19

Conceptualization and Measurement: Basic Distinctions and Guidelines

Gerardo L. Munck, Jørgen Møller and Svend-Erik Skaaning

One of the key aims of the social sciences is to describe the social world. Descriptions are one of the most powerful products of the social sciences. Based on descriptions, countries are ranked as being more or less democratic or respectful of human rights or corrupt; the level of violence over time within and between particular groups is gauged; political parties are compared on a left–right spectrum; citizens are held to have more or less liberal or religious values, and so on. Much of what we know about the social world is due to research that seeks to provide descriptions. In addition, research oriented to offering descriptions provides important input for research that aims at explaining the social world.

In this chapter, we offer an overview of the issues involved in producing the data that are used in descriptions. The overview is divided into three main sections. We begin by focusing on the task of conceptualization. *Concepts* play a fundamental but frequently unappreciated role in the production of data. We clarify the components of concepts, discuss how concepts can be organized and distinguish among different kinds of conceptual systems. We next turn to measurement, distinguishing between the production of *data on indicators* and *data on indices*. The notions of indicators and indices are sometimes used interchangeably. However, the tasks and choices involved in producing data on indicators and indices, respectively, are distinct and better addressed one at a time. Thus, in the second section we focus on data on indicators, and address the task of selecting indicators, designing measurement scales and collecting data. Subsequently, in the third section, we turn to data on indices, where we develop a key distinction between two kinds of indices – those that combine data on multiple units and those that combine data on multiple indicators measuring different properties in one unit – and discuss key options concerning these two kinds of indices.

Data can be good or poor, and we are also concerned with ensuring that data are of high

quality. Thus, we discuss not only what is involved in *producing* data but also what is involved in *evaluating* descriptions. Ideally, as we suggest, evaluations would feed back into the production of data, but frequently evaluations are carried out as a post-production task. To this end, we discuss various criteria that are relevant to an evaluation of data. However, because this chapter focuses on concepts and the link between concepts and measures, and does not provide a full discussion of measurement, we emphasize the criterion of validity and conceptualize it more broadly than is customary.

We provide many examples to illustrate our points about methodology. However, one of our recurring examples is democracy. This is a concept that has been the center of attention in much of the methodological literature.[1] Moreover, it is a concept that is central to a broad body of substantive research in political science and other disciplines.

CONCEPTUALIZATION

Concepts are the building blocks of the social sciences, as they are of all sciences. There is no theory without concepts, there is no description without concepts, and there is no explanation without concepts. Thus, concept formation – conceptualization – has logical priority in research because it is a key input in all subsequent steps, including those concerned with the production of data. Moreover, though quantity and quality are mutually complementary, every quantitative concept presupposes a qualitative concept. Indeed, as Sartori (1970: 1038) put it, because we cannot measure something if we have not specified its meaning, 'concept formation stands prior to quantification'. Or, more broadly, as Bunge (1995: 3; 2012: 122) argues, 'concept formation precedes empirical test' and 'in concept formation quality precedes quantity' (see also Lazarsfeld and Barton, 1951: 155–6). Thus, researchers need to focus on the formation of concepts and to recognize the qualitative foundations of all research.

There are no rules on how to form a concept, just as there are no rules that can be followed to create a theory. Concepts are formed through a combination of induction and deduction. As suggested by Adcock and Collier (2001: 531–3), the decisions about a concept that is to be used in the social sciences are frequently made in light of a dialogue with 'the broader constellation of meanings and understandings associated with a given concept', or what they label the 'background concept'. Moreover, the link between conceptualizing and theorizing is very close: as Kaplan (1964: 53) notes, 'proper concepts are needed to formulate a good theory, but we need a good theory to arrive at the proper concepts'. Concept formation, like theory building, is largely an art.

Nonetheless, the product as opposed to the process of concept formation can surely be assessed. Concepts can be clear or vague. Concepts can be well formed or poorly formed. Concepts can be more or less elaborate and systematized. Indeed, there are various features of concepts that are used to distinguish good from bad concepts. At the very minimum, it is important to be clear about the various parts of a concept and the ways in which the sense or meaning of a concept is organized, which are two matters we address next.

Term, Sense and Reference

A concept consists of three interrelated elements. The first is a *term*. This is a sign that designates the *sense* or connotation of a concept – the part of the concept that is frequently understood as its meaning – and the latter in turn refers to the objects that are included in the *reference* or denotation of a concept (see Figure 19.1).

Most of the discussion about concepts rightly focuses on concepts' *sense*, which is

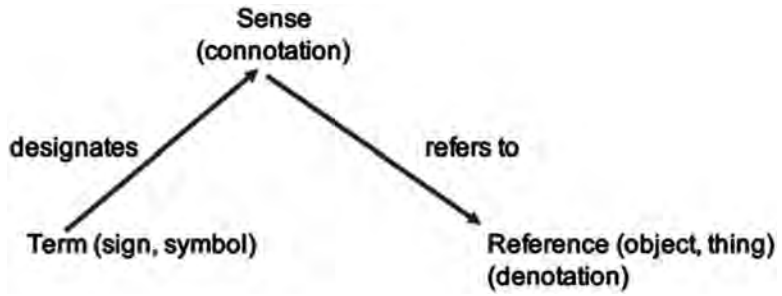

Figure 19.1 The parts of a concept: term, sense and reference

Note: This depiction is an adaptation of what is commonly known as the semantic triangle or the Ogden/Richards triangle (Ogden and Richards, 1923: 11).

given by the conceptual attributes that represent properties of objects and the relationship among conceptual attributes. Indeed, the meaning of a concept can largely be taken to be conveyed by a concept's sense, and debates about concepts focus mainly on this part of a concept. For example, debates about the concept of democracy since the work of Schumpeter (1942) hinge on matters such as what the conceptual attributes of democracy are and what the relationship among conceptual attributes is (Collier and Levitsky, 1997). We discuss this aspect of concepts more fully below. However, first, a few brief comments regarding a concept's term and reference are in order.

First, the role played by the term of a concept might seem rather simple. But Dahl's (1971) effort to introduce the term 'polyarchy', so as to avoid the possible confusion created by the multiple uses given to the term 'democracy', shows that terminological issues are not trivial. Indeed, there are many terms that are given different meaning. Furthermore, understanding how a term is used requires some knowledge of the broader semantic field in which it is embedded. For example, though the term 'regime' has a different meaning in the fields of comparative politics and international relations, the difference is clarified once the term is placed within the semantic field of these two fields of research. Thus, while terminological matters are not the most important ones, they certainly deserve some attention (Sartori, 2009 [1975]: 61–9; 2009 [1984]: 111–15, 123–5).

Second, the idea of the *reference* of a concept needs to be clarified at the outset. A concept's reference (aka the domain of a concept) is all objects to which a concept refers and is thus related to the unit of analysis of a study. In contrast, a concept's *extension* is those objects which actually have certain properties. It is important to grasp the distinction between reference and extension, and the relationship between them. Though statements about reference rely on theoretical concepts and do not presuppose that of truth, statements about extension rely on empirical concepts and do presuppose that of truth. For example, it is one thing to say that democracy is a series of properties of political communities and another to say country x is a democracy. Indeed, the latter is an empirical claim, which could be factually true or false and can only be addressed once data has been collected, and hence is not strictly a conceptual matter (Bunge, 1974a: ch. 2; 1974b: 133–53; 1998a [1967]: 73–80).[2] Thus, we start our discussion here by considering theoretical concepts and purely conceptual operations, before turning to empirical concepts and empirical operations, such as the construction of indicators and data collection.

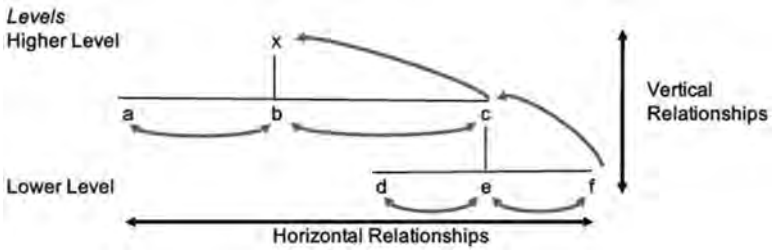

Figure 19.2 The structure of a concept: levels and relationships

The Attributes and Structure of a Concept

Turning to a more detailed discussion of a concept's sense, it is critical to recognize that a concept's sense is conveyed by (i) the conceptual attributes that represent properties of objects and (ii) the relationship among conceptual attributes or, for short, the structure of a concept. For this reason, the meaning of concepts is not fully conveyed by a simple listing of conceptual attributes, a common feature of definitions.

