5
Making Measures Capture Concepts: Tools for Securing Correspondence between Theoretical Ideas and Observations

Bernhard Miller

Introduction

'Perhaps the most fundamental barriers to good comparative research are measurement and the problems of comparability of measures.' (Peters, 1998, p. 80) A quick glance at the contents of this book reveals that this is a bold statement. Given the sheer number of challenges we face designing our research projects, it might even be an overstatement. But whether or not we share Peters' view, measurement as the link between theory and empirical reality is the backbone of empirical research and therefore at the core of research design, irrespective of whether research is quantitative or qualitative (based on large-n or small-n), or, for that matter, whether it is comparative or not. The central role of measurement in research design goes some way to explain the skepticism of one distinguished commentator on the subject who is 'doubtful, that any amount of study ... can teach you how to measure social phenomena, though it can conceivably be helpful in understanding exactly what is achieved by a proposed method of measurement or measuring instrument' (Duncan, 1984, p. 154). This is what this chapter sets out to do.

I proceed by addressing two sets of issues: First: What are the challenges we face devising measures? And: Which tools can we employ to help us solve them? Particularly for the readership of this volume it will be helpful to look at research design problems from the vantage point of what is to be achieved. Accordingly, measurement should be understood in functional terms as the process of arriving at persuasive empirical tests of research hypotheses (Geddes, 2003, p. 157). More narrowly, and as a part of this process, measurement is 'the assignment of numbers to objects or events according to rules' (Stevens, 1951, p. 22).[1] While this definition is coined for quantitative researchers, it can be generalized to qualitative research. Measurement attributes (relative) *values* to observations according to pre-defined rules. In the following, I stress the design of measures – and only discuss in passing issues of measurement error (Brady and Collier, 2004; King, Keohane and Verba, 1994).

This chapter is written for the researcher engaged in political science research outside of survey research[2] and combines insights from both a more abstract as well as from an application oriented perspective on measurement. It is structured as follows: In the first section I discuss measurement as the process of operationalization, validity- and reliability-testing. In the second part I discuss how careful index-construction can help alleviate problems of measurement and provide a compact list of practical advices for researchers. The application part ties these recommendations to my own research and illustrates their usefulness. The conclusion summarizes the main points.

Design problem

Measurement problems are manifold and affect both large-n and small-n studies. For large-n studies, measurement is often said to be reductionist, based on inadequate indicators and thus resulting in poor data quality (Geddes, 2003, p. 216; Collier, Brady and Seawright, 2004b, p. 206). Measurement in small-n studies on the other hand has drawn criticism because of potential subjective biases (King, Keohane and Verba, 1994; Geddes, 2003). There is little in these contributions, however, which translates directly into practical tips for the research process (Thomas, 2005).

As Wonka (Chapter 3) stresses, the necessary preconditions for measurement are clear and unambiguous concepts. Measurement needs to proceed from there. To discuss just how, we need to delineate what we mean by measurement. Everyone has an intuitive idea of what measurement is. This intuition probably involves commonplace measures such as temperature. The process of establishing temperature is quite simple: You pick a thermometer and know the temperature. Changes in the object of interest are easily observed, readily quantified, reliably reproduced, and can easily be documented. Most readers would sigh in relief if measurement in our discipline was as straightforward.[3] In order to discuss differences, difficulties and remedies, I shall use the temperature example to take a closer look at the elements of which measurement consists (Figure 5.1). The concept we are interested in is temperature – different degrees of warm and cold in our environment. In and by itself the concept is unobservable and of little help. Research into the characteristics of mercury has enabled inventors to link the volume of this metal to changes in temperature. This step is called operationalization and results in an indicator (change in volume). Operationalization is often conflated with measurement (Brady, 2004). I suggest to maintain the distinction, however, as there is a specific set of problems associated with operationalization justifying the term. Once we measure temperature it is important to establish whether the value is in fact related to how warm or cold it is. That is, the validity needs to be established (it could be possible for example that our measure only works for certain ranges of temperature). Finally, in order to be able to produce a reliable measure we need to make sure that its results are reproducible, that is, any researcher must be able to arrive at the same result using the same measurement under the same circumstances.

