this sense, lexical scales allow one to measure concepts that are not a good fit for latent variable modeling techniques, including those built upon Guttman assumptions.

Second, rankings in lexical scales are based not just on logical ordering requirements, but also considerations of functional dependence and conceptual centrality. These may, or may not, conform to a Guttman scale. Returning to our main 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. And, while one might, for instance, consider competitive elections more central to the concept to 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 would not necessarily correspond to empirical patterns in a way that would fit the Guttman model, because observing that a country has universal suffrage tells us little about whether or not it has competitive elections.

Finally, it is important to re-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 imply no deterministic or probabilistic relationship between components. Thus, it is impossible to test the validity of a lexical scale using the empirical attribute distribution. This means, moreover, that it is impossible to inductively construct a lexical scale from a set of observed attributes because no pattern of observations is more or less consistent with any particular lexical ordering of observed traits.

On one hand, this may be regarded as a weakness of the lexical scale, since researchers cannot justify their conceptual choices by appealing to the empirical record. On the other hand, our examples demonstrate that inductively constructed measures of concepts that are lexically structured may be invalid, precisely because such concepts need not imply any particular relationship between observables. 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

---

9 While lexical scales based purely on dependence are observationally indistinguishable from Guttman scales, the two approaches to scale construction are conceptually quite different. In particular, ranking decisions based on logical requirement differ subtly from the relationships between latent traits and observable implications of those traits defined by a Guttman scale. In a lexical scale, higher-level traits that logically require lower level traits cannot occur without the substantive attributes that they require. For example, you simply cannot have formal legality without rule by law. Furthermore, these requirements are traits constituent to the concept rather than empirical manifestations of latent traits.

10 A lexical scale based purely on logical requirements would be observationally equivalent to a Guttman scale. Thus one could, in principle, empirically validate the aspects of a lexical scale that rely only on such distinctions. One cannot validate ordering decisions based on functional dependence or centrality using the empirical record.
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. 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. Nonetheless, if the analyst drops attributes of a concept from an index (because they 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 by no means immune to problems of concept/construct validity.

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 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 set 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

---

11 Technically, one could use knowledge of a lexical score to predict something about the distribution of observable attributes. So, knowing that a case scores 5 on the generic scale in Table 1 tells you that the case satisfies all of the conditions. But the logic of lexical scaling is not based on a data generating process that maps latent values into observed characteristics, or responses.

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

13 Authors’ choices of indicators to include in an index are often somewhat arbitrary (Goertz 2006; Haig & 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 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.
number of aggregation principles are embedded in a complex statistical model. On the other hand, 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 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 & Vreeland 2010; Pemstein, Meserve & Melton 2010; Treier & 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 & 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. 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.

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 a Lexical index has any claim to ontological priority over 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.

---

14 With respect to the Freedom House indices, Cheibub, Gandhi & 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 $5^{10} = 9,765,625$ possible ways to obtain a sum of scores between zero and 40 in political rights, and $5^{15} = 30,517,578,125$ 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.”
VI. References


**Table 1:**

Generic Lexical scale

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.</td>
<td>$\sim A$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>$A$</td>
<td>$\sim B$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>$A$</td>
<td>$B$</td>
<td>$\sim C$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>$A$</td>
<td>$B$</td>
<td>$C$</td>
<td>$\sim D$</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>$A$</td>
<td>$B$</td>
<td>$C$</td>
<td>$D$</td>
<td>$\sim E$</td>
</tr>
<tr>
<td>5.</td>
<td>$A$</td>
<td>$B$</td>
<td>$C$</td>
<td>$D$</td>
<td>$E$</td>
</tr>
</tbody>
</table>

0-5 = ordinal scale. $A-E$ = conditions that are satisfied. $\sim A-E$ = conditions that are not satisfied. Relationships are deterministic except where cells are undefined (empty).
Figure 1:
Lexical Scale in Tree-Diagram (Taxonomic) Format

~A       A
   ~B       B
      ~C       C
         ~D       D
             ~E       E
Table 2:

Generic Guttman Scale

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>~B</th>
<th>~C</th>
<th>~D</th>
<th>~E</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>~A</td>
<td>~B</td>
<td>~C</td>
<td>~D</td>
<td>~E</td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>~B</td>
<td>~C</td>
<td>~D</td>
<td>~E</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>B</td>
<td>~C</td>
<td>~D</td>
<td>~E</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>~D</td>
<td>~E</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>~E</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
</tr>
</tbody>
</table>

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