Listing the defining attributes of a concept is useful. It puts the focus on what conceptual attributes should be *included* in a concept. Moreover, inasmuch as a definition also clarifies what conceptual attributes should be *excluded* from a concept (even though they are included by some scholars), such an exercise is critical. For example, one of the ongoing concerns in the discussion about the concept of democracy is how to strike the right balance between expanding the concept of democracy beyond the sphere of elections. This can, for example, be done by adding attributes considered to be part of democracy (e.g. horizontal accountability), and expanding the concept of democracy in such a way that what might be considered extraneous attributes are included in the concept of democracy (e.g. the economic equality of citizens) (Munck, 2016).

However, it is important to note that any such list offers an incomplete sketch of a concept. Indeed, inasmuch as more than one conceptual attribute is posited, it is necessary to inquire about the structure of a concept, which is given by the relationships among conceptual attributes at the same level (horizontal relationships) and at different levels (vertical relationships) (see Figure 19.2). That is to say, the meaning of a concept might not be conveyed by each attribute taken individually, in an additive manner, and the structure of a concept could be key to its meaning. Thus, to fully and correctly grasp the meaning of a concept, it is crucial to appreciate that concepts can be – indeed, usually are – *conceptual systems*, which in turn are part of larger conceptual systems or semantic fields.[3]

It is also important to distinguish among different kinds of conceptual systems that connect and systematize multiple concepts that share, at least partially, their sense or their reference.[4] The simplest form is the *typology*, which unifies a series of concepts of different connotation but at the same level and of the same scope by proposing the underlying dimensions of multiple concepts. An example of such conceptual systems is Aristotle's (1995 [c. 330 BC]: Book III, chs 6 and 7) classical typology of political regimes, which relies on the underlying dimensions of the number of persons who exercise power and the ends they seek. Another is Dahl's (1971: 7) modern typology of political regimes, which relies on the underlying

Figure 19.3 Conceptual systems I: typologies

Source: Dahl (1971: 7).

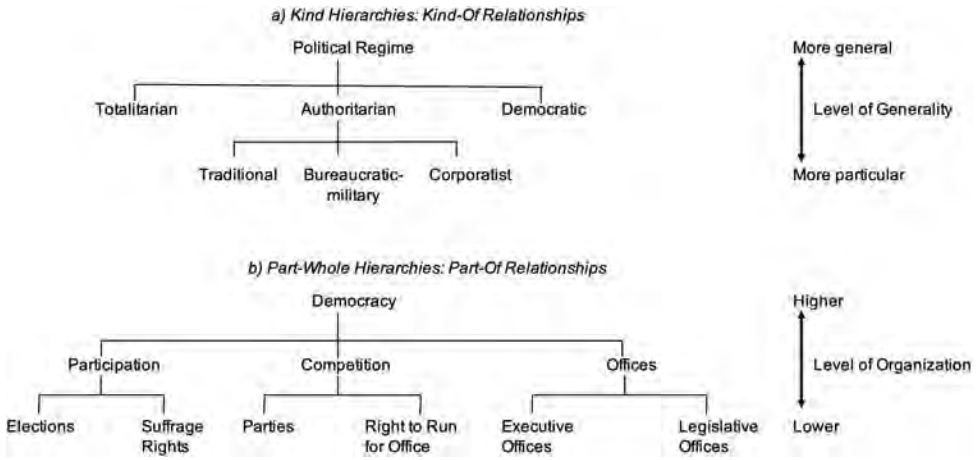

Figure 19.4 Conceptual systems II: two hierarchical structures

Note: The example of a kind hierarchy draws on Linz (1975); the example of a part-whole hierarchy draws on Schumpeter (1942), Dahl (1971), and Przeworski et al. (2000).

dimensions of contestation and participation (see Figure 19.3).[5]

A different, and more complex, conceptual system is the *taxonomy*, which connects concepts at different levels in a hierarchical structure, in one of two ways. One hierarchical structure, sometimes called a *kind hierarchy,* organizes concepts that partition collections of objects into groups and subgroups, and yields a classic taxonomy. An example of this kind of conceptual system is Juan Linz's (1975) encompassing and nuanced classification of 20th-century political regimes (see Figure 19.4, panel a), which are defined in terms of the underlying dimensions of pluralism, ideology, mobilization and leadership. But there is another hierarchical structure, sometimes called a *part-whole hierarchy,* that organizes concepts that decompose wholes into parts and also connects parts to the whole.[6] A classic example of such a hierarchy is the conceptualization of democracy that decomposes a whole – democracy – into parts at various levels of organization (see Figure 19.4, panel b).

Evaluation

Concepts are not true or false. Nonetheless, not all concepts rest on an equally sound foundation; some have been carefully elaborated and justified, while others are merely stipulated without much in the way of reflection. Without making any claim to exhaustiveness, we propose a set of criteria to assess whether concepts are good or bad.

Most basically, concepts have to be *intelligible*. This means that we should be able to answer the following questions: what is the concept designed by the term or symbol used for? What are the conceptual attributes? What is the structure of a concept, that is, what are the relationships among conceptual attributes? What is the reference of a concept?

Several criteria regarding a concept's term can be highlighted (Sartori, 2009 [1975]: 61–9; 2009 [1984]: 111–15, 123–25, 132–3; Gerring, 1999). Most critical is the criteria of *terminological univocality*, that is, the avoidance of terminological ambiguity introduced through the use of homonyms, terms with multiple meanings, and synonyms, multiple terms with the same meanings. Other criteria concern the *fit* of the term with the terminology used in prior research, and the *familiarity* and *resonance* of the term.[7]

Another criterion is *logical formation*. Inasmuch as any concept consists of more than one conceptual attribute, it is important to ask whether the proposed conceptual system fulfills, for a given domain, two logical requirements. First, they should be mutually exclusive, meaning that no concept or conceptual attribute at the same level overlaps with the meaning of another concept or conceptual attribute; and they should be collectively exhaustive, meaning that no concept or conceptual attribute that is part of a conceptual space is excluded. In addition, inasmuch as any concept consists of more than one conceptual attribute at different levels, whether or not the conceptual attributes are logically organized by level of generality or organization is a key consideration (Lazarsfeld and Barton, 1951: 156–9).

Yet another key criterion, which deserves some elaboration, is *conceptual validity*, understood here with reference to the sense of the concept and both the conceptual attributes and the structure of a concept.[8] The inclusion and exclusion of conceptual attributes is a key decision in the formation of a concept. The same goes for any decision regarding the relationship among conceptual attributes. And each decision can and should be assessed in terms of the extent to which the decision is theoretically justified.[9]

It bears noting that the assumption underpinning this criterion – that concepts can and should be assessed in light of their theoretical justification – is not universally accepted. On the one hand, many scholars posit that a number of concepts, and especially those that have an obvious normative connotation, are 'essentially contested' and that they will always remain 'open' in the sense that a research community will never agree on a definitive definition (Gallie, 1956; Gray, 1978). From this relativist perspective, any claim that a certain concept is more theoretically justified than another can be portrayed as arbitrary or subjective. This perspective could even lead to the view that since disputes over the meaning of concepts cannot be resolved, any effort at measurement is futile, in that claims about what is measured cannot be settled.

However, it is not obvious that, for example, democracy, seen as the 'essentially contested' concept *par excellence*, merits such a characterization (Bobbio, 1989: ch. 4, 2003: 42; Beetham, 1994: 27; see also Arblaster, 2002: 6–10). Though disagreements about the concept of democracy persist, it is clear that the research by Schumpeter (1942) and Dahl (1971) has led to widespread consensus about the core meaning of democracy in research on democratization (Munck, 2009: 16–23, 2016). The same can be said about other concepts with strong normative

resonance. For example, Waldron (2002) observes that while the institutional or political arrangements required by the rule of law – another concept frequently characterized as essentially contested – are subject to disagreement, there is actually considerable consensus about its basic formal–legal requirements, such as that laws are prospective, open and clear and that there is congruence between official action and declared rule (see also Collier et al., 2006: 228–30; Møller and Skaaning, 2014: ch. 1).

