Do you know your data? Criteria for dataset quality

– Work in progress, April 2, 2008 –

Paper prepared for delivery at the workshop

“*The numbers we use, the world we see*”, ECPR Joint Sessions 2008, Rennes

Ingo Rohlfing, PhD

Research Associate

Chair for Empirical Social and Economic Research

Department of Management, Economics and Social Sciences

University of Cologne

Herbert-Lewin-Str. 2

D-50931 Köln

rohlfing@wiso.uni-koeln.de

For comments I am grateful to Hans-Jürgen Andreß, André Kaiser, Henning Lohmann, and Hawal Shamon. All errors are my own. Valuable research assistance was provided by Caroline Wehner. The paper is part of the project “*The quality of macrodata in the social sciences*” at the University of Cologne.

1. Introduction

Macrodata is widely used in comparative research both in qualitative and quantitative analysis (cf., Bollen, Entwisle, and Alderson 1993; Munck and Snyder 2007). In contrast to the widespread application, there is little general reflection about the quality of available datasets. One of the most comprehensive discussions on data quality can be found in Merritt and Rokkan's (1966) edited volume on the *Yale Political Data Program* that was suspended some time ago. However, the volume is exclusively concerned with microdata, which is only one specific type of data (Kittel 2006), and focuses on its utility for cross-national research as opposed to data quality in the first place (but see Friedrich 1966). One also finds some discussions of available data in particular fields of research like democracy (e.g., Munck and Verkuilen 2002) or taxation (Lieberman 2002) and presentations of datasets such as in the newsletter of the *APSA Comparative Politics Section* (e.g., Ledet 2006). These studies, on which I will partially draw in my paper, are based on an implicit or explicit notion of what good macrodata is. As will be seen, however, they are incomplete in some important respects. Moreover, they are tailored to a specific issue area or dataset, which limits the transferability of the underlying criteria to other fields of research. Finally, the current literature offers some reflection on the positive and negative incentives data users and producers face for the improvement of datasets (Herrera and Kapur 2007). What is lacking hitherto is a comprehensive and broadly applicable framework specifying criteria that macrodata should meet. The development of such guidelines is important because the mindless use of existing data may be an impediment for scientific progress. The increasing sophistication in quantitative and qualitative methodology will be of limited value when the employed data is of poor quality.

My paper aims to promote the discussion on data quality and to contribute to the improvement of datasets by developing a *unified* framework with which one can assess the quality of available macrodatasets. It details the tasks a data producer faces in the generation of high-quality data and enables the data user to determine whether a dataset has been properly constructed. I will argue that the quality of data should be determined on five dimensions each of which subsumes a set of criteria: *conceptualization*, *operationalization*, *data collection and processing*, *aggregation,* and *representativeness*. I will explain why each of these dimensions is an integral element of good macrodata and what potential sources for inferior data are. Moreover, I will describe instruments with which can discern whether a dataset conforms to a standard and present available remedies. The treatment of each dimension is supplemented by a discussion of existing datasets. Pointing to the deficiencies of a dataset is not meant as a critic of those who have produced the data. The generation of a dataset is always to be welcomed because weak data is better than none at all. However, generating valid inferences makes it essential to recognize shortcomings of available datasets and the resulting implications (cf., King, Keohane, and Verba 1994, pp. 28-31).

In section two, I will begin with a discussion of what macrodata is in order to clarify to what type of data my criteria particularly apply to. In section three, I will briefly introduce my unified framework for the assessment of data quality. Sections four to eight comprise the detailed discussion of the criteria data should meet. In section nine, I will turn to the implications of low-quality data for the generation of measurement error. Contrary what one may believe, I contend that the consequences of inferior data are far from straightforward because the mismeasurement of cases does not necessarily follow a violation of one or more quality criteria. The last section concludes the paper by emphasizing the need of high-quality data for the accumulation of knowledge in the social sciences.

2. What is macrodata?

It is common in the social sciences to distinguish between three levels of analysis and, correspondingly, three types of data: *micro*, *meso*, and *macro*. The assignment of units of analysis to one level is important in empirical research if one aims to understand the effects different levels have on each other. For example, it is a common assumption that macrovariables like properties of the nation state have effects on the action of collective and individual actors operating within this state (Kittel 2006; Rokkan 1966, pp. 19-20).[1] Microdata capture information about individuals and small groups like households (e.g., Duncan and Hill 1989; Mahler 2004). Mesodata refers to collective and cooperative actors like companies and political parties (Scharpf 1997, pp. 51-68). There is some degree of ambiguity involved in the distinction between micro- and mesodata since there is no solid basis for assigning small groups either to the micro- or mesolevel. A similar problem holds for the differentiation between meso- and macrodata. It is undisputed that the nation state is located at the macrolevel (Kittel 2006). The more difficult and ambiguous question is whether subnational entities like regions and communities belong to the macrolevel as well or are better placed at the mesolevel.[2]

[1] One may argue that there are no absolute criteria for the specification of the level of analysis, but that it depends on the point of view. In this perspective, it would be possible to conceive of states in an international relations study as operating on the microlevel. However, I consider absolute criteria more appropriate, which is also the dominant view in the literature {Kittel, 2008 #3499}.

[2] It seems reasonable to conceive of entities right below the level of the nation state as "macro" (e.g., the states in the United States and Germany), while units at an even lower level, like cities or municipalities, are more appropriately located at the mesolevel (cf., Teune 1968, p. 126).

One can further distinguish between *aggregated* and *genuine* macrodata as two different types of data.[3] Genuine macrodata yield information about properties of the nation state, for example, whether a country belongs to the group of federal or unitary states (cf., Lijphart 1999, chap. 10). Such information cannot be aggregated from the micro- and mesolevel, but is an inherent characteristic of a state. Conversely, aggregated data is derived from summing up micro- or mesolevel information (cf., Scheuch 1966). Classic variables that can be subsumed under this category are unemployment rates (e.g., Carlsen 2000) and public expenditure (e.g., Kittel and Obinger 2003). One specific manifestation of aggregated macrodata are summary statistics of public surveys, which for example underlie the *World Governance Indicators* dataset of the World Bank (Kaufmann, Kraay, and Mastruzzi 2007a).[4] Given that there is a vast amount of literature dealing with the specific problems of survey research (e.g., Blasius and Thiessen 2006; King et al. 2003), I leave this specific subtype of aggregated macrodata aside in the following.[5]

Both genuine and aggregated macrovariables should be subsumed under the rubric of macrodata because both potentially have effects on lower-level units of analysis. One can examine how federalism, a genuine macrovariable, affects the national fiscal policy through the incentives it creates for political actors on different territorial dimensions (e.g., Braun, Bullinger, and Wälti 2002). In a similar vein, unemployment rates are regularly used as control variables in most of the quantitative studies on how parties shape the welfare state in an era of globalization (e.g., Garrett and Mitchell 2001; Kittel 1999). Table 1 gives examples for aggregated and genuine data for all three levels of analysis and corresponding types of data.

Table 1: Examples for genuine and aggregated micro-, meso-, and macrodata

	Microdata	Mesodata	Macrodata
Genuine	Personal political attitudes	Centralization of sectoral wage bargaining	Type of political regime (federal vs. unitary)
Aggregated	Income of households	Sectoral unemployment rate	Gross domestic product

[3] An additional distinction concerns the one between subjective and objective data (cf., Bollen 1993). Since it is orthogonal to the micro-meso-macro dimension, I do not discuss it further here but return to it later in the paper when discussing specific criteria.

[4] The dataset and related material can be accessed via http://info.worldbank.org/governance/wgi2007/ (last accessed 01/16/08).

[5] Once the survey data is generated it is easy to aggregate it, for example by calculating the mean.

All dimensions of high-quality data to be discussed in the following apply to both types of macrodata. For this reason, I will simply refer to macrodata in the remainder of this paper.[6]

3. A framework for high-quality macrodata at a glance

The comprehensive perspective on data quality that I propose is depicted in figure 1.[7] It partially draws on and is inspired by Adcock and Collier's (2001) treatment of measurement validity and Goertz' (2006) approach toward concept formation. The assessment of any dataset follows the steps a data producer should perform when generating a dataset.[8] The first task involves the specification of the background concept – democracy, war, etc. – in the form of a systematized concept.[9] The consistency of the *conceptualization* is high when three points are given. First, concept specificity must be maximal, which means that the understanding of the background concept must be made explicit. This point may seem evident, but it is often far from obvious what the definition of a concept is (Goertz 2006, ch. 4). Second, the same definition should apply to all units under analysis at any point in time so as to achieve conceptual homogeneity. Third, concept fit should be maximized by excluding irrelevant and including all relevant attributes of a concept. The second task calls for the *operationalization* of the attributes constituting the systematized concept. The first standard that is at stake here is attribute specificity, which requires explicating the indicators with which the concept is measured. In addition, the content validity of the indicators should be high, that is, they should be proper measures of what they are supposed to measure. The third part aims at proper *data collection and processing* and involves two subdimensions. First, the selected sources must provide a pool of information allowing one to generate valid indicator scores. The second aspect concerns the reliability of the scoring procedure. It is to be ensured

[6] Depending on the research interest, it may be useful to distinguish between aggregated and genuine macrophenomena. For example, changing the state of an aggregated macrolevel phenomenon like unemployment rates demands a different political strategy than the change of a genuine phenomenon like federalism that is not an aggregate of the behavior of individual or collective actors.