The scheme in Figure 5.1 shows that measurement can usefully be described as a process (Carmines and Zeller, 1994, p. 2). With concepts specified we know what we theoretically want to observe. Latent variables are our theoretical constructs while observed variables are empirical manifestations thereof.[4] The latent variable is what a researcher

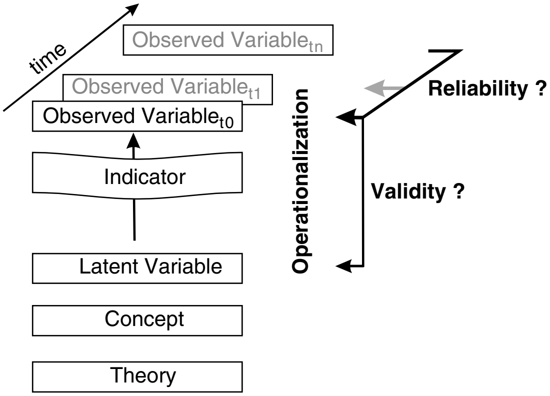

Figure 5.1 The measurement process

would ideally like to observe (politicians' true policy preferences for example). Usually, direct observations are not possible, however, and indicators need to be identified through which the researcher arrives at her observable variable. This part of the measurement process is called operationalization. After the data are collected, the researcher needs to make sure that they are valid (how closely does the observed correspond to the latent variable?) and reliable (are the values for the observed variable identical if measurement is repeated?). What is worth emphasizing is that the measurement process is identical whether the research is large-n or small-n, quantitative or qualitative (King, Keohane and Verba, 1994, p. 152). As Brady stresses, '*qualitative* comparisons are the basic building blocks of any approach to measurement' (2004: 63, italics original; see also the discussion on typologies in Lehnert, Chapter 4). The notion that measurement is more of a quantitative playground thus lacks a basis. What differs is the operationalization (Brady, 2004, p. 65, fn 14) and the means to assess validity and reliability. As questions of measurement apply to all dependent and independent variables it is also insubstantial whether the design is outcome- or factor-centric. To discuss the research design problem let us begin with a list of criteria for good measures and then proceed by discussing the difficulties involved in achieving them.

How to design a good measure

In the ensuing part, I outline criteria for the design process. The design aspect is particularly central in political science as there are many extensive theoretical concepts which are not directly observable (the size of a constituency is observable, corporatism and democracy are not). Virtually all measures in the social sciences are derived measures, i.e. they are based on another indicator (Hempel, 1952). Corporatism for example is – among other things – measured as the degree of union concentration (Siaroff, 1999). Given the mostly complex relation of concepts and empirical reality, the central criterion for a good measure is that it be based on theoretically sound foundations. If there is no theoretical blueprint we have nothing to evaluate our measure against. Furthermore: the more precise the theoretical framework, the easier to develop means for testing it. The literature offers standard demands – data need to be comparable, results must be valid and reliable – but more concrete criteria are hard to come by.

Operationalization

A central challenge to measurement is that variables might not be observable (King, Keohane and Verba, 1994). King and colleagues recommend restricting research to observable concepts. This, however, is not helpful – amongst other things because our research should be determined primarily by theoretical concerns. The task then is to find sensible observable variables for our concepts.

'Observable' as a term evidently entails some ambiguities. I therefore resort to the distinction between latent and observed variables (Bollen, 1989). For example, an actor's preferences are a latent variable – what is observable on the other hand, are revealed preferences only. Parties do not write their preferences into a manifesto without trying to anticipate voter reactions. Thus, what is revealed might not correspond to what we want to measure. The deviations which might exist need to be explored and taken into account. I present a simple measurement model to illustrate the assumptions operationalization entails. A measurement model is a formal representation of the relation between latent and observed variables to elucidate I provide an example below:

$$x_j = \lambda_{ji}\, \xi + \delta_i$$

x_j is the observed variable which is composed of the latent variable ξ, scaled by λ_{ji} to assure for comparability over all j. δ_i is the error term. As a modification, one could introduce another systematic error term to account – in the above example – for deviations of revealed and true preferences which might vary between different objects of observation.

Take, for example, 'terrorism prevention' (adapted from Rohwer and Pötter, 2002). We might be interested in how much effort states put into protecting their citizens from terrorist activity and consider as terrorism prevention all assets spent on military and police projects above the average level of spending before terrorism was on the agenda. Our observed variable then is the amount of money spent. However, it makes little sense to treat every additional € spent on the prevention of terrorism as equally important in every country. In small countries, citizens profit more from increased spending than in larger states (because longer borders and more people are harder to protect). Therefore our measure is scaled by the population or area for λ_{ji}. The error term δ_i serves to remind the researcher that each measurement might to some degree be erroneous.[5] In small-n research, it is relatively easy to explicate the substantial effect of potential measurement errors by describing the myriad of influences to which an indicator is subjected; both qualitative and quantitative researchers should make more use of this. In large-n research, there are methods to correct for measurement error (Bollen, 1989). The advantage of measurement models is that they explicate assumed causal relations and encourage thinking about alternative explanations or different causal relations between latent and observed variables.