On the other hand, a common epistemology, empiricism, holds that knowledge only concerns observable properties and that empirical concepts but not theoretical ones are acceptable (Bridgman, 1927; Carnap, 1936, 1937). From this perspective, the suggestion that concepts could be assessed in light of theory would be deemed unjustified and all work on theoretical concepts would be no more than a distraction from, and even a hindrance to, the real work of measurement (King et al., 1994: 25, ch. 2, 109–10). However, the distinction between, and mutual irreducibility of, theoretical and empirical concepts is well established (Kaplan, 1964: 54–60; Sartori, 2009 [1975]: 83–4; Laudan, 1977: chs 1 and 2). And the shortcomings of the empiricists' endeavor to reduce the theoretical to the empirical are evident (Bunge, 2012: ch. 13). Indeed, the main concepts in the social sciences are theoretical as opposed to empirical. Key examples are society, economy, class, ideology, politics, state, power, rights, constitutionalism, democracy, rule of law, welfare and peace. Few scholars are willing to remain silent about these concepts.

In short, these two extremes can and should be avoided. *Contra* Gallie, many key concepts have been theoretically developed enough to have some shared meanings, and measurement does not have to wait until all conceptual disputes are resolved. *Contra* empiricists, the banishment of theoretical concepts is simply a self-defeating position that is hard to consistently maintain. Thus, the validation of concepts by reference to theory is both viable and central.

MEASUREMENT I: DATA ON INDICATORS

Turning from conceptualization to measurement opens up a whole new series of challenges. Theoretical concepts refer to at least some imperceptible facts. Thus, inasmuch as social scientists seek to describe and explain the world, they must address some complicated empirical operations involved in measurement (see Figure 19.5). First, to bridge theoretical concepts and facts, they must develop *indicators*, which relate observable properties to unobservable ones, and propose how to draw distinctions based on indicators. Second, to produce data, they must engage in *data collection*, which assigns (qualitative or quantitative) values to indicators in light of observables about objects. In other words, they must design and use measuring instruments. Thus, though any attempt to produce data must begin with a clear idea of *what* is to be measured, the distinct issues involved in *how* to measure some theoretical concept –

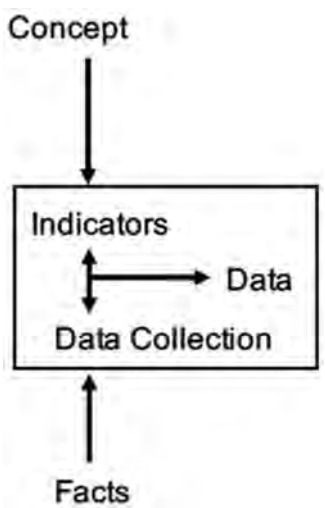

Figure 19.5 The concept–fact interface

the development of indicators and, relatedly, of measurement scales, and data collection and coding – deserve scrutiny.

Indicators

The general challenge in developing indicators, sometimes called operationalization, is to build a bridge between unobservables and observables, that is, to link theoretical concepts that refer to facts about properties of objects with empirical concepts, an observable property of the same object (Bunge, 1998b [1967]: 192–6).[10]

Due to the nature of indicators, probably the hardest challenge in the design of a measuring instrument relates to what is usually called *content validity* – the extent to which one or more indicators capture the correct and full sense or content of the concept being measured (Adcock and Collier, 2001: 536–40). The goal of measurement is to generate data that can be used to evaluate the truth of claims about facts (e.g. the US is a democracy in 2019). But data will be useful for this purpose only inasmuch as any data collected on some indicators can be linked back to the concept (e.g. the concept of democracy in this example) used in a factual claim. Building such bridges is anything but an easy task, especially when the concept of interest is multidimensional, that is, has many conceptual attributes.

This task is made harder because researchers also have to be concerned about measurement *equivalence*, the extent to which an indicator captures the same sense or content of the concept being measured in different contexts (Przeworski and Teune, 1970: chs 5 and 6; Adcock and Collier, 2001: 534–6; Davidov et al., 2014). Often it is not obvious that the same indicator will have similar meanings in different countries, for different persons and in different time periods. This means, first, that it is often necessary to use several indicators to measure a concept in order to capture different nuances of the concept and increase the reliability of any measures, and second, that different contexts sometimes call for different indicators to capture the same facts. This can be understood by invoking the distinction between common and system-specific indicators (Przeworski and Teune, 1970: ch. 6). Common indicators work in the same way across different contexts, while system-specific indicators vary across contexts but are functional equivalents, meaning that they are in principle substitutable. For example, actions considered corrupt in one place are considered appropriate behavior elsewhere, so asking similar questions about corruption will not provide equivalent measures. In survey research, questions should preferably have the same meaning for all respondents, but linguistic, cultural and other differences make it difficult to establish measurement equivalence (see also Locke and Thelen, 1995; van Deth, 1998).

At the same time, the search for system-specific indicators can lead to an excessive, even paralyzing, emphasis on the unique and can open the door to relativism. For example, most current global datasets on democracy rely on common indicators and hence could be criticized for not taking into account how different conceptual attributes of democracy should be adapted to different contexts. In addition, some of these datasets have been criticized, and rightly so, for having a Western bias, in that specific Western institutions are treated as universal standards for assessing other countries. However, it is clear that an attempt to factor in ideas from the literature on 'non-Western democracies', especially the argument that democracy takes a different form in non-Western societies, amounts to a rejection of any standard to compare countries around the world. What might at first glance seem a rather simple empirical operation – the design of indicators – actually hides many potential pitfalls. Indeed, for these reasons, the development of indicators that offer a basis for testing factual claims has been recognized as an important accomplishment

(Harré, 1981), and the development of broad cross-national measures, such as those used to measure economic activity around the world, are celebrated (Vanoli, 2005).

Measurement Scales

The design of indicators is inextricably linked with another task, namely, the design of the measurement scales used to discriminate among cases. The standard options are well known: there are, most basically, nominal, ordinal, interval and ratio scales. Moreover, the standard way of comparing these scales barely needs mention: the move from nominal toward ratio scales involves a gain in precision. Thus, all else being equal, scales designed to collect data that is more precise and informative are preferable. Or, as is sometimes argued, inasmuch as nominal and ordinal scales are treated as qualitative scales, quantification is a sign of progress.

However, we add a caveat to this conventional wisdom that is suggested by the debate about whether democracy is best seen as a question of either–or or more-or-less (Sartori, 1987; Bollen, 1990; Collier and Adcock, 1999). In this debate, many authors have suggested persuasively that nominal and ordinal scales are sometimes preferable, in that they actually capture better the concept of interest. For example, the common idea of a democratic transition suggests that some changes are actually qualitative in nature and hence that nominal scales are appropriate (Przeworski et al., 2000: 18). Likewise, a common argument in the literature on democratization is that the extension of the right of suffrage evolved one social group at a time, a change well captured by an ordinal scale. Thus, it is important to note that decisions regarding measurement scales are made in the context of specific concerns and concepts, and hence that, as Collier and Adcock (1999: 537) suggest, these decisions should be justified through 'specific arguments linked to the goals of research' rather than by reference to the superior information of certain scales when considered in the abstract.[11]

Data Collection

Once a researcher has designed an indicator or a series of indicators, each with their own measurement scale, the distinct task of data collection – the gathering and categorization of relevant information about a phenomenon of interest for a researcher – can begin in earnest. In this regard, we caution against a narrow view of the possible kinds of data and sources of data, and hence a narrow view of the challenges involved in data collection and the problems that might emerge in the course of data collection. Indeed, an overreliance on data from data-rich countries or time periods (e.g. the US in current times) would likely introduce bias into our knowledge of the social world. Moreover, in thinking about data collection, we draw attention to three questions: (1) When and where was it created? (2) Who created it? (3) For what purposes was it created? Answers to these questions provide the background information required to carry out systematic source criticism (*Quellenkritik*), which is the process of evaluating whether information (of all kinds) is more or less valid, reliable or relevant for a particular purpose.

Sources of qualitative data

A key distinction is frequently made between primary sources and secondary sources. Primary sources provide direct or firsthand evidence about events, objects or persons. They include historical and legal documents, eyewitness accounts, interviews, surveys, audio and video recordings, photographs, speeches, diaries, letters, art objects and various kinds of online communications (e.g. emails, tweets, posts, blog entries). Secondary sources provide some kind of interpretation and analysis of events, conditions or experiences. Hence, newspaper articles and reports can be either

primary or secondary sources, depending on whether they provide information about facts or analysis and interpretation.