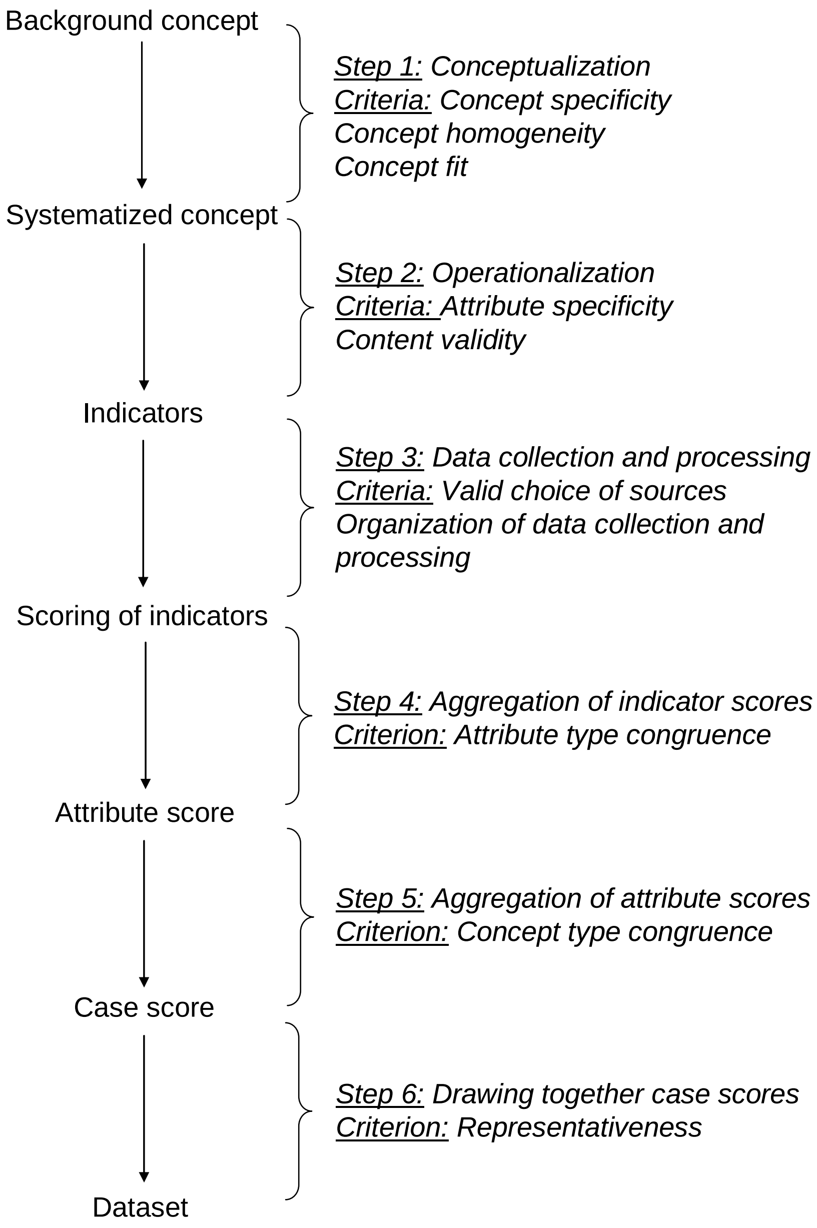
[7] These criteria are not unique for macrodata inasmuch as they can be equally well applied to mesodata. In principle, these dimensions are also relevant for microdata, as the extensive discussion about the equivalence of indicators in different cultural contexts (content validity) (Hambleton and Patsula 1998; King et al. 2003) and sample selection techniques (use of sources) exemplify (Deininger and Squire 1996; Duncan and Hill 1989). Nonetheless, I will only refer to macrodata in the following because there are some specific issues involved that have received little attention in the past.

[8] The straight sequence of steps in my guideline should not imply that the generation of data is as straightforward. On the contrary, the collection of data, and the conduct of research in more general, often involves moving back and forth between different tasks {cf., \Ragin, 2000 #473}. For example, the understanding of a concept may profoundly change once one performs field research, thus requiring going back to the second step and modify the underlying definition. In a stylized perspective, however, one can order the order the tasks as done in figure 1.

[9] The background concept is similar to what Goertz (2006) calls the first level of a concept. The systematized concept is similar to the second level.

by a proper organization of the data collection process and internal quality controls by the data producer.

Figure 1: A unified perspective on data quality


Background concept
Step 1: Conceptualization
Criteria: Concept specificity
Concept homogeneity
Concept fit
Systematized concept
Step 2: Operationalization
Criteria: Attribute specificity
Content validity
Indicators
Step 3: Data collection and processing
Criteria: Valid choice of sources
Organization of data collection and processing
Scoring of indicators
Step 4: Aggregation of indicator scores
Criterion: Attribute type congruence
Attribute score
Step 5: Aggregation of attribute scores
Criterion: Concept type congruence
Case score
Step 6: Drawing together case scores
Criterion: Representativeness
Dataset

After having scored indicators, the fifth step requires the *aggregation* of indicator scores to a single attribute score. The criterion that is at stake here asks for the attribute type congruence of the aggregation. It is achieved when the aggregation of indicator scores conforms to the way the attribute was operationalized in step two (cf., Goertz 2006, ch. 4). A similar perspective is taken in the next step where one needs to check whether the aggregation of attribute scores to case scores conforms to the way the concept was specified in the first step.

Finally, one takes a comprehensive view on the whole dataset and determines its *representativeness* and suitability for generalization. This requires, first, to assess whether it represents a sample and, if so, whether it is biased or random. In addition, it is mandatory to determine the extent and nature of missing data for the units included in the dataset.

In the remainder of this paper, I will discuss each of the criteria in more detail. The implications of low-quality data for measurement and causal inference, which is at the heart of all treatments of data quality {Adcock, 2001 #979}, will be discussed insofar as they are specific to the criterion under scrutiny. Beyond these particularities, violating a criterion belonging to the first six tasks *potentially* results in measurement error.[10] In contrast to how this issue is generally treated in the literature, I argue that low-quality data does not necessarily create measurement error. Explaining this ambiguous link between data quality and the scoring of cases requires a discussion of concept formation and aggregation, so I leave this issue for a later section.

4. Consistency of conceptualization

4.1. Concept specificity

The clear definition of the subject one is interested in – peace, capitalism, etc. – is important because of two reasons. Substantively, it communicates the understanding of a concept to other researchers and promotes the discourse about different conceivable definitions. Methodologically, the specification of a concept has far reaching implications for all subsequent stages of the measurement process (Adcock and Collier 2001). The need for an explicit definition of a concept may seem so obvious that it does not deserve to be treated as an own criterion. However, available conceptualizations often are less transparent than it may seem at first sight. Definitions frequently lack concept specificity in terms of an explication of the concept’s *attributes* and the *links* that are in place between them.

The definition of governance that underlies the *World Governance Indicators* (WGI) can be taken as an example for suboptimal concept specificity. At one place, the WGI’s documentation says that governance is “*the process by which governments are selected and replaced, the capacity of the government to formulate and implement sound policies, and the respect of citizens and the state for the institutions that govern economic and social interactions among them*” (Kaufmann, Kraay, and Mastruzzi 2004, p. 254). Later on, six clusters of governance are introduced, one of which is “political stability and the absence of violence” (Kaufmann, Kraay, and Mastruzzi 2004, p. 254). The absence of violence is not

[10] Task seven calls for the analysis of the whole dataset, implying that the scoring and mismeasurement of individual cases do not matter.

explicitly part of the first definition, leaving the data user somewhat uncertain about whether "political stability and the absence of violence" is considered an attribute of good governance or not. Similar ambiguities become apparent when one compares the other five clusters with the attributes contained in the definition presented above {cf., \Rohlfing, 2008 #3500}. In the face of such conceptual vagueness, it is useful to examine the data producer's choice of indicators and to discern what they are measures of. The rationale for this strategy is that the employed measures represent the (explicit or implicit) understanding of the underlying concept. For this reason, one can perform a conceptual backward induction from the selected indicators to the attributes that they actually measure and choose the conceptualization achieving the highest degree of content validity (see below) (cf., Goertz 2006, ch. 4).[11] In case such a conceptualization is not offered by the data producer, one needs to develop a proper conceptualization oneself.

The second important subdimension of concept specificity concerns the verbal linkages between attributes in terms of "and" and "or". One can for example say that democracy requires free elections *and* political liberties. Goertz (2006, ch. 2) speaks of *necessary/sufficient* conceptualizations when "AND" is in place between the attributes.[12] For reasons of convenience and so as to use a more intuitive label, I will call necessary/sufficient concepts AND-concepts in the following. Substantively, the AND-perspective on democracy means that a country needs to hold free elections and grant full-fledged political liberties in order to qualify as democratic.[13] If one of the two attributes is missing, the country is not democratic irrespectively of how it performs on the other dimension. On the other hand, a concept is of the *family resemblance* type when the attributes are related to each other by "or" (cf., \Collier and Mahon 1993). I label this concept type OR-concept in the remainder of this paper. An OR-concept exclusively subsumes sufficient attributes, meaning that an empirical referent only needs to meet one condition. Returning to the example just given, a country can also be said to be democratic when it holds free elections or grants political liberties.[14] In this view on democracy, a country qualifies as democratic if it holds fair elections or allows for political liberties. If one of these two attributes is present, it does not matter for the classification of the country how it performs on the other dimension. The substantive difference between the AND- and OR-perspective on democracy is obvious. In addition, the

[11] Applying this strategy to the WGI shows that the six clusters represent the real understanding of governance.
12 It is custom to capitalize the verbal "and" and "or" in order to denote that it is meant as the logical AND used in Boolean Algebra (cf., Ragin 2000), which matters for the aggregation of scores.
[13] The attributes are individually necessary and jointly sufficient, thus "necessary/sufficient".
[14] One can also think of mixed concept structures in which a subset of the attributes is necessary, but not sufficient. It is not a problem to deal with hybrid concepts. However, I do not address them in order to keep the discussion comprehensible.

linkage between attributes has implications for the aggregation of attribute scores to case scores. Since the choice of the aggregation technique is a quality criterion of its own, I will reserve the discussion of this point for a later section.