Finding indicators is often the driving force behind measurement. Tensions between the theoretically desirable and the empirically available are therefore often barely disguised. Unless correspondence between concept and measure can be taken for granted, however, a measure is not worth much. Therefore criteria for operationalization should be documented. In cases where different indicators can be used – and there is no compelling theoretical reason for or against one – the differences between the operationalizations need to be documented. These alternative operationalizations should be presented and interpreted. This strengthens the robustness of the results and enhances the confidence we can have in them rather than weakening them.

Operationalization is often discussed in close connection with scale types. Which is the appropriate scale to use for a variable? The literature distinguishes between nominal, ordinal, interval and ratio scales (Stevens, 1946).[6] The higher the scale, the more information it contains. Yet higher scales are not inherently better – there is no *inherent* reason for assigning ordinal values to different interest groups or religions for that matter. Obviously, there will be research questions suggesting other scales. However, the issue must not be stressed too much, because ultimately it is not about the substantive meaning of a measure but only a technical characteristic. Geddes (2003, pp. 70–1), for example, uses a dichotomous measure of regime type as she is interested in the beginning and end of regimes only. Scales are important when it comes to the difference between small- and large-n research. While King and colleagues rightly assert that the scales apply to both approaches (1994, p. 151), qualitative analysis of interval scales would seem difficult if not impossible. Language does not lend itself to precise differentiation.[7] Measures in small-n studies can often be justified more thoroughly and therefore offer the potential for more accurate measurement. It is misleading to say that measures need to be more tightly specified in large-n research (Peters, 1998, p. 81). Thus, the differences between these two types of operationalization are worth exploring a bit further. Geddes (2003, p. 144) seems to suggest that quantitative operationalization is mainly restricted to picking ‘off-the-shelf’ measures, whereas qualitative operationalization forces the researcher to specify clear criteria. This differentiation clearly is not helpful. In both research approaches, any ambiguity in the design of a measure must frustrate our efforts to provide an intersubjectively relevant argument which lends itself to testing and developing hypotheses.

What must be stressed, however, is that any operationalization will be more plausible when it corresponds closely to a given concept. A nominal concept (war – yes or no) should be operationalized nominally – even though one could certainly find interval indicators (e.g., number of civilian casualties). Intermediate values would have no theoretical basis and would thus render the indicator devoid of sensible interpretation – or even lead to wrong conclusions.

To sum up, operationalization, in order to assure close correspondence between a concept and the measure needs to specify how latent and observed variables are related and whether or not there might be alternative ways to operationalize a concept. The researcher should also explicate all potential deviations between latent and observed variables.

Validity

The next three sections discuss how to scrutinize data once it has been generated. Once suitable indicators have been identified the question is: How can we make sure we have a measure which actually measures the concept we are interested in. King and colleagues recommend that researchers adhere to data and not 'allow to let unobserved or unmeasurable concepts get in the way' of achieving validity (1994, p. 25). It is not clear what implications one can draw from this recommendation. However, it seems to suggest that there are concepts which are easily observable, and thus are preferable to use in measurement. Yet, as we have seen, there are not many directly observable concepts. The best way to avoid this problem, then, is to use theoretically well-grounded and explicit operationalization as sketched in the last section. The validity tests which the literature suggests to test a measure are usually (Bollen, 1989; Duncan, 1984; Rohwer and Pötter, 2002)[8]: (1) Content validity; (2) criterion validity; and (3) construct validity. As I will demonstrate, there are significant problems with two of these validity tests – particularly for small-n studies but also medium large-n studies below the level of surveys.[9]

1. To test for *content validity* the analyst judges whether – or to what degree – a measure reflects its underlying concept. In technical terms, the researcher looks at a sample of measured values and then draws inferences as to how closely they correspond to the concept. The ensuing critique (Carmines and Zeller, 1994, p. 14), then, is that without random sampling the representativeness of the sample cannot be assumed. Furthermore, there is no criterion for when content validity is achieved, particularly given more abstract or extensive constructs (Carmines and Zeller, 1994, p. 14). A qualitative approach allows for a more straightforward interpretation which is also tied closely to the subject matter. Logically, there is no way to ever test for or find exact correspondence between observed and latent variables – simply because the latent variable can never be observed. Suggesting anything different is misleading. The central idea underlying content validity is to test to which degree an indicator reflects its domain. This invites qualitative testing by experts on a given subject. This intersubjective 'test' will thus probably fail to yield clear results, but it is ideally suited to test the correspondence of indicators and concepts – in a given context. Only experts can weigh in other factors which might potentially distort the results. However, there are trade-offs: Published expert opinion might not speak directly to the research question and thus needs to be interpreted with great care. This is the case if interviewed experts are potentially subject to time constraints or unable to answer in terms of the analytical categories provided to them. These are substantial problems and need to be approached carefully. In contrast, 'objective' validity tests, as discussed below, define such criteria but cannot provide insight into the substantial value of an indicator. The more complex a measure, the harder to match its correspondence to qualitative data – while measures such as disproportionality indices can be validated relatively easily, it is harder to assess measures of democracy, which take into account a number of different dimensions.[10] This approach is thus compatible with both large-n and small-n research.