The ideas associated with systematic source criticism and the distinction between primary and secondary sources have their origin in the academic discipline of history. Historical data presents a number of attractions for social scientists, including more variation on key variables, the ability to investigate how similar causal mechanisms play out in different contexts and the ability to analyze path dependency. However, the more social scientists delve back in time, the more they come to depend on the prior work of trained historians, who have produced the narrative accounts that social scientists use either to code historical datasets or to produce in-depth historical narratives.

This raises an important but often ignored challenge: that social scientists will be prone to solely enlist or overly emphasize 'works by historians using implicit theories about how events unfold and how people behave very similar to the theory under consideration' (Lustick, 1996: 607). To mitigate this risk, social scientists first need to recognize that historical work cannot be seen as theoretically neutral. The implicit or explicit theoretical and historiographical perspectives of historians (e.g. the Marxist or Annales schools) color the ways they interpret their findings. Social scientists must therefore build a representative body of historical data from which to draw inferences. This means that they need a deep knowledge about the development of historiography and the debates of historical work in a particular field (Lustick, 1996; Lange, 2013: 141–8).

Different guidelines have been developed to ensure this. Lustick (1996) proposes four strategies:

- *Explain variance in historiography*: Assume a normal distribution among historical works and then identify the consensus.
- *Be true to your school*: Identify a particular historical tradition or school as superior for the purpose at hand and then accept this interpretation, knowing how it differs from other interpretations.
- *Quasi-triangulation*: Limit the readings of history to those interpretations that have a broader support across historical schools.
- *Explicit triage*: Argue why some historical studies are better than others given the task at hand.

Møller and Skaaning (2019: 6) endorse Lustick's argument that social scientists need to systematically consider differences between historical interpretations, but they criticize the notion that the average or consensus interpretation is less biased. Instead, to avoid selection bias in the sources of data, they suggest that social scientists should factor in the 'shape of the distribution within historiography' in three ways:

- *Aim for conceptual consistency*: Prioritize historical interpretations that are based on similar concepts as those being considered by the social scientist.
- *Clarify the vantage point of historical accounts*: Prioritize historical interpretations that are relatively atheoretical or where the thesis conflicts with the thesis that the social scientist is interrogating.
- *Prioritize updated evidence*: Prioritize historical interpretations that are based on newer evidence.

These three criteria are anchored in a simple Bayesian logic and they enable social scientists to heed what has been termed 'the Ulysses Principle', that is, to figuratively tie oneself to the mast in order to take precautions against influencing the evidence that is used to examine descriptive or causal propositions (Møller and Skaaning, 2019).

This principle, it bears noting, is not only relevant when dealing with historical sources. Recent methodological debates have emphasized the possibility of going deep more generally by shifting the focus from the macro level of analysis to the micro level, so as to probe mechanisms (Beach and Pedersen, 2016). While there are different ways of doing this, they all force social scientists to deal with qualitative data sources, such as interviews, archives, newspapers, organization records

and reports and participants' observations (see Tilly, 2008; McAdam et al., 2008). This requires not only a close familiarity with the data, but also careful consideration about how to avoid bias in the identification and reading of qualitative sources. If one takes out the historical part of the criteria mentioned above, they are applicable for processing many different kinds of qualitative data.

Sources of quantitative data

Shifting focus to quantitative data, these normally take one of five forms:

1 *Hand-coded data*, such as the CIRI Human Rights Database, the Manifesto Project Database, the Uppsala Conflict Data Program, the Freedom House data on Political Rights and Civil Liberties and the Polity IV Project, where researchers or their assistants code events or conditions based on some predefined criteria.
2 *Machine-coded data*, such as the Integrated Crisis Early Warning System, the Global Database of Events, Language, and Tone (GDELT), and the Fragile States Index, where researchers develop automated algorithms that can categorize behavior, conditions or opinions.
3 *Ordinary survey data*, such as the World Values Survey, the Afrobarometer and various national election studies, where a sample (often representative) of people belonging to a particular group (citizens of a nation, employees in a firm, parliamentarians, members of an organization, etc.) is enlisted to respond to a number of questions about opinions and behavior.
4 *Expert survey data*, such as parts of the Varieties of Democracy dataset, the Chapel Hill Expert Survey, the Perceptions of Electoral Integrity dataset and the Quality of Government Survey, where experts are enlisted to answer questions about a certain topic about which they have special competence.
5 *Administrative data*, such as election turnout and vote share, roll call votes, number of state employees and government financial and economic statistics, which have been collected by national public agencies and international organizations (e.g. the UN, the World Bank, the IMF and the OECD).

In situations where we are interested in measuring not opinions but the actual condition of, say, different aspects of democracy or the prevalence of corruption, another distinction has received much attention: namely, the difference between fact-based and judgment-based indicators. Those favoring fact-based (directly observable and verifiable) indicators emphasize that such data are more transparent and replicable and therefore broadly recognizable. They criticize judgement-based and perception-based data for being based on fuzzy and unsubstantiated inferences and personal biases.

Users and producers of judgement-based indicators have responded to this criticism by pointing out that fact-based indicators are often unable to capture all relevant nuances of particular phenomena. The preference for fact-based data rests, according to Schedler (2012: 28), on two conditions, which are often not fulfilled: '(1) transparent empirical phenomena whose observation do not depend on our judgmental faculties and (2) complete public records on those phenomena'. For example, some aspects of democracy, such as freedom of expression, are not easily observable. More generally, '[s]ome empirical phenomena we cannot observe in principle, others we cannot observe in practice' (Schedler, 2012: 28). In a nutshell, the problem is that directly observable empirical information is often incomplete, inconsistent or insufficient.

Different types of evidence can of course be used simultaneously to answer particular research questions. Just as researchers can make use of methods triangulation in order to appraise theoretical expectations, they can also carry out data triangulation and take advantage of the strengths and shortcomings of different kinds of sources of data (Skaaning, 2018). In general, the combination of information from different kinds of data increases our ability to capture related, but distinct, aspects of the variable in question. In addition, relying on multiple indicators can reduce the impact of idiosyncratic measurement errors associated with single indicators and facilitates systematic assessment of how reliable the data are.

There has recently been much talk of new data collection methods, based on increased computer power and a plethora of new information that is accessible online – what has been referred to as 'big data'. Web scraping of information from, for example, newspapers or Wikipedia or social media (Twitter, Facebook) allows scholars to build large datasets. One partial novelty here is to treat text – including alterations of text on Wikipedia and the like – as data. These newer sources of data collection are addressed in other chapters of the *Handbook*. Hence, all we note here is that the issues of conceptualization and measurement discussed in this chapter are also relevant for these new data collection enterprises.

On coding

One of the more versatile means of producing systematic data – whether quantitative or qualitative, whether on variables or causal mechanisms, whether for a large-scale or a small-scale project, whether for the current period or times long past – is hand-coding by a single scholar or a team of scholars. Even though this is only one among various means of assigning values to indicators, given its important role in the social sciences we offer some comments about this procedure.

The production of hand-coded data normally proceeds in particular stages. Relevant information is gathered, after which a coder evaluates the evidence on one or more issues and translates it into a score based on more or less explicit and precise standards or coding rules. Despite careful attention to the selection of sources, training of coders and documentation of coding procedures, specific biases can still influence the scores (Bollen and Paxton, 1998, 2000).

The accessibility and selection of sources is a major issue. Evidence has been through a filtering process in which some information passes through and some is filtered out. This process is likely to introduce problems because the filters are selective in non-random ways, meaning that the information is generally neither complete nor representative.

If the patterns of incomplete data are not random, descriptive and explanatory analyses using the data will be biased. For instance, Casper and Tufis (2003) have demonstrated that some of the most prominent democracy measures are not genuinely interchangeable, even though they are all anchored in Dahl's (1971) definition of polyarchy and even though they are highly correlated (between .85 and .92). One reason for this could be systematic missingness. For example, relevant information is frequently not available for poor countries and autocracies. Missingness can be evaluated by simple tests of non-random missingness (see e.g. Ríos-Figueroa and Staton, 2012), where one examines whether there are significant differences between the scores for units covered by the data and those units that are not covered on other variables expected to be related to the outcome that is being researched.

Another issue is how the coders or respondents process the evidence. They can introduce random and systematic measurement errors by interpreting the sources differently, either because they base their evaluation on different pieces of (relevant or irrelevant) information, because they weight the same evidence differently or because they have different understandings of the concepts and scales that are used. More generally, various actors in the 'data supply chain' respond to different incentives and have variable capabilities that influence – and sometimes consciously manipulate – the production of data (Herrera and Kapur, 2007).