4.2. Concept homogeneity

Concept homogeneity denotes that the same definition is applied across units of analysis and time (Przeworski and Teune 1966). Concept homogeneity is important to achieve in order to make the data comparable on the cross-section and time-series dimension. A case in point for inconsistent definitions are national unemployment rates. Many countries in the world today follow the definition of the International Labor Organization, which divides the adult population into people that are employed, unemployed, and inactive. A lack of concept homogeneity on the cross-section dimension derives from different understandings of what an employed, unemployed, and inactive person is (Brandolini, Cipollone, and Viviano 2006). The United States, for instance, treat persons as inactive that haven’t applied for a job during the last four weeks. Canada, on the other hand, grants a longer period of time for applying for a job before classifying a person as inactive. Ceteris paribus, it follows that Canada’s unemployment rate will always be higher than the one of the United States (Sorrentino 2000). An example for a time-series inconsistency is a labor market directive of the European Commission from 2000. The directive defines unemployment more restrictively as compared to the years before 2000. The new definition caused a decline in official unemployment figures because many people that were previously classified as unemployed were now considered inactive (Garrido and Toharia 2004).

If concept homogeneity is low, the best way to increase it is to go back to the raw data and to generate measures that are comparable across time and/or countries. A problem may be that the raw data does not allow one to produce comparable figures. There is no way to retrospectively calculate the share of people that haven’t applied for a job during the last eight weeks when one has only asked for applications made during the last four weeks. If the appropriate raw data is unavailable, one can instead statistically adjust the data according to the different definitions. This is actually the way that was taken in the field of unemployment (Sorrentino 2000). The statistical adjustment procedure tries to make scores comparable by taking the differences between various definitions of unemployment into account. One problem of this remedy is that there will always remain some uncertainty about how definitional discrepancies are best translated into adjustment procedures. Moreover, one has to decide in the first place about what definitional differences one should try to adjust for. With respect to unemployment data, it has been criticized that some essential aspects were ignored in the adjustment process and that the adjusted scores are misleading (Sorrentino 2000). In the last end, the adjustment of scores involves a couple of arbitrary decisions that introduce an inherent degree of uncertainty into the comparability of the new scores.

4.3. Concept fit

The third criterion on the conceptualization dimension concerns the maximization of *concept fit*, which calls for the inclusion of all relevant attributes and the exclusion of all irrelevant attributes from the definition (Sartori 1970).[15] The applicability of this criterion depends on whether it is possible to unambiguously identify the core meaning of a background concept (Adcock and Collier 2001, pp. 538-540; Munck and Verkuilen 2002, pp. 7-12). For example, research on governance suffers from a lack of shared understanding of what governance is (Kaufmann, Kraay, and Mastruzzi 2007b). Such a lack of agreement makes it difficult to say whether a specific definition is too narrow or too broad or both. However, even when there is disagreement about the appropriate conceptualization, there most often is a certain degree of consensus about some attributes that should be excluded and included. There is widespread consent, for instance, that social justice is not a defining characteristic of democracy (Munck and Verkuilen 2002, p. 9).

The maximization of concept fit must be achieved on the basis of the current theoretical and empirical state of the art. Theory tells one what the relevant attributes of a background concept are and whether some attributes carry the same meaning and thus are redundant. The empirical state of the art is equally important because the qualitative in-depth analysis of cases allows one to assess whether a specific attribute should be subsumed under a concept or not (Adcock and Collier 2001, p. 539). For example, tariff levels are only a partial measure of protectionism because countries also impose non-tariff barriers (NTB) to trade. The close empirical analysis of different states would be suitable to identify protectionism through NTBs, which would result in a definition of protectionism that includes tariff levels and NTBs.

5. Consistency of operationalization

5.1. Attribute specificity

The equivalent to concept specificity in the operationalization stage is *attribute specificity*. Attribute specificity is given when the data producer details what indicators were selected and how they are related to each other. Making the operationalization of each attribute transparent

[15] The problem of redundant attributes (Munck and Verkuilen 2002, p. 8) is similar to including an irrelevant attribute.

is relevant for several reasons. First, it may be that indicators are redundant. An example for redundant indicators is the measurement of government effectiveness through the *World Governance Indicators* (Kaufmann, Kraay, and Mastruzzi 2004, p. 255). Two of the indicators measuring this attribute are the quality of public service provision and the quality of the bureaucracy. These indicators are redundant because it is not obvious in what way the quality of public service provision should be different from the quality of the bureaucracy whose objective is to provide public services.

In addition, attribute specificity demands it to detail how the indicators are related to each other (presuming that the attribute is measured through multiple indicators). As is the case for attributes, one can link indicators to each other through an “AND” or “OR”. For example, the World Governance Survey measures the socializing dimension of governance, capturing the way citizens raise and become aware of public issues, through five indicators. The indicators are freedom of expression, freedom of peaceful action, freedom from discrimination, opportunity for consultation, and public duties (Hyden and Court 2002, p. 31). The indicators are connected through an “AND”, turning the socializing dimension into what I call an AND-attribute.

It is equally possible to consider indicators as substitutable and to relate them to each through an “OR”, in the case of which one is dealing with OR-attributes. One can say that one attribute of a developed welfare state are public services that are specifically aimed at children. Whether these services are provided in the form of child benefits or day care is not important. This implies that a country only needs to pay child benefits or maintain daycare in order to qualify as a developed welfare state with respect to this attribute. Again, the attribute type matters for the aggregation of indicators scores to attribute scores, which will be discussed later in this paper in the section of aggregation.

5.2. Content validity

An indicator exhibits content validity when it is an adequate measure of the attribute under which it is subsumed (Carmines and Zeller 1979, pp. 20-22). A lack of content validity arises when an indicator is either too broad or too narrow. Measurement of participation, for example, should not be based on voter turnout because voting is mandatory in some countries. Moreover, people may feel obliged to cast their vote in authoritarian countries in order to avoid repressive measures (cf., Bogaards 2007). Thus, turnout probably overestimates the extent of participation in these countries. On the other hand, electoral participation is witnessing a decline in many democratic countries (Dalton and Wattenberg 2000). At the same time, people get engaged in civil groups pursuing specific interests like environmental protection. In this view, turnout may also underestimate participation, making the indicator inappropriate for the measurement of this concept.

Solutions to the problem of missing content validity are *context-specific* and *adjusted indicators* (Adcock and Collier 2001, pp. 535-536). The former refers to the use of different indicators for the scoring of different cases (Przeworski and Teune 1966, pp. 555-565; van Deth 1998). Voter turnout could be a measure of participation in those countries for which one can credibly argue that it is a valid indicator. Another indicator is then selected for countries for which turnout is assumed to be misleading. The choice of context-specific indicators should be based on theoretical reasoning in combination with the close analysis of cases in order to get an idea of what an indicator measures in a specific context. Another way of achieving content validity is the adjustment of the original scores. With respect to participation, one could estimate a lower turnout for those countries in which voting is compulsory. As discussed above, the substantial problem with this technique is that there are no clear-cut rules about how to adjust the observed score. The adjusted score probably is closer to the true score, but some measurement error is hard to avoid.

6. Consistency of data collection and processing

6.1. The choice of sources

The actual scoring of indicators requires the gathering and processing of raw data. The extent and nature of information that is needed depends on the concept one aims to measure. The scoring of countries on a federalism-unitarism dimension only calls for the analysis of whether the subfederal level enjoys autonomous policy-making capacity. Such information can be easily found in a country's constitution or related documents regulating the policy-making authority between territorial levels. On the other hand, it is much more difficult to obtain a valid figure of a country's involvement in major cross-border incidents (Azar et al. 1973). Not all events are documented in easily accessible sources and, moreover, it may even be difficult to *examine* these sources because of limited resources (Thies 2002). In the following, I decide for taking a comprehensive perspective on data collection and processing in order to fully detail the pitfalls that are linked to this task. Which of the issues to be discussed in this section apply in a particular constellation depends on the concept and cannot be addressed in the abstract here. Thus, it is up to the data user to carefully examine the aspects that are at stake given the own research context.

The use of raw data for the scoring of cases potentially suffers from a *source coverage problem* (Azar et al. 1973) that has three facets: *source sampling*, *source sampling bias*, and *low intersource reliability*. Source sampling is present when the information a source contains represents a sample of all the information that is related to an empirical phenomenon, which is likely to be the rule (Smith 1969; Thies 2002).[16] The analysis of a subset of information is a point of concern when one is interested in the magnitude or number of events inasmuch as they tend to be underestimated. Assume that you aim to measure the democratic quality of countries by counting the number of violations of democratic principles that are reported in a certain newspaper (cf., Bollen, Entwisle, and Alderson 1993). It is rather likely that not all oppressive measures are reported somewhere because the incumbent regime controls the local media and the flow of information to some degree at least.[17] For this reason, a count of newspaper reports will probably yield a figure that is too low.[18] Auxiliary knowledge about how and what information makes it into the newspapers may allow one to conclude that the obtained information and the corresponding democracy score is the lower bound of the true score. However, in many cases one does not know the size of the population, which is a marked difference to ordinary survey research where the population size is generally known (e.g., the number of inhabitants of a country). Thus, it is relatively difficult to estimate the uncertainty of the figures derived from a sample of primary and secondary sources, which are often used for the generation of macrodata.

The second dimension of the source coverage problem is closely related to the first one and concerns source sampling bias. The presence of a source sampling bias is important to determine when one is interested in the *forms* an event takes because some types of events may be more likely to be encountered in a source than others. Moreover, a bias may undermine the calculation of measures of central tendency like the sample mean. In the realm of primary sources, this is particularly a problem if some documents are kept secret inasmuch as these can be expected to be systematically different from publicly available sources. A source bias also concerns secondary sources like historical books that take a specific, biased perspective on a subject matter (Lustick 1996). A similar problem pertains to newspapers that do not report events of minor importance (Smith 1969). With respect to democracy research, for example, serious violations of democratic procedures are more likely to be documented

[16] Source sampling refers to the collection of sources for the scoring of one case, whereas the traditional sampling bias asks for how a group of cases were selected from the population (cf., Collier and Mahoney 1996). It is conceivable that source sampling undermines the scoring of cases, but that the set of scored cases is a random sample of all cases.