In addition, the practical procedures in the specific coding processes can introduce method effects. For example, scores can be influenced by how many units and questions the coders process, whether and when revisions can be made or whether they code across cases or over time. All of these factors tend to influence the implicit reference points in the minds of coders and thus the scores that are generated through exercises in coding.

On a more general level of abstraction, the reproducibility of measurement procedures is an important aspect of social science. This requires a systematic approach to data collection, precise descriptions of the procedures and transparency of these issues. Ideally, researchers should be able to reproduce or replicate the scores, and then assess the results of independent coding exercises. For example, where multiple, overlapping indicators exist, if the same variable is coded by several coders for the same units, one can assess the extent to which they generate consistent and converging data. In such cases, inter-coder reliability tests are valuable tools to assess whether the assumptions about consensus among coders are met (Gwet, 2014).

One way to do this is to employ Item-Response Theory (IRT) modeling techniques. These use patterns of agreement between the scores from different coders/indicators (and sometimes also other kinds of information, such as coder characteristics) to identify variations in reliability and systematic bias, and use this information to reduce measurement error in connection to latent concepts and to generate systematic estimates of uncertainty.

Evaluation

An evaluation of measuring instruments and the data on indicators produced by using these instruments, much as with concepts, hinges first of all on *intelligibility*. If an independent scholar is not able to comprehend how the data was produced, what decisions were made to produce the data, and what the reasons were for at least the key decisions, the data cannot be properly scrutinized. In other words, without transparency, there is no possibility of replication and no way of assessing reliability and validity.

The demand for transparency has traditionally been directed mostly at quantitative data, but it has recently been pushed by the DA-RT (Data Access and Research Transparency) initiative within the American Political Science Association with respect to qualitative research as well. One of the tools that has been proposed is data repositories that allow researchers to store qualitative data in a systematic way. This enables scholars to document their evidentiary record and makes it possible for other scholars to acquaint themselves with what is written in the sources that are referred to for evidence. For instance, the use of active citation gives readers a quick way to assess if a particular observation or interpretation does indeed seem to be supported by the work that is referenced (Lupia and Elman, 2014).

There are many other criteria that could be used to assess measuring instruments and data on indicators. As noted, measuring instruments can be more or less *versatile*, that is, they can be better or worse suited to generate data on various concepts in different domains (that is, temporal and spatial units). Data can be more or less *reliable,* that is, yield the same results when repeated measures are carried out independently. Data can have more or less *measurement error*, and identifying the sources of such error and providing estimates of uncertainty is part of best practice.

Importantly, in contrast to the evaluation of concepts, the evaluation of data on indicators can rely on empirical tests, using the data that has been produced and other available data (Cronbach and Meehl, 1955; Campbell and Fiske, 1959; Adcock and Collier, 2001; Seawright and Collier, 2014; McMann et al., 2016). For example, in a test of *convergent-discriminant validity* a researcher examines to what degree a new measure converges with established measures of the same concept and diverges from established measures of different concepts. In turn, in a test of *nomological validity* a researcher examines to what degree a new measure is able to reproduce well-established relationships among variables. Thus, it is important that researchers take advantage of the various empirical tests that can yield information that is relevant to an assessment of data.

However, the value of such tests depends very much on the current state of empirical knowledge. That is, a test of convergent-discriminant validation requires that a researcher can take for granted that the other measures, the standards with which the measure of interest is compared, are valid. In turn, a test of nomological validation requires that a researcher can take for granted that the established relationship is valid. Yet frequently this is not the case, and hence these tests may simply not be relevant. Moreover, the proponents of new measures frequently challenge existing conceptualizations or explanations, making agreement with prior knowledge an improper standard.

Thus, it is critical to stress the centrality of the question of *content validity*, that is, the extent to which one or more indicators capture the correct and full sense or content of the concept being measured (Adcock and Collier, 2001: 536–40). Assessing the validity of data is complex, because it concerns the link between observables and unobservables. Moreover, unlike estimates of convergent-discriminant and nomological validity, it cannot be quantified through an analysis of the data. However, it is important to recognize some key points about content validity. First, the question of content validity is distinctive. Second, it has priority in an evaluation of measurement validity, in the sense that it should be addressed first, during the process of indicator construction, and that it affects the data that are used in tests of convergent-discriminant and nomological validity. Third, it is an important consideration regardless of the kind of data (quantitative or qualitative) that is produced.

MEASUREMENT II: DATA ON INDICES

Data analysis for the purpose of description and explanation frequently relies on data on indicators. However, the production of data on indicators frequently raises a new question: how might these data on indicators be combined? Indeed, there are many reasons why a scholar may want to develop what can generically be called *indices*, which combine data on indicators. The production of indices involves complex considerations, several of which are of a technical nature, and there is a large literature on index formation (e.g. Lazarsfeld, 1958; Lazarsfeld and Menzel, 1961; Blalock, 1982: ch. 7; Bollen and Lennox, 1991; Nardo et al., 2005; Greco et al., 2019). Thus, our discussion is necessarily cursory. Nonetheless, we draw attention to some key distinctions and options that have not always been addressed with clarity in the recent literature, and introduce some considerations that are ignored by the literature on measurement that pays little or no attention to the connection between theoretical concepts and measurement.

At the broadest level, drawing on the distinction between two of the core parts of a concept, its sense and reference (see above), it is possible to distinguish between two kinds of indices: (i) indices that combine data on the same indicator (measuring the same property) in multiple units (e.g. percentage of people in the world earning less than 2 dollars a day), and (ii) indices that combine data on multiple indicators (measuring different properties) in one unit (e.g. how democratic is the US) (see Figure 19.6).[12] In addition, building on these two kinds of indices, megaindices can be, and frequently are, built (e.g. proportion of countries in the world that are democracies, proportions of country dyads in the world that are democratic dyads, etc.). However, the core issues and options concern these two basic situations.

Combining Data on Units

In the social sciences, the lowest level of analysis is the individual, and hence the most fine-grained data that are collected are data on properties of individuals. From this basic starting point, it is possible to combine data

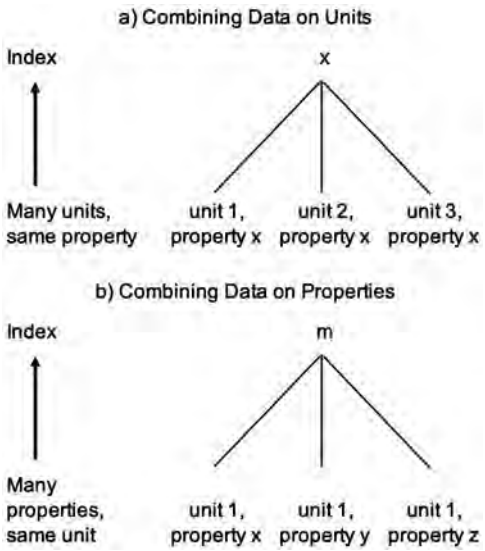

Figure 19.6 The production of data on indices: two basic situations

on units all the way up to the highest possible level of analysis, the world system. However, there are two different ways, corresponding to two different social properties, in which data on units can be combined, and the index that is produced is different depending on which option is chosen (Lazarsfeld, 1958: 111–12; Lazarsfeld and Menzel, 1961: 426–8).

When the data on different units (e.g. individuals, firms or states) concerns a property possessed by each unit (e.g. income or life), an index that represents an *aggregate or resultant property* is generated. Examples are GDP, GDP per capita, percent of GDP accounted for by trade, global GDP, number of deaths in war, homicides per 100,000, proportion of the population that supports democracy and percentage of votes won by candidates in an election. In turn, when the data on different units concerns a property a unit has by virtue of a relationship among units (e.g. relative income, capital–labor relations or trading relationship between states), an index that represents a *relational or structural property* is generated. Examples are income inequality, polarization of the class structure, conflict levels of industrial relations, judicial independence, state legitimacy, trade dependence between countries and eigenvector centrality.

These are not the only social properties. Indeed, as Lazarsfeld (1958: 112–13; Lazarsfeld and Menzel, 1961: 428–9) pointed out long ago, there is a third kind of social property: *global or emergent properties*. These properties are not based on information about lower level units because they are not possessed by each lower level unit either independently of other units or due to a relationship with other units. Examples of global or emergent properties are crowd behavior, national culture, social cohesion, political stability and the dominant mode of production. The measurement of such social properties does not proceed by combining data on the same property in multiple units.