[17] Moreover, it is likely that the nature of the reported oppressive measures are systematically different from the non-reported ones, since major violations of democratic principles are more likely to make it into the news. This is not important here because we are just interested in the magnitude of an event and not the forms it may take.

18 In rare instances, the selected sources may also lead one to overestimate the magnitude. O’Kane (1993) shows that data on coups d’états suffer from an upward bias because political leaders have an interest to fake a coup so as to strengthen the domestic power position.

than minor incidents.[19] As is the case for source sampling, the careful qualitative assessment of the source selection rule is a suitable means to get an idea of the presence and extent of a source selection bias.

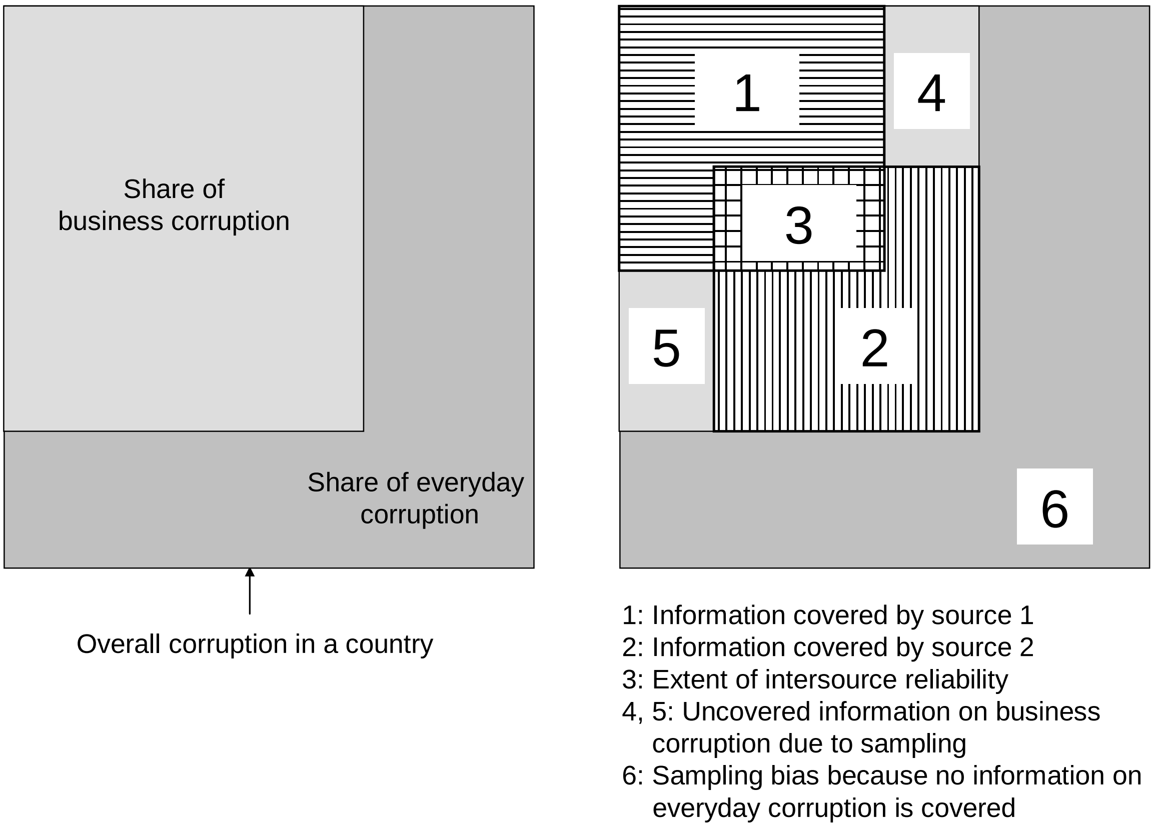
The third dimension of the source coverage problem asks for the extent of *intersource reliability*. Intersource reliability captures the degree to which different sources contain the same information (Azar et al. 1973; Jackman and Boyd 1979; Sommer and Scarritt 1999). Just like a source may only cover a subset of all relevant information, different sources might capture divergent parts of the population of information. The higher intersource reliability is, the less the choice of a specific source matters for the data one will generate. For example, version six of the *Worldwide Governance Indicators* (WGI) draws on some sources that were not used for the construction of earlier versions. At the same time, some sources drawn on for previous compilations of the WGI were not utilized anymore for version six. The founders of the WGI recognize that this may be problematic because the WGI scores one obtains may depend on the raw data one uses. In this case, however, the changes in the sources are not an issue because the old and the revised indicators correlate strongly with each other (Kaufmann, Kraay, and Mastruzzi 2007a, pp. 9-10).

One cure to the problem of low intersource reliability may seem obvious, namely, to pool the information gathered from different sources (Bowman, Lehoucq, and Mahoney 2005; Lustick 1996). Pooling raw data is superior to the use of a single source because the joint set of information necessarily covers a larger share of the population. In absolute terms, however, it may be that even a pool of raw data only represents a small subset of the population. For a similar reason, high intersource reliability does not tell anything about the validity of each source. High reliability merely denotes that the sources cover the same information, which may just be a small and biased sample of the population. A case in point for this problem are democracy indices that draw on similar raw data like news reports for the scoring of countries. The fact that some democracy indices correlate high with each other is a sign of high reliability, but dubious validity because they all suffer from similar source sampling problems and biases as described above (Bollen 1993). An additional practical issue with this instrument is that it may be infeasible for extensive datasets insofar as it is unlikely that one is able to draw on multiple sources for each country-year. Figure 2 summarizes the discussion of

[19] In the literature on the source coverage problem, an additional point concerns the measurement of subcategories of a variable (Doran, Pendley, and Antunes 1973). Subcategories represent a multicategorical definition of the background concept. One distinguishes, for instance, between types of democracies instead of simply differentiating democracies from non-democracies (cf., Collier and Levitsky 1997; Goertz 2006, pp. 119-120). Thus, subcategories are nothing special and can be easily accommodated in my framework in the conceptualization and operationalization stage.

source sampling, sampling bias and inter-source reliability by the example of the measurement of corruption.

Figure 2: Source sampling bias and inter-source reliability[20]


Share of
business corruption
Share of everyday
corruption
Overall corruption in a country
1
4
3
5
2
6
1: Information covered by source 1
2: Information covered by source 2
3: Extent of intersource reliability
4, 5: Uncovered information on business
corruption due to sampling
6: Sampling bias because no information on
everyday corruption is covered

The figure on the left represents the corruption that takes place in a country. For reasons of convenience, I assume that corruption can be divided into two subtypes: business corruption (light gray), related to economic activities, and everyday corruption (dark gray), which citizens experience in relation to the bureaucracy, doctors, etc. The goal is to measure the magnitude of the overall corruption in a country as well as the extent and share of its two variants. In order to achieve this goal, one draws on two national newspapers (shaded with horizontal (source 1) and vertical bars (source 2). The right figure shows that the information they provide on business corruption covers a good extent of all business corruption that is going on. Assuming that the non-covered information (zone 4 and 5) is not systematically different from the raw data one obtains, the loss of information about business corruption can be attributed to source sampling. The effect is a modest underestimation of the true size of business corruption and, even if one would perfectly measure everyday corruption, the overall extent of corruption. The relatively small overlap between the two sources (area 3) shows that

[20] The assumptions underlying this figure are that one uses two sources and that neither of the two overestimates the true number of events, which, as I explained above, may be the case.

the intersource reliability is low. This may prompt the researcher to question the utility of the two sources for her purposes. In this hypothetical example in which we have full information, however, one can see that the low degree of reliability should not be a point of concern because altogether the two newspapers provide a good deal of information about business corruption. The more serious problem is that the sources do not coverage everyday corruption at all (area 6). This means that one will completely fail to measure this subtype and, as a consequence of that, strongly underestimate its overall size.

The available literature on the use of sources conceives of the source coverage problem as one of subjective data. This view is incomplete inasmuch as there are also problems in using objective data (Russett 1966, p. 98; Scheuch 1966, p. 139).[21] For example, Greece provided information about its public debt in the years preceding its accession to the European currency union in 2001 that was systematically too low. The aim to face lift objective data does not necessarily require faking it because it is often less objective than it seems. The GDP, for instance, is considered an archetypical case of objective data. It is, however, a composite measure of objective data and an estimated component. This component covers, among other things, the extent of illicit employment. Again, Greece can serve as an example insofar as it increased its preliminary estimate of its GDP of 2005 upward by 25 percent. The substantial rise in the GDP was explained with a revised estimation of the subjective share of the GDP that was generated through smuggling, bribery and other activities for which objective data was unavailable.

The collection of multiple sources and pooling of information is mostly infeasible for objective data, since it is generally provided by one official source. Yet, there are two other instruments that can be employed. First, one can try to assess of how the raw information was collected and in what direction the source coverage problem may work (George and Bennett 2005, ch. 5; Lustick 1996). This is a time-consuming endeavor when the dataset has broad coverage, but it is indispensable if one wants to get an approximate idea about the quality of the data. A careful consideration of the employed sources seems to be rarely done by data producers. For instance, constructors of democracy indices often do not consider whether the sources they draw on may yield a biased picture about the true state of democratic development of a country (Bollen 1993; Bowman, Lehoucq, and Mahoney 2005).

Second, one can draw on the expertise of country experts and ask them for an assessment of the data and the sources from which it was derived (Bowman, Lehoucq, and

[21] In my context, subjective data is data that is generated by an expert, for example a democracy score for a country-year on the basis of various reports about this country. Objective data captures information that is believed to be “hard” like the GDP.

Mahoney 2005). The increase in measurement validity one will obtain by the close inspection of cases may come at the expense of a declining coverage of the dataset. It is unlikely that one will be able to find enough experts so as to maintain the dataset's time-series cross-section extension if it is too extensive. This means that one faces a breath-depth trade off when deciding about whether to solely rely on easily accessible sources or to additionally draw on the knowledge of country experts. Moreover, it should be taken into account that country experts may have their own biases in assigning scores to cases, which is a central point in the following section.

6.2. The organization of data collection and processing

Besides the assessment of the data quality, a data user should also determine whether and, if yes, what quality controls the data producer had in place when generating the data. One specific risk involved in using subjective data is that the assignment of scores to indicators by experts is biased. Biased experts create so called *method factors* because of which the generated scores are systematically different from the true scores (Bollen and Paxton 1998).[22] One reason that was found to be influential in the production of democracy scores is the personal familiarity of the expert with the cases. Other sources of method factors may be the political affinity of an expert or her dependency on a funding institution, which would like to see that some cases are assigned better or worse scores than they would receive from an independent expert (Bollen 1993; Bollen and Paxton 2000; Herrera and Kapur 2007).

Data users can get a hint at the presence of method factors by assessing the organization of the measurement process. The probability of systematic measurement error is high when there is only one expert and no reliability tests or other quality checks. Conversely, confidence in the quality of the scoring procedure can be higher when multiple experts are involved and when internal quality checks are performed. However, the validity of measurement is still in danger because all experts may display the same bias. Such a situation occurs, for example, when all experts are dependent on the funding of another institution whose interests interfere with the proper scoring of cases. A more sophisticated way of examining the presence and extent of method factors is the use of structural equations models. They can be used to estimate the source, extent, and direction of the subjective bias (see for an application Bollen 1993 and Bollen and Paxton 1998). Moreover, it may be possible to determine specific cases suffering from method factors and try to correct for it (Bollen and Paxton 2000). However, the application of these models for the analysis of method factors

[22] Broadly seen, method factors also include the inappropriate use of sources as discussed in the previous section (Bollen 1993, pp. 1212-1214). In this section, I only discuss some factors that may occur even when there are no problems with the collection of sources.

quickly becomes intricate and statistically demanding, so one should be careful in using these tools and avoid their mindless application (Bollen 1993). Even if the experts are completely unbiased, some safeguards are necessary because the raw information may open room for interpretation. Different coders probably come to different coding decisions on the basis of the same sources, thus decreasing the reliability of the scoring procedure (Bollen 1993; LaPalombara 1968). For instance, it may be ambiguous to classify a certain type of cross-border incident as noteworthy or negligible. Because of this, the development of high-quality datasets always demand internal controls of the reliability and validity of the generated scores.

7. Consistency of aggregation

After having assessed of how scores are assigned to indicators, one needs to determine how these scores are aggregated.[23] The generation of case scores through the aggregation of indicator scores is a two-step involving two closely related criteria. In the first step, the aggregation of indicator scores to attribute scores needs to satisfy the criterion of *attribute type congruence* (presuming that one is dealing with multi-indicator attributes). Attribute type congruence is given when the aggregation technique conforms to the rule the attribute type calls for, since the type of attribute determines what the viable aggregation procedure is. The minimum-scoring indicator should be taken as the attribute score under AND-attributes, while the maximum-scoring indicator matters for the scoring of OR-attributes (Goertz 2006, chs. 2 & 4). In the second step of the aggregation stage, the attribute scores can be aggregated to a single case score. Depending on the research question and subject matter, one may decide not to perform this step (Munck and Verkuilen 2002, pp. 22-26).[24] If one decides to aggregate, however, it is mandatory to achieve *concept type congruence*. Similarly to attribute type congruence, this criterion is met when the aggregation of attribute scores is congruent with the nature of the concept in terms of AND- and OR-concepts.[25]

A non-social science example helps to understand why the concept and attribute type determines how to aggregate indicator scores. Assume you are interested in the “health of a person”, one attribute of which is the absence of a fatal disease. Another attribute could be the

[23] Contrary to what one may believe, the aggregation of scores also matters for genuine macrodata. Assume you are interested in whether a country is decentralized or centralized (cf., Hooghe and Marks 2003). The autonomy of regional entities is measured through their financial and legislative autonomy that are operationalized through one indicator each. In this instance, one needs to aggregate the indicator scores so as to know whether the country in question is centralized or decentralized. Recall that a macrovariable is genuine when it cannot be aggregated from the behavior of actors on the micro- and mesolevel. Thus, the need to aggregate indicator scores is fully compatible with the notion of a genuine macrovariable.

[24] For instance, the World Governance Indicators provide information for six dimensions of governance for each country-year because of the data producers’ belief that it is not meaningful to generate case scores (Kaufmann, Kraay, and Mastruzzi 2007a).

[25] As mentioned before, it is possible to follow other aggregation techniques when the concept and its constituent attributes are appropriately defined and operationalized.

absence of chronic illnesses. “Absence of fatal disease” is operationalized by not having cancer, HIV, and other forms of serious diseases. More precisely, “absence of fatal disease” is an AND-attribute, since one should not have cancer and HIV and other forms of diseases for receiving a score of one on this attribute. This means that if one has a score of one on the indicator “cancer”, it does not matter that one does not have HIV and all other fatal diseases. In such a constellation, no one would say that fatal disease are mostly absent simply because one “only” has cancer. Similarly, “health” is an AND-concept because having a deadly disease is essential for being healthy. Again, it would be strange to say that one is mostly healthy because one just suffers from one fatal sickness and does not have chronic diseases too. Therefore, one would say a person with a deadly disease is not healthy, that is, one would assign a score of zero and not a score that is close to one because of a lack of other diseases. The reverse logic can be applied to OR-concepts and OR-attributes when the concept is “sickness” and the attribute “presence of fatal diseases”. Cancer is sufficient for having a deadly disease, independently of whether one does not have other forms of serious illnesses. In a similar vein, cancer is sufficient for being sick, notwithstanding that one may not have chronic diseases.

The same logic of aggregation applies to social science concepts. When one says that a country has high military capabilities when it has nuclear weapons *or* a large conventional force, then it suffices to receive a positive score on one indicator (assuming the measures are dichotomous). The score on the other indicator does not matter at all). The situation is different when one operationalizes high military capabilities through a nuclear force *and* a large conventional army. In this instance, a country with no nuclear weapons cannot qualify as having strong capabilities even if the number of troops is very large and vice versa. For the reasons mentioned in the non-social science example above, it would be a clear case of inconsistent aggregation to say that this state has moderate capabilities simply because it owns a strong nuclear force.

With respect to existing social science datasets, Goertz (2006, ch. 4) shows that the *Polity* data measuring the democratic quality of countries lacks concept type congruence. The *Polity* index adds the scores of five attributes, which contradicts the underlying AND-conceptualization of democracy that requires taking the minimum-scoring attribute. The *World Governance Survey* (WGS) lacks attribute type congruence as well as concept type congruence. The WGS concept of governance is of the AND-type, since it comprises six attributes that are all considered essential elements of good governance. The attributes are of the AND-type as well and are operationalized through five indicators each. Since the minimum rules on the level of indicators and attributes, it follows that the case score should be equal to score of the minimum indicator. However, the WGS score for a country-year is the sum of all indicator scores, thereby producing attribute type and concept type incongruence of aggregation. When one ranks the countries according the attribute sums and the minimum-scoring attribute, the rank correlation (Kendall's tau) is just .7 for the year 2000 and .65 for 1995, which are the two years for which governance scores are available. This example underscores the importance of aggregating indicator scores properly.

It is occasionally recognized in the literature that proper aggregation is an important element of data quality (e.g., Munck and Verkuilen 2002). However, it is mostly neglected that the aggregation rule must perfectly correspond to the type of attribute and concept, thus limiting the number of viable aggregation techniques to one (Goertz 2006, ch. 4). A case of lacking attribute type congruence is the aggregation of indicators underlying the World Governance Survey (WGS). The WGS conceptualizes governance through six attributes, each of which is measured through five indicators. The WGS adds the indicator scores instead of taking the minimum indicator (Court, Hyden, and Mease 2002, pp. 11-12).

When a dataset lacks attribute and/or concept type congruence, one can adjust the concept and attribute type to the applied aggregation technique or vice versa. With regards to the *Polity* data, this means that one needs to modify the underlying definition of democracy if one believes that adding attribute scores actually reflects the way democracy should be defined. This could be achieved by saying that the democratic quality of a country is determined by its joint performance on the five dimensions of democracy that underlie the *Polity* data.[26] The important difference to the current definition lies in the qualification "joint performance", which can be translated in an aggregation rule that adds attribute scores. On the other hand, one can create concept type congruence by leaving the conceptualization intact and taking the minimum of all five attributes as the country score. This strategy makes it necessary to get a hand on the non-aggregated data so as to determine the minimum-scoring attributes. If the raw data cannot be obtained (which is not the case for the *Polity* data), one needs to determine the consequences of the inappropriate aggregation technique. This point is related to measurement and aggregation under different concept and attribute types and is left for the section after the next one.

[26] The dimensions are competitiveness of political participation, regulation of political participation, competitiveness of executive recruitment, openness of executive recruitment, and constraints on chief executive.

8. Representativeness

The criteria that I have discussed so far are concerned with assigning scores to individual cases. In the seventh part, it is necessary to broaden the view and to assess the representativeness of the generated dataset. Taking a comprehensive perspective is necessary because resource constraints and limited access to sources often make it impossible to generate scores for all cases. As a consequence of that, one has to deal with *missing data*. Missing data can manifest itself in two different ways that are not mutually exclusive. First, some units for which one aims to make causal inferences are not included in the sample (Hug 2003). Second, there are missing observations for units that are included in the dataset (Allison 2001).