Combining Data on Properties

A second kind of index is produced by combining data on multiple indicators (measuring different properties) in one unit. To be sure, the production of such indices does not need to be limited to one unit. For example, though some scholars have developed an index of democracy for one country, it is common for scholars to produce indices covering many countries or even the entire world. The increase in the number of units opens some important possibilities, such as tests of dimensionality. But the point is that the focus of such index production is on the question of how data on multiple indicators, each linked with different conceptual attributes, should be combined.

This challenge has been the subject of considerable debate, and different scholars have different views about how such a challenge should be addressed. Nonetheless, in broad strokes, the key choice a researcher faces is whether aggregation, that is, the combination of data on multiple indicators, should be

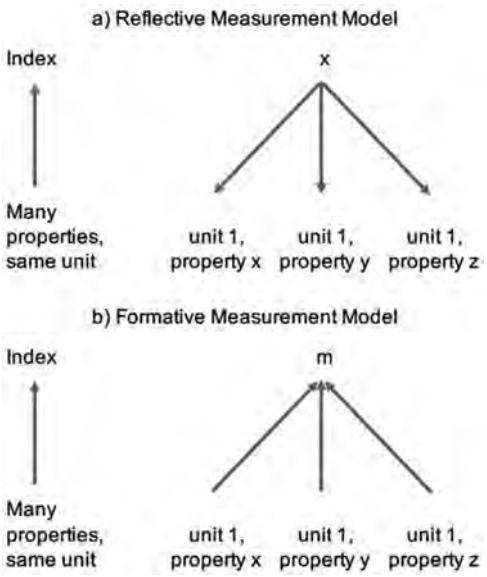

Figure 19.7 Combining data on properties: reflective and formative measurement models

based on what has been called a reflective or a formative aggregation model.[13]

These aggregation models differ both conceptually and substantively. In a *reflective model*, the concept is understood as the common 'cause' of the indicators used to measure it. Hence, 'causation' runs from the latent concept to the indicators (see Figure 19.7, panel a). Changes in the latent trait (not directly observed) are therefore expected to 'cause' a change in the indicator scores, but not vice versa, and that change in the latent variable should simultaneously bring about variation on all indicators. It follows that indicators should have a high positive correlation. This indicates that the multiple indicators and hence conceptual dimensions tap into a single underlying dimension. If so, indicators can be seen as partially interchangeable and dropping one of these indicators would not alter the meaning of the index that is produced. A good example is Teorell's (2010: 164–5) socioeconomic modernization index, which he constructs, through the use of principal components analysis, by combining information on nine indicators: net output of the non-agricultural sector as percentage of GDP, gross secondary school enrolment ratio, urban population as percentage of total population, life expectancy at birth, infant mortality rate, the log of GDP per capita, radios per capita, televisions per capita and newspaper circulation per capita. The indicators all load highly on a common latent dimension, which lends support to the index construction.

The assumptions behind a *formative model* are different. A latent concept is construed as the summary of the relevant variation in a set of indicators that are understood as constitutive of a particular concept. In other words, a latent concept is composed of conceptual attributes that are individually important for the meaning of the concept. In this case, 'causation' flows from the indicators to the latent concept (see Figure 19.7, panel b). In contrast to reflective models, in formative models the correlation among indicators is considered irrelevant and, since the indicators are understood as defining attributes, excluding one or more of them will fundamentally alter the meaning of the concept that is to be captured. To illustrate, contestation (or competitive elections) and inclusive suffrage are often conceived as the two essential features of representative government (Dahl, 1971; Coppedge et al., 2008). These two conceptual attributes are not necessarily highly correlated with each other. Today, many countries have universal adult suffrage but not much contestation, and historically many countries had a high degree of contestation but highly restrictive voting rights. However, only including indicators that capture either suffrage or contestation would critically alter the core concept that is being measured. Measuring one property cannot substitute for the measurement of another property, and dropping the data on one of the multiple properties would radically alter the meaning of the index that is produced.

Evaluation

The constructors of indices must tackle some distinct choices, beyond those that go into the

production of data of indicators, when they consider whether and how to aggregate data from indicators. In combining data on the same indicator in multiple units, analysts need to be aware of what social property is being measured, and hence whether the appropriate procedure is being used. In turn, in combining data on multiple indicators (measuring different properties) in one (or more) units, they have to be aware at least of the choice between reflective or formative aggregation models. However, as Lazarsfeld (1958: 113) noted, it is by no means self-evident how an analyst should proceed. Indeed, at times it is not even clear whether an analyst faces the challenge of combining data on units or on properties.

Given this uncertainty, the temptation to rely on default options might be strong. But this temptation should be resisted. As with the evaluation of data on indicators, empirical tests, using the data that have been produced and other available data, can be conducted and used to inform the construction of indices. Indeed, various empirical checks can be of help (Bollen and Bauldry, 2011). However, one cannot simply make the data speak for itself. Thus, no matter which of the options is seen as more suitable for a given aggregation task, whatever procedure is used to form an index through the combination of data on indicators needs to be justified theoretically. What this means is that, to ensure what has been called concept-measure consistency (Goertz, 2006: ch. 4), which might be thought of as a counterpart or aspect of the criterion of *content validity* discussed above, what is needed is a theory about how multiple indicators should be combined (Goertz, 2006: 53–65, ch. 5; Munck, 2009: 30–2, 49–51). Indeed, much as with data on indicators, data on indices are valid if they fulfill two criteria: (i) a theoretical concept has been formed in a conscious and careful manner, that is, a theory has been articulated to justify what conceptual attributes are included and excluded, how the included conceptual attributes relate to each other and what the referents of conceptual attributes are; and (ii) the way in which data on indicators is combined matches the concept that is being measured.

CONCLUSION

The social sciences, in contrast to disciplines such as logic and mathematics, are factual sciences, given that they refer to facts about the concrete world. Thus, empirics and, more narrowly, measurement, understood as the production of data, are essential parts of social science research. However, empirics should be distinguished from empiricism. Empiricism is a one-sided epistemology that holds that experience is the only source of knowledge and that, in the context of measurement, asserts that theoretical concepts are not different from empirical concepts or that theoretical concepts can be reduced to empirical ones. The history of science reveals the limitations of empiricism. Indeed, a widely recognized indicator of progress is the replacement of classification schemes based on concepts that represent secondary, observable properties with ones based on primary, non-observable properties of things. For example, the conceptualization of chemical elements based on atomic number and electron configuration rather than observable properties such as color or smell, and the classifications in biology based on molecular differences rather than observable morphological traits.

Thus, counter to an empiricist approach to measurement, this chapter places the focus squarely on theoretical concepts and insists on the link between theoretical concepts and measures. Indeed, we have sought to draw attention to various ways in which a clear idea of *what* theoretical concept is to be measured is needed to make decisions regarding *how* to measure that theoretical concept. And to that end, we started by addressing what concepts and conceptual systems are, and then highlighted how both the production of data

on indicators and indices should consider the link between concepts and measures.

We do not seek to convey the message that the link between concepts and data should be the only concern in any measurement project. Other matters are also important. Moreover, not every project on measurement has to be conjoined with a project on conceptualization. There can surely be a division of labor between researchers who seek to form concepts and researchers who produce data. However, for data to be used to ascertain the truth of the kind of factual claims that are routinely made in the social sciences, decisions regarding the production of data *must be* guided by ideas regarding the sense and reference of concepts as well as their structure. Measures that ignore these matters are of limited value and, inasmuch as they are interpreted as measures of theoretical concepts, potentially erroneous.

Notes

1 Collier and Levitsky (1997); Bollen and Paxton (1998, 2000); Collier and Adcock (1999); Munck and Verkuilen (2002); Goertz (2006); Gerring et al. (2019).

2 This distinction between reference and extension is frequently overlooked. Indeed, it is not addressed in the influential work on the social sciences by Sartori (1970, 2009 [1984]: 102–6). However, it actually is consistent with Sartori's (2009 [1975]: 84) clarification that 'the rules of transformation along a ladder of abstraction … apply to observational, not to theoretical, concepts'. That is to say, though the intension and extension of a concept varies inversely (Sartori, 1970: 1040–4; Collier and Mahon, 1993: 846), this statement applies only to empirical concepts and not to theoretical concepts. For more on the distinction between theoretical and empirical concepts, see Kaplan (1964: 54–60).