The first variant of the missing data problem can only be in place when the data user aims to generalize beyond the cases included in the dataset. Thus, whether a dataset represents a sample and, if so, whether it is randomly drawn cannot be assessed independently of the research interest a data user has.[27] For example, Crepaz and Moser (2004) select 15 OECD member states for their analysis of the impact of veto points on public expenditure. The choice of OECD countries seems to be correlated with the dependent variable. [28] High-expenditure countries are overrepresented and low-expenditure countries are more frequently missing, which implies that the selection rule involves a bias.[29] This would not be a problem if Crepaz and Moser would not generalize beyond their set of countries. Unfortunately, they do neither discuss whether generalization is intended, nor do they assess whether the extension of causal inferences is warranted in the light of case selection. However, it seems justified to infer that they intend to generalize because it is not explicitly ruled out. Moreover, they discuss their results by referring to the capacity of "the state" to redistribute incomes, i.e., their insights are not limited to the OECD members under scrutiny.

The quantitative analysis of biased sample is an issue because the results are of dubious validity. The problem is that the size and signs of the coefficients may be different from those they would have in an analysis of a random sample (Bartels 1995, p. 10).[30] Case study research is equally undermined by a sampling bias. The status of a case as typical and deviant depends on the distribution of cases in which it is embedded (Collier and Mahoney 1996). The relative nature of a case's status may undermine case selection because a biased

[27] The mere collection of data does not involve the generation of any descriptive or causal inferences. For this reason, the representativeness of a dataset can only be seen relative to claims a data user aims to make.

[28] The dataset covers Austria, Australia, Belgium, Canada, Denmark, Finland, France, Germany, Italy, Japan, The Netherlands, Norway, Sweden, the United Kingdom, and the United States.

[29] The bias is probably rooted in limited data availability for low-expenditure countries.

[30] These effects refer to bias on the dependent variable. Sampling bias on the independent variable is without problems in a regression analysis (Collier and Mahoney 1996).

sample is systematically different from a random sample. The more dissimilar the distributions of cases are, the more likely it becomes that a case has a different status in the two samples and that one chooses a wrong case given the own research interest (Rohlfing 2008).[31] It may be possible to roughly determine in what way the bias works, e.g., whether countries with high GDPs have a higher probability of getting selected. This may render it possible to make a qualified hunch about how case selection is affected. It is possible to make the same assessment in a regression study. However, the adverse and unpredictable effect of a bias on regression estimation remains, so qualitative knowledge about the nature of the bias is less useful in large-n analysis (Bartels 1995, p. 10).

Even when one analyzes the population or a random sample, one confronts a missing data problem when scores are lacking for the included units. Three types of missingness can be distinguished (Allison 2001, pp. 3-5; King et al. 2001, p. 2-3). Data is *missing completely at random (MCAR)* when the missingness on one variable is not systematically related to the scores units take on other variables. This means, for example, that countries with high and low GDP are equally likely to lack observations. Data is *missing at random (MAR)* when two conditions are given: first, data is systematically missing and, second, one knows of what nature the missingness is. In the case of GDP, MAR is in place when poor countries lack observations more often than wealthy countries and that one knows that missingness is systematically related to the level of the GDP or any another variable. Conditioning on a third variable would be in place when countries with high illiteracy rates lack data for the GDP more often than states with lower levels of illiteracy.[32] Finally, data can be *missing not at random (MNAR),* also called *non-ignorable (NI)* missingness. MNAR signifies that the missingness is systematic. In contrast to MAR, however, it is not predictable by conditioning on third variables. MNAR would be present if it would not be known that GDP data of low-GDP countries is missing more often than data for high-GDP states (Herrera and Kapur 2007).

There are three basic remedies for the missing data problem: listwise deletion, imputation, and running a statistical model taking the missing data problem into account (Allison 2001; Hug 2003; King et al. 2001; König, Finke, and Daimer 2005). The feasibility of each instrument hinges on the type of missingness. When data is MCAR, it is sufficient to delete cases listwise, i.e., cases for which data is lacking are dropped from the dataset. The

[31] Random sampling is not necessarily a solution because of sampling variability, which denotes that cases are differently distributed in random samples (cf., Gerring 2007, ch. 5).

[32] The relation between other variables and the variable with missing values does not have to be causal (King et al. 2001, p. 3).

more involved techniques of imputation and sophisticated statistical estimation are not necessary. On the other hand, listwise deletion may bias estimates in a regression analysis when data is MAR and MNAR. Imputation and statistical analysis are more appropriate in the face of these types of missingness. For this reason, one should be very careful in considering whether the data is most likely to be MCAR, MAR, or MNAR and what the most appropriate technique for the type of missingness at hand is. Since the discussion about the three remedies of missing data is as extensive as it is involved, I leave it with these general remarks and refer to the references for a more detailed discussion of missing data (cf., Allison 2001; Hug 2003; King et al. 2001; König, Finke, and Daimer 2005). Table 2 summarizes the discussion so far.

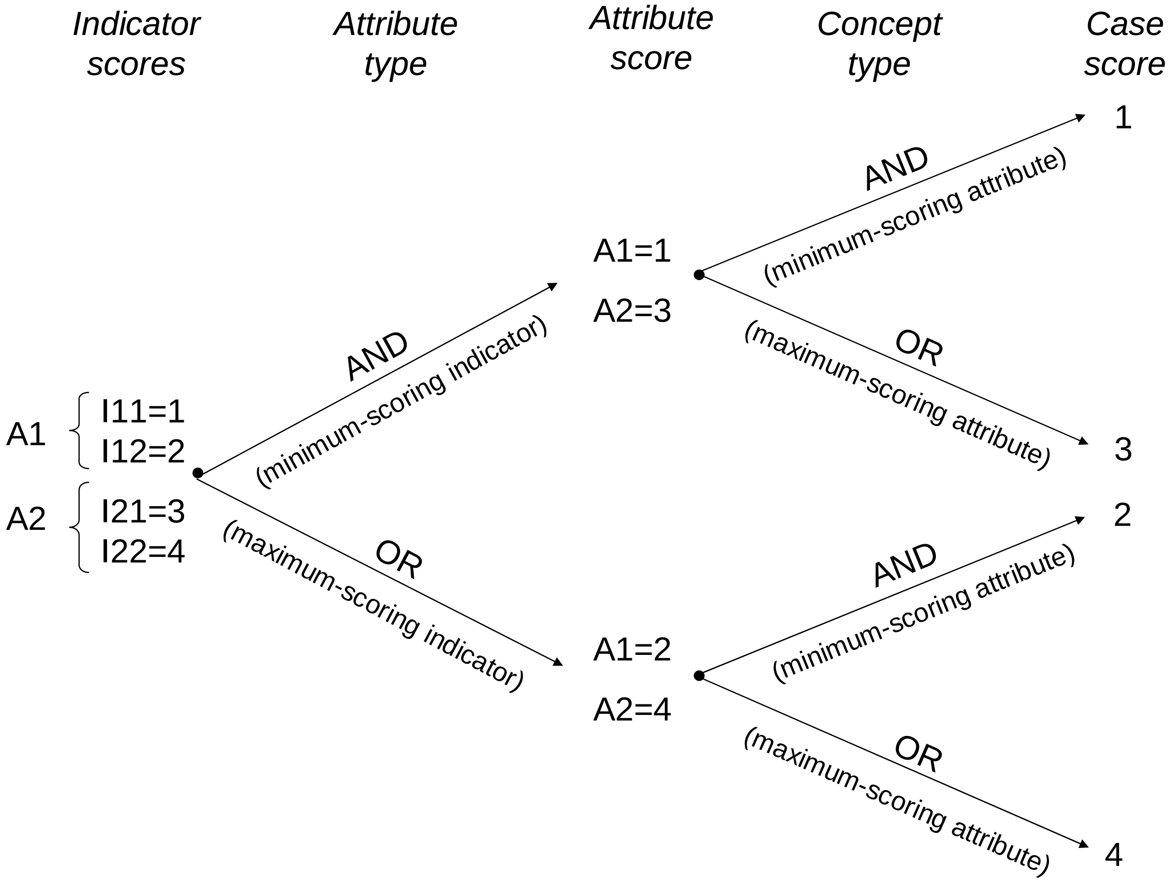
Table 2: A guideline for the assessment of data quality

Criterion	Question
Conceptualization	
Concept specificity	Are the attributes and the concept type unambiguously specified?
Concept homogeneity	Is the definition the same across time and units?
Concept fit	Are irrelevant attributes excluded and relevant attributes included?
Operationalization	
Attribute specificity	Are the indicators and the attribute types unambiguously specified?
Content validity	Do the indicators measure what they are supposed to measure?
Data collection and processing	
Use of sources	Do the sources cover a sample of or all information?
	If it is a sample, is it randomly drawn or biased?
	Are different sources used over time and across units? If so, are they substitutable?
Organization	Did the data producer install internal quality checks (reliability tests, etc.)?
Aggregation	
Attribute type congruence	Does the aggregation of indicator scores correspond to the attribute type?
Concept type congruence	Does the aggregation of attribute scores correspond to the concept type?
Representativeness	
Sampling	Are the included units a sample? If yes, is it biased or random?
Missing data	Is data missing for the included units? If so, is it MCAR, MAR, or MNAR? What remedy was taken by the data producer?

9. Measurement and measurement error

In the light of the previous discussion, it is now possible to detail the conditions under which the violation of one of the first six quality criteria may create measurement error.[33] For example, using indicators with low content validity may result in scores that are different from those one would obtain through valid indicators (Adcock and Collier 2001). In the following, I contend that contrary to what one may believe and how measurement is discussed in the literature, the violation of a quality criterion does *not* necessarily cause measurement error on the case level. The ambiguous link between data quality and measurement error is rooted in the dependence of the aggregation technique on the type of concept and attributes.

Figure 3: Types of attributes, concepts, and the scoring of cases


Indicator scores
Attribute type
Attribute score
Concept type
Case score
A1
I11=1
I12=2
A2
I21=3
I22=4
AND
(minimum-scoring indicator)
OR
(maximum-scoring indicator)
A1=1
A2=3
A1=2
A2=4
AND
(minimum-scoring attribute)
OR
(maximum-scoring attribute)
AND
(minimum-scoring attribute)
OR
(maximum-scoring attribute)
1
3
2
4

As explained above, concepts and attributes can be of the AND-type and OR-type. There is no inherent link between the concept and attribute types (Goertz 2006, pp. 50-53), implying that are four possible combinations: AND/AND, AND/OR, OR/AND, and OR/OR, where the first part refers to the type of attribute and the second element denotes the concept type.[34] A simple hypothetical example helps to clarify how and in what way the scoring of cases depends on

[33] The seventh criterion – the representativeness of the data – is not relevant in this context inasmuch as measurement error occurs on the level of individual cases.

[34] The attribute type is given first because the aggregation process begins with the generation of a score for each attribute.

the combination at hand. Assume you have two attributes A_1 and A_2 that are operationalized through the indicators I_{11}, I_{12} and I_{21}, I_{22}, respectively. I_{11} is scored 1, I_{12} equals 2, a score of 3 is assigned to I_{21}, and I_{22} receives a score of 4. Depending on the mix of the attribute and concept type, the score of the case is 1, 2, 3, or 4.[35]

The contingency of scoring on the concept and attribute type underscores the importance of being specific in the conceptualization and operationalization and in meeting the criteria of attribute type and concept type congruence of aggregation. Furthermore, the example shows that measurement error on the level of indicators and attributes does not automatically result in the wrong scoring of a case. This ambiguity can be understood by distinguishing between *indicator-related*, *attribute-related*, and *case-related* measurement error. Indicator-related mismeasurement occurs when one assigns a wrong score to an indicator, for example because of the source coverage problem. Imagine one assigns a score of 1.5 to indicator I_{11} and that A_1 is an OR-attribute. The indicator is mismeasured but does not produce inaccurate score for attribute A_1 because indicator I_{12} still represents the maximum. Now, assume that one measures a score of 2.5 instead of 1 on indicator I_{11}. The score on I_{11} then is the maximum and creates measurement error on the attribute A_1 too. However, the attribute-related measurement error will be without consequences on the case level when the concept is of the OR-type as well. The reason is that attribute A_2 remains the maximum-scoring attribute with a score of 4. This means that indicator I_{11} needs to receive a score larger than 4 in order to produce measurement error on all levels under an OR/OR-structure. Because of these complexities, it is far from straightforward to say in the abstract what the consequences of inferior data on the indicator and attribute level are for the scoring of a case.

The contingent scoring procedure has further implications for the identification of the extent and nature of the measurement error in the dataset. Different cases take different scores on the same indicators, implying that the effects of indicator- and attribute-related measurement error may vary from case to case. It is possible that all cases, a subset of them, or no cases at all are affected. Furthermore, it follows that it may be difficult, if not impossible to tell whether a measurement mistake on the indicator and attribute levels produce *systematic* or *random* measurement error or neither of both on the level of cases. This discussion should not imply that one can be less concerned about data quality inasmuch as case-related measurement error does not occur automatically. The correct measurement of a concept is always important because knowing the correct scores for indicators and attributes is

[35] For reasons of simplicity, I assume that the attributes are either of the AND- or OR-type.

a desirable goal of its own. Moreover, the generated data may be used for the measurement of a different concept in the future and that the mismeasured indicator/attribute then matters for the scoring of cases.

In addition, identifying the presence of measurement error and, if it is in place, its nature is of tremendous importance because of the adverse consequences for causal inference. In a *regression analysis*, non-systematic measurement error renders the estimators of a coefficient's variance inefficient. Systematic measurement error tends to be more problematic for regression analysis, whereas the precise effects depend on whether it occurs on the dependent or the independent variable and whether the included variables are correlated with each other and the error term. When measurement error on Y is uncorrelated with the regressors, the efficiency of the independent variables decreases. On the other hand, one will obtain a biased estimator for an independent variable if the measurement error on Y is correlated with this variable. The estimators of all independent variables are biased and inconsistent when one independent variable is mismeasured and correlated with all other independent variables, which is generally the case (Gujarati 2004, pp. 524-528).[36] These effects of measurement error on regression output can be ignored when the error is small. Since the magnitude of the measurement error is generally unknown, however, it is a dubious practice to assume measurement error away, as is regularly implicitly done in empirical research.

Systematic and non-systematic measurement errors also have implications for *Fuzzy-Set Analysis* and *Qualitative Comparative Analysis (fs/QCA)*, which has become an increasingly popular cross-case technique in the social sciences. The precise functioning of fs/QCA cannot be presented here (cf., Ragin 1987, 2000, 2008). What matters is that the solution one derives by minimizing the truth table through Boolean Algebra depends on how the cases are distributed across the rows capturing all logically possible configurations of scores on the causes of interest. Non-systematic measurement and systematic measurement error may render one row empty or may move cases to empty rows. As a consequence of that, the way the truth table is minimized and the solution one obtains may be different from the true solution.[37]

[36] The estimator of a coefficient is only biased when the independent variables are not correlated with each other (Herrera and Kapur 2007).

[37] It is regularly pointed out that fs/QCA should be build on the intimate knowledge of cases (Ragin 2000). This may diminish the probability that one uses indicators with low content validity when one constructs the dataset oneself. However, case knowledge does not save one from making mistakes in the other parts of the data development process like the aggregation of scores. Moreover, case knowledge may be limited when one uses an exiting dataset and is impossible to acquire on a broad basis when using a dataset containing hundreds or thousands of observations (Bennett and Elman 2006).

Finally, *qualitative case studies* are undermined by measurement error because the scores of cases on macrovariables provide the basis for case selection and causal inference (George and Bennett 2005, chaps. 8-9; Lieberson 1991; Lijphart 1971; Mill 1843 [1974]). Assume, for example, that one selects two cases for a method-of-difference design (cf., Przeworski and Teune 1970), i.e., the cases only differ on the outcome and one independent variable. This choice will be fallacious when the dependent variable of one case receives a wrong score because of measurement error and, once the measurement error is removed, the two cases display the same scores on the outcome.

Because of the far reaching implications of systematic and non-systematic measurement error on causal inference, it would be fallacious to ignore potential or existing deficiencies of existing datasets. In some instances it may be easy to identify the implications of a shortcoming. In other cases, however, it is much more difficult to determine whether and what kind of measurement error is present. For instance, an indicator having low content validity is not necessarily the maximum- or minimum-scoring indicator of all cases. In such a situation, the only feasible way is determine on a case-by-case basis whether the low-quality indicator affects the scoring of attributes and cases and what the aggregate picture is. Although this may be a time-consuming endeavor, this is the price one has to pay for the assessment and improvement of data quality.

10. Conclusion

In this paper, I developed general guidelines with which one can approach macrodatasets and determine their quality. Because of the increasing number of easily available datasets, it may happen that an inquiry in the quality of competing datasets yields an ambiguous picture. One dataset might perform better than others in some respect, while at the same time failing to meet another criterion. The data user then has to decide which of the criteria are the most important given the own research question, pick a dataset and try to correct for the adverse effects following from the existing shortcomings. Making such a decision requires much transparency and documentation on behalf of the data providers. For example, it needs to be reported what sources one used for assigning scores of what cases. Moreover, a transparent handling of the missing data problem makes it necessary to provide the raw data and not only the refined data that lacks any missing values. The data user will be unable to recognize the missing data problem when the data producer only publishes refined data and does not detail how the missing data problem was solved. Although some progress toward the transparency and improvement of datasets has been achieved in the past, more attention needs to be paid to data quality and more resources spent on the improvement of datasets.

References

Adcock, Robert, and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review* 95 (3):529-546.

Allison, Paul David. 2001. *Missing data*. Thousand Oaks: Sage.

Azar, Edward E., Stanley H. Cohen, Thomas O. Jukam, and James M. McCormick. 1973. The problem of source coverage in the use of international events data. *International Studies Quarterly* 16 (3):373-388.

Bartels, Larry M. 1995. Symposium on *Designing Social Inquiry*, Part 1. *The Political Methodologist* 6 (2):8-11.

Bennett, Andrew, and Colin Elman. 2006. Qualitative research: Recent developments in case study methods. *Annual Review of Political Science* 9:455-476.

Blasius, Jörg, and Victor Thiessen. 2006. Assessing data quality and construct comparability in cross-national surveys. *European Sociological Review* 22 (3):229-242.

Bogaards, Matthijs. 2007. Measuring democracy through election outcomes: A critique with African data. *Comparative Political Studies* 40 (10):1211-1237.

Bollen, Kenneth. 1993. Liberal Democracy - Validity and Method Factors in Cross-National Measures. *American Journal of Political Science* 37 (4):1207-1230.

Bollen, Kenneth A., Barbara Entwisle, and Arthur S. Alderson. 1993. Macrocomparative research methods. *Annual Review of Sociology* 19:321-325.

Bollen, Kenneth A., and Pamela Paxton. 1998. Detection and determinants of bias in subjective measures. *American Sociological Review* 63 (3):465-478.

Bollen, Kenneth A., and Pamela Paxton. 2000. Subjective measures of liberal democracy. *Comparative Political Studies* 33 (1):58-86.

Bowman, Kirk, Fabrice Lehoucq, and James Mahoney. 2005. Measuring political democracy: Case expertise, data adequacy, and Central America. *Comparative Political Studies* 38 (8):939-970.

Brandolini, Andrea, Piero Cipollone, and Eliana Viviano. 2006. Does the ILO definition capture all unemployment? *Journal of the European Economic Association* 4 (1):153-179.

Braun, Dietmar, Anne-Béatrice Bullinger, and Sonja Wälti. 2002. The Influence of Federalism on Fiscal Policy Making. *European Journal of Political Research* 41:115-145.