3 Bunge (1998a [1967]: 82–9); Sartori (2009 [1984]: 118–25); Thagard (1992); Collier and Levitsky (2009).

4 For useful discussions about concepts and conceptual systems, see Bunge (1998a [1967]: chs 2 and 3), Bailey (1973, 1994), Marradi (1990) and Collier et al. (2012).

5 2x2 typologies have been hugely influential in social science, both for descriptive and explanatory purposes. But typological property spaces are often much more complicated, as they can contain more than two dimensions and as each of these dimensions can be divided into more than two classes. For example, one could add the rule of law, the effective power to govern and/or the guiding ideology as separate dimensions to the regime typology in Figure 19.3 if one has good theoretical reason to do so. Or one could subdivide contestation and/or participation into different levels, say low, medium and high. The problem with such operations is that the property space can quickly become too complex to be useful for theorizing and empirical analysis. Thus, the essence of forming a typology is to first identify the dimensions and the classes on each dimension, and then to reduce the property space in order to focus on the most important types (Lazarsfeld and Barton, 1951: 169–80; Elman, 2005; Collier et al., 2012).

6 On the distinction between kind and part-whole hierarchical structures, see Thagard (1990, 1992: 7–8, 27–33) and Collier and Levitsky (2009).

7 For an exemplary critical analysis of the term 'authoritarianism', as used in the study of political regimes, see Przeworski (2017).

8 Though the idea of measurement validity is ubiquitous in the literature on measurement, the distinction between conceptual validity and measurement validity is rarely made; for exceptions, see Jackson and Maraun (1996) and Billiet (2016: 196–200). Yet, inasmuch as the idea that there are theoretical concepts apart from their measures is accepted, as is the case here, this distinction is crucial.

9 For exemplary justifications of the concept of democracy, see Dahl (1989) and Saward (1998).

10 Inasmuch as some observable property of another object is lawfully related to the observable property of an object under consideration, the observable property of another object could be used as an indicator.

11 There is an associated issue that crops up frequently in the measurement of democracy. Scholars have good reasons to want qualitative *and* quantitative distinctions. However, one common practice – the derivation of qualitative distinctions from quantitative distinctions – deserves scrutiny. Indeed, such exercises tend to rely on a rather arbitrary assertion, usually made with little reference to the concept of democracy, that some point on a scale can be treated as the dividing line between democracy and non-democracy. It is preferable to start with qualitative distinctions and then refine these measures by adding quantitative distinctions.

12 The problem of combining data also occurs if multiple scores are generated for a single indica-

tor in the same unit (e.g. when multiple coders are used in data based on expert rating) or if data are generated for multiple indicators of the same conceptual property in the same unit (e.g. when a battery of indicators are used to measure some psychological trait). Here we take as our starting point data which can already be treated as data on conceptual properties.

13 On reflective and formative aggregation models, see Blalock (1982: ch. 7); Bollen and Lennox (1991); Edwards and Bagozzi (2000); Coltman et al. (2008); Bollen and Bauldry (2011); Edwards (2011).

REFERENCES

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

Arblaster, Anthony. 2002. *Democracy.* 3rd ed. Buckingham and Philadelphia: Open University Press.

Aristotle. 1995 [c. 330 BC]. *Politics*. Oxford: Oxford University Press.

Bailey, Kenneth D. 1973. 'Monothetic and Polythetic Typologies and Their Relation to Conceptualization, Measurement, and Scaling.' *American Sociological Review* 38(1): 18–32.

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

Beach, Derek, and Rasmus Brun Pedersen. 2016. *Causal Case Study Methods: Foundations and Guidelines for Comparing, Matching, and Tracing.* Ann Arbor, MI: University of Michigan Press.

Beetham, David. 1994. 'Key Principles and Indices for a Democratic Audit', pp. 25–43, in David Beetham (ed.), *Defining and Measuring Democracy*. London: Sage.

Billiet, Jaak. 2016. 'What Does Measurement Mean in a Survey Context?', pp. 193–209, in Christof Wolf, Dominique Joye, Tom W. Smith and Yang-chih Fu (eds), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.

Blalock, Hubert M. 1982. *Conceptualization and Measurement in the Social Sciences.* Beverly Hills, CA: Sage.

Bobbio, Norberto. 1989. *Democracy and Dictatorship: The Nature and Limits of State Power*. Minneapolis, MI: University of Minnesota Press.

Bobbio, Norberto. 2003. *Teoría general de la política.* Madrid: Editorial Trotta.

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

Bollen, Kenneth A., and Shawn Bauldry. 2011. 'Three Cs in Measurement Models: Causal Indicators, Composite Indicators, and Covariates.' *Psychological Methods* 16(3): 265–84.

Bollen, Kenneth A., and Richard Lennox. 1991. 'Conventional Wisdom on Measurement: A Structural Equation Perspective.' *Psychological Bulletin* 110(2): 305–14.

Bollen, Kenneth A., and Pamela Paxton. 1998. 'Detection and Determinants of Bias in Subjective Measures.' *American Sociological Review* 63(3): 465–78.

Bollen, Kenneth A., and Pamela Paxton. 2000. 'Subjective Measures of Liberal Democracy.' *Comparative Political Studies* 33(1): 58–86.

Bridgman, P. W. 1927. *The Logic of Modern Physics.* New York, NY: The Macmillan Company.

Bunge, Mario. 1974a. *Treatise on Basic Philosophy* Vol. 1. *Semantics I: Sense and Reference.* Dordrecht, Holland: D. Reidel Publishing Company.

Bunge, Mario. 1974b. *Treatise on Basic Philosophy* Vol. 2. *Semantics II: Interpretation and Truth.* Dordrecht, Holland: D. Reidel Publishing Company.

Bunge, Mario. 1995. 'Quality, Quantity, Pseudoquantity, and Measurement in Social Science.' *Journal of Quantitative Linguistics* 2(1): 1–10.

Bunge, Mario. 1998a [1967]. *Philosophy of Science.* Vol. 1. *From Problem to Theory.* New Brunswick, NJ: Transaction Publishers.

Bunge, Mario. 1998b [1967]. *Philosophy of Science* Vol. 2. *From Explanation to Justification.* New Brunswick, NJ: Transaction Publishers.

Bunge, Mario. 2012. *Evaluating Philosophies*. New York: Springer.

Campbell, Donald T., and Donald W. Fiske. 1959. 'Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix.' *Psychological Bulletin* 56(2): 81–105.

Carnap, Rudolf. 1936. 'Testability and Meaning.' *Philosophy of Science* 3(4): 419–71.

Carnap, Rudolf. 1937. 'Testability and Meaning.' *Philosophy of Science* 4(1): 1–40.

Casper, Gretchen, and Claudiu Tufis. 2003. 'Correlation versus Interchangeability: The Limited Robustness of Empirical Finding on Democracy Using Highly Correlated Data Sets.' *Political Analysis* 11(2): 196–203.

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

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, Jody LaPorte and Jason Seawright. 2012. 'Putting Typologies to Work: Concept Formation, Measurement, and Analytic Rigor.' *Political Research Quarterly* 65(1): 217–32.

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

Collier, David, and Steven Levitsky. 2009. 'Democracy: Conceptual Hierarchies in Comparative Research', pp. 269–88, in David Collier and John Gerring (eds), *Concepts and Method in Social Science: The Tradition of Giovanni Sartori.* New York: Routledge.

Collier, David, and James E. Mahon. 1993. 'Conceptual "Stretching" Revisited: Adapting Categories in Comparative Analysis.' *American Political Science Review* 87(4): 845–55.

Coltman, Tim, Timothy M. Devinney, David F. Midgley and Sunil Veniak. 2008. 'Formative versus Reflective Measurement Models: Two Applications of Formative Measurement.' *Journal of Business Research* 61(12): 1250–62.

Coppedge, Michael, Angel Alvarez and Claudia Maldonado. 2008. 'Two Persistent Dimensions of Democracy: Contestation and Inclusiveness.' *Journal of Politics* 70(3): 632–47.

Cronbach, Lee Joseph, and Paul E. Meehl. 1955. 'Construct Validity in Psychological Tests.' *Psychological Bulletin* 52(4): 281–302.

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

Dahl, Robert A. 1989. *Democracy and Its Critics*. New Haven, CT: Yale University Press.