Carlsen, F. 2000. Unemployment, Inflation and Government Popularity: Are there Partisan Effects? *Electoral Studies* 19 (2-3):141-150.

Carmines, Edward G., and Richard A. Zeller. 1979. *Reliability and Validity Assessment*. Beverly Hills, Calif.: Sage Publications.

Collier, David, and Steven Levitsky. 1997. Research Note: Democracy with Adjectives: Conceptual Innovation in Comparative Research. *World Politics* 49:430-451.

Collier, David, and James E. Mahon. 1993. Conceptual Stretching Revisited - Adapting Categories in Comparative-Analysis. *American Political Science Review* 87 (4):845-855.

Collier, David, and James Mahoney. 1996. Insights and pitfalls: Selection bias in qualitative research. *World Politics* 49 (1):56-91.

Court, Julius, Goran Hyden, and Ken Mease. 2002. Assessing governance: Methodological challenges. *World Governance Survey Discussion Paper 2*.

Crepaz, Markus M. L., and Ann W. Moser. 2004. The Impact of Collective and Competitive Veto Points on Public Expenditure in the Global Age. *Comparative Political Studies* 37 (3):259-285.

Dalton, Russell J., and Martin P. Wattenberg. 2000. *Parties without Partisans: Political Change in Advanced Industrial Democracies*. Oxford: Oxford University Press.

Deininger, Klaus, and Lyn Squire. 1996. A new data set measuring income inequality. *World Bank Economic Review* 10 (3):565-591.

Doran, Charles F., Robert E. Pendley, and George E. Antunes. 1973. Test of cross-national event reliability: Global versus regional data sources. *International Studies Quarterly* 17 (2):175-203.

Duncan, Greg J., and Daniel H. Hill. 1989. Assessing the quality of household panel data: The case of the panel study of income dynamics. *Journal of Business & Economic Statistics* 7 (4):441-452.

Friedrich, Carl J. 1966. Some general theoretical reflections on the problems of political data. In *Comparing nations: The use of quantitative data in cross-national research*, edited by R. L. Merritt and S. Rokkan. New Haven; London: Yale University Press.

Garrett, Geoffrey, and Deborah Mitchell. 2001. Globalization, government spending and taxation in the OECD. *European Journal of Political Research* 39 (2):145-177.

Garrido, Luis, and Luis Toharia. 2004. What does it take to be (counted as) unemployed? The case of Spain. *Labour Economics* 11 (4):507-523.

George, Alexander L., and Andrew Bennett. 2005. *Case studies and theory development in the social sciences*. Cambridge, Mass.: MIT Press.

Gerring, John. 2007. *The Case Study Method: Principles and Practices*. Cambridge: Cambridge University Press.

Goertz, Gary. 2006. *Social science concepts: A user's guide*. Princeton: Princeton University Press.

Gujarati, Damodar N. 2004. *Basic econometrics*. 4th ed. Toronto: McGraw-Hill.

Hambleton, Ronald K., and Liane Patsula. 1998. Adapting tests for use in multiple languages and cultures. *Social Indicators Research* 45 (1-3):153-171.

Herrera, Yoshiko M., and Devesh Kapur. 2007. Improving data quality: Actors, incentives, and capabilities. *Political Analysis* 15 (4):365-386.

Hooghe, Lisbet, and Gary Marks. 2003. Unraveling the central state, but how? Types of multi-level governance. *American Political Science Review* 97 (2):233-243.

Hug, Simon. 2003. Selection Bias in Comparative Research: The Case of Incomplete Data Sets. *Political Analysis* 11 (3):255-274.

Hyden, Goran, and Julius Court. 2002. Governance and development. *World Governance Survey Discussion Paper 1*.

Jackman, Robert W., and William A. Boyd. 1979. Multiple sources in the collection of data on political conflict. *American Journal of Political Science* 23 (2):434-458.

Kaufmann, Daniel , Aart Kraay, and Massimo Mastruzzi. 2007a. Governance Matters VI: Governance indicators for 1996-2006. *World Bank Policy Research Paper 4280*.

Kaufmann, Daniel, Aart Kraay, and Massimo Mastruzzi. 2004. Governance matters III: Governance indicators for 1996, 1998, 2000, and 2002. *World Bank Economic Review* 18 (2):253-287.

Kaufmann, Daniel, Aart Kraay, and Massimo Mastruzzi. 2007b. The worldwide governance indicators project: Answering the critics. *Working Paper*.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review* 95 (1):49-69.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton: Princeton University Press.

King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon. 2003. Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review* 97 (4):567-583.

Kittel, Bernhard. 1999. Sense and Sensitivity in Pooled Analysis of Political Data. *European Journal of Political Research* 35 (4):533-558.

Kittel, Bernhard. 2006. A Crazy Methodology? On the Limits of Macro-Quantitative Social Science Research. *International Sociology* 21 (5):647-677.

Kittel, Bernhard, and Herbert Obinger. 2003. Political parties, institutions, and the dynamics of social expenditure in times of austerity. *Journal of European Public Policy* 10 (1):20-45.

König, Thomas, Daniel Finke, and Stephanie Daimer. 2005. Ignoring the Non-ignorables? Missingness and Missing Positions. *European Union Politics* 6 (3):269-290.

Kono, Daniel Y. 2006. Optimal obfuscation: Democracy and trade policy transparency. *American Political Science Review* 100 (3):369-384.

LaPalombara, Joseph. 1968. Macrotheories and microapplications in comparative politics: A widening chasm. *Comparative Politics* 1 (1):52-78.

Ledet, Richard. 2006. The Quality of Government Institute's Cross-Sectional and Cross-Sectional Time-Series Dataset. *APSA Comparative Politics Newsletter* 17 (2):29-31.

Lieberman, Evan S. 2002. Taxation data as indicators of state-society relations: Possibilities and pitfalls in cross-national research. *Studies in Comparative International Development* 36 (4):89-115.

Lieberson, Stanley. 1991. Small Ns and big conclusions: An examination of the reasoning in comparative studies based on a small number of cases. *Social Forces* 70 (2):307-320.

Lijphart, Arend. 1971. Comparative Politics and the Comparative Method. *American Political Science Review* 65 (3):682-693.

Lijphart, Arend. 1999. *Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries*. New Haven: Yale University Press.

Lustick, Ian S. 1996. History, historiography, and political science: Multiple historical records and the problem of selection bias. *American Political Science Review* 90 (3):605-618.

Mahler, Vincent A. 2004. Economic Globalization, Domestic Politics, and Income Inequality in the Developed Countries: A Cross-National Study. *Comparative Political Studies* 37 (9):1025-1053.

Merritt, Richard L., and Stein Rokkan, eds. 1966. *Comparing nations: The use of quantitative data in cross-national research*. New Haven; London: Yale University Press.

Mill, John Stuart. 1843 [1974]. *A system of logic*. Toronto: University of Toronto Press.

Munck, Gerardo L., and Jay Verkuilen. 2002. Conceptualizing and measuring democracy - Evaluating alternative indices. *Comparative Political Studies* 35 (1):5-34.

Munck, Gerardo L., and Richard Snyder. 2007. Debating the direction of comparative politics: An analysis of leading journals. *Comparative Political Studies* 40 (1):5-31.

O'Kane, Rosemary H.T. 1993. The ladder of abstraction: The purpose of comparison and the practice of comparing African coup-d'état. *Journal of Theoretical Politics* 5 (2):169-193.

Przeworski, Adam, and Henry Teune. 1966. Equivalence in cross-national research. *Public Opinion Quarterly* 30 (4):551-568.

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

Ragin, Charles C. 1987. *The Comparative Method: Moving Beyond Quantitative and Qualitative Strategies*. Berkeley: University of Berkeley Press.

Ragin, Charles C. 2000. *Fuzzy-set social science*. Chicago: University of Chicago Press.

Ragin, Charles C. 2008. *Rethinking social inquiry*. Chicago: Chicago University Press.

Rohlfing, Ingo. 2008. What you see and what you get: Pitfalls and problems of nested analysis in comparative research. *Comparative Political Studies*.

Rokkan, Stein. 1966. Comparative cross-national research: The context of current efforts. In *Comparing nations: The use of quantitative data in cross-national research*, edited by R. L. Merritt and S. Rokkan. New Haven; London: Yale University Press.

Russett, Bruce M. 1966. The Yale Political Data Program: Experience and prospects. In *Comparing nations: The use of quantitative data in cross-national research*, edited by R. L. Merritt and S. Rokkan. New Haven; London: Yale University Press.

Sartori, Giovanni. 1970. Concept misformation in comparative politics. *American Political Science Review* 64 (4):1033-1053.

Scharpf, Fritz W. 1997. *Games real actors play*. Boulder: Westview Press.

Scheuch, Erwin K. 1966. Cross-national comparisons using aggregate data: Some substantive and methodological problems. In *Comparing nations: The use of quantitative data in cross-national research*, edited by R. L. Merritt and S. Rokkan. New Haven; London: Yale University Press.

Smith, Raymond F. 1969. On the structure of foreign news: A comparison of the New York Times and the Indian White Papers. *Journal of Peace Research* (1):23-36.

Sommer, H., and J. R. Scarritt. 1999. The utility of Reuters for events analysis in area studies: The case of Zambia-Zimbabwe interactions, 1982-1993. *International Interactions* 25 (1):29-59.

Sorrentino, Constance. 2000. International unemployment rates: How comparable are they? *Monthly Labor Review* 123 (6):3-20.

Teune, Henry. 1968. Measurement in comparative research. *Comparative Political Studies* 1 (1):123-138.

Thies, Cameron G. 2002. A Pragmatic Guide to Qualitative Historical Analysis in the Study of International Relations. *International Studies Perspectives* 3 (4):351-372.

van Deth, Jan W. 1998. Equivalence in comparative political research. In *Comparative Politics: The Problem of Equivalence*, edited by J. W. van Deth. New York: Routledge.