Davidov, Eldad, Bart Meuleman, Jan Cieciuch, Peter Schmidt and Jaak Billiet. 2014. 'Measurement Equivalence in Cross-National Research.' *Annual Review of Sociology* 40: 55–75.

Edwards, Jeffrey R. 2011. 'The Fallacy of Formative Measurement.' *Organizational Research Methods* 14(2): 370–88.

Edwards, Jeffrey R., and Richard P. Bagozzi. 2000. 'On the Nature and Direction of Relationships between Constructs and Measures.' *Psychological Methods* 5(2): 155–74.

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

Gallie, Walter B. 1956. 'Essentially Contested Concepts.' *Proceedings of the Aristotelian Society* 56: 167–98.

Gerring, John. 1999. 'What Makes a Concept Good? A Criterial Framework for Understanding Concept Formation in the Social Sciences.' *Polity* 31(3): 357–93.

Gerring, John, Daniel Pemstein and Svend-Erik Skaaning. 2019. 'An Ordinal, Concept-driven Approach to Measurement: The Lexical Scale.' *Sociological Methods and Research*. DOI: https://doi.org/10.1177/0049124118782531

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

Gray, John. 1978. 'On Liberty, Liberalism and Essential Contestability.' *British Journal of Political Science* 8(4): 385–402.

Greco, Salvatore, Alessio Ishizaka, Menelaos Tasiou and Gianpiero Torrisi. 2019. 'On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness.' *Social Indicators Research* 141(1): 61–94.

Gwet, Kilem L. 2014. *Handbook of Inter-Rater Reliability*. 4th ed. Gaithersburg, MD: Advanced Analytics.

Harré, Rom. 1981. *Great Scientific Experiments: Twenty Experiments that Changed Our View of the World*. Oxford: Phaidon Press.

Herrera, Yoshiko, M. and Devesh Kapur. 2007. 'Improving Data Quality: Actors, Incentives, and Capabilities.' *Political Analysis* 15(4): 365–85.

Jackson, Jeremy S. H., and Michael Maraun. 1996. 'The Conceptual Validity of Empirical

Scale Construction: The Case of the Sensation Seeking Scale.' *Personality and Individual Differences* 21(1): 103–10.

Kaplan, Abraham. 1964. *The Conduct of Inquiry: Methodology for Behavioral Science.* Scranton, PA: Chandler Publishing Co.

King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Lange, Matthew. 2013. *Comparative-Historical Methods.* Los Angeles, CA: Sage.

Laudan, Larry. 1977. *Progress and Its Problems: Toward a Theory of Scientific Growth.* Berkeley: University of California Press.

Lazarsfeld, Paul F. 1958. 'Evidence and Inference in Social Research.' *Daedalus* 87(4): 99–130.

Lazarsfeld, Paul F., and Allen H. Barton. 1951. 'Qualitative Measurement in the Social Sciences: Classification, Typologies, and Indices', pp. 155–92, in Daniel Lerner and Harold D. Lasswell (eds), *The Policy Sciences: Recent Developments in Scope and Method*. Stanford, CA: Stanford University Press.

Lazarsfeld, Paul F., and Herbert Menzel. 1961. 'On the Relation between Individual and Collective', pp. 422–40, in Amitai Etzioni (ed.), *Complex Organizations: A Sociological Reader*. New York: Holt, Rinehart and Winston.

Linz, Juan J. 1975. 'Totalitarianism and Authoritarian Regimes', pp. 175–411, in Fred Greenstein and Nelson Polsby (eds), *Handbook of Political Science* Vol. 3*, Macropolitical Theory*. Reading, MA: Addison-Wesley Press.

Locke, Richard M., and Kathleen Thelen. 1995. 'Apples and Oranges Revisited: Contextualized Comparisons and the Study of Comparative Labor Politics.' *Politics and Society* 23(3): 337–67.

Lupia, Arthur and Colin Elman. 2014. 'Openness in Political Science: Data Access and Research Transparency.' *PS: Political Science and Politics* 47(1): 19–42.

Lustick, Ian S. 1996. 'History, Historiography, and Political Science: Multiple Historical Records and the Problem of Selection Bias.' *American Political Science Review* 90(3): 605–18.

Marradi, Alberto. 1990. 'Classification, Typology, Taxonomy.' *Quality & Quantity* 24(2): 129–57.

McAdam, Doug, Sidney Tarrow and Charles Tilly. 2008. 'Methods for Measuring Mechanisms of Contention.' *Qualitative Sociology* 31(4): 307–31.

McMann, Kelly M., Daniel Pemstein, Brigitte Seim, Jan Teorell and Staffan I. Lindberg. 2016. *Strategies of Validation: Assessing the Varieties of Democracy Corruption Data. V-Dem Working Paper* 2016: 23.

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

Munck, Gerardo L. 2016. 'What Is Democracy? A Reconceptualization of the Quality of Democracy.' *Democratization* 23(1): 1–26.

Munck, Gerardo L., and Jay Verkuilen. 2002. 'Conceptualizing and Measuring Democracy: Evaluating Alternative Indices.' *Comparative Political Studies* 35(1): 5–34.

Møller, Jørgen, and Svend-Erik Skaaning. 2014. *The Rule of Law: Definitions, Measures, Patterns and Causes.* New York: Palgrave Macmillan.

Møller, Jørgen, and Svend-Erik Skaaning. 2019. 'The Ulysses Principle: A Criterial Framework for Reducing Bias When Enlisting the Work of Historians.' *Sociological Methods and Research*. DOI: https://doi.org/10.1177/0049124118769107

Nardo, Michela, Michaela Saisana, Andrea Saltelli, Stefano Tarantola, Anders Hoffman and Enrico Giovannini. 2005. *Handbook on Constructing Composite Indicators.* Paris: OECD Publishing.

Ogden, C. K., and I. A. Richards. 1923. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and the Science of Symbolism*. London: Routledge.

Przeworski, Adam. 2017. 'A Conceptual History of Political Regimes: Democracy, Dictatorship, and Authoritarianism.' *Studia Socjologiczno-Polityczne. Seria Nowa* 7(2): 9–30.

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

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

Ríos-Figueroa, Julio, and Jeffrey Staton. 2012. 'An Evaluation of Cross-National Measures of Judicial Independence.' *Journal of Law, Economics, and Organization* 30(1): 104–37.

Sartori, Giovanni. 1970. 'Concept Misformation in Comparative Politics.' *American Political Science Review* 64(4): 1033–53.

Sartori, Giovanni. 1987. *The Theory of Democracy Revisited* Part 1: *The Contemporary Debate*. Chatham, NJ: Chatham House Publishers.

Sartori, Giovanni. 2009 [1975]. 'The Tower of Babel', pp. 61–96, in David Collier and John Gerring (eds), *Concepts and Method in Social Science: The Tradition of Giovanni Sartori.* New York: Routledge.

Sartori, Giovanni. 2009 [1984]. 'Guidelines for Concept Analysis', pp. 97–150, in David Collier and John Gerring (eds), *Concepts and Method in Social Science: The Tradition of Giovanni Sartori.* New York: Routledge.

Saward, Michael. 1998. *The Terms of Democracy.* Cambridge: Polity Press.

Schedler, Andreas. 2012. 'Judgment and Measurement in Political Science.' *Perspectives on Politics* 10(1): 21–36.

Schumpeter, Joseph A. 1942. *Capitalism, Socialism, and Democracy*. New York: Harper and Brothers.

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

Skaaning, Svend-Erik. 2018. 'Different Types of Data and the Validity of Democracy Measures.' *Politics and Governance* 6(1): 105–16.

Teorell, Jan. 2010. *Determinants of Democratization*. New York: Cambridge University Press.

Thagard, Paul. 1990. 'Concepts and Conceptual Change.' *Synthese* 82(2): 255–74.

Thagard, Paul. 1992. *Conceptual Revolutions*. Princeton, NJ: Princeton University Press.

Tilly, Charles. 2008. 'Describing, Measuring, and Explaining Struggle.' *Qualitative Sociology* 31(1): 1–13.

van Deth, Jan W. (ed.). 1998. *Comparative Politics: The Problem of Equivalence*. London: Routledge.

Vanoli, André. 2005. *A History of National Accounting.* Amsterdam: IOS Press.

Waldron, Jeremy. 2002. 'Is the Rule of Law an Essentially Contested Concept (in Florida)?' *Law and Philosophy* 21(2): 137–64.