2. Criterion validity is a large-n test. Its logic, however, can also be applied to small-n research as it assesses the degree to which a measure is related to another relevant measure. In other words, the test is based on correlation with another *existing* indicator (Taagepera and Grofman, 2003). A correlation in and by itself is useless to determine if the indicator actually measures the intended concept. Unless we know that the reference indicator is valid, the validity problem comes full-circle. In order to apply criterion validity, the researcher should either use the existing literature or qualitative information to assess correspondence to the concept. Alternatively, one can, of course, correlate a measure to another one if either of the two is highly contextual. Epstein and O'Halloran (1999), for example, measure the amount of discretion the US Congress delegates to the executive branch by manually coding the pertinent legislation. Another measure for the same concept is based on the length of statutes (Huber and Shipan, 2002).[11] Finding a high (negative) correlation between the two indicators would therefore validate the Huber and Shipan measure as we can be sure that there is a content validity to the Epstein and O'Halloran indicator. Criterion validity is therefore a useful test but one which needs to be applied carefully and with explicit reference to the concept under investigation.

3. The final test is the so-called *construct validity*. These are tests based on *hypotheses*. To elucidate: If we hypothesize a positive relation between the frequency of back-ache and visits to the doctor, finding such a relation would lead us to conclude that the indicator produces valid results. The logic of construct validity thus is identical to theory testing. It compares a finding to theoretical expectations (however derived) and a match is considered to corroborate the measure. The measure, however, is not in any way more or less valid than before! Since this test lacks an *independent* confirmation of validity, this method has a serious logical flaw – it tests for plausibility of an indicator, not for its validity and thus does not serve the purpose of a test (for similar critiques, see Bollen, 1989; Rohwer and Pötter, 2002).[12] The deficits of this test lead me to recommend avoiding it.

While not explicitly discussed as a validity test in the literature, an analyst can also use outlier analysis as a qualitative technique to gauge validity. Usually outliers are seen as a good way to test hypotheses or explore alternative explanations (Collier, Brady and Seawright, 2004b). A similar logic holds for validity tests. In any distribution, extreme values should reveal clearly identifiable differences particularly in case-by-case comparisons. While it is hard to validate exact differences between cases, the outlying nature facilitates finding evidence for or against the validity of these cases. This technique lends itself to quantitative research in particular. Extreme values on a quantitative measure should correspond to rather distinguishable characteristics in qualitative sources. A single case in a typology cell in small-n research would be another example for application of this test. Brady (2004, p. 63) in any case argues that, ultimately, all measures are based on qualitative comparisons.

To conclude: The best, if most demanding, test of validity is based on qualitative information which allows the researcher to assess in detail the correspondence of a measure to a concept. This content validity, however, needs to be at the basis of other tests based on correlations as well. Tests which do not allow for empirically tying a measure back to a concept are logically unsuited to assess validity.

Comparability

To note that observations need always be valid *given* a specific context is crucial. Any comparative study – both over time and across contexts – needs to address the problem of comparability as we already saw in the Peters quote at the beginning of this chapter. In different settings the relation between observed and latent variables might vary or, put differently – a perfectly valid measure in one context might measure another thing entirely somewhere else. Federalism, for example, might mean completely different things to people in France – without any pertinent experience – and Germany where it is a subject of permanent discussion. The literature has come to speak of the 'problem of equivalence' (van Deth, 1998). There are two broad strategies to assure equivalence (Rathke, Chapter 6; van Deth, 1998): One is to assure that indicators are exactly identical, that is to make sure no problem of equivalence exists. As pointed out, this strategy is hardly a promising remedy. The second strategy is to either choose concepts at a (higher) level of abstraction at which equivalence exists between contexts or to rely on inference. Increasing the level of abstraction (Wonka, Chapter 3) will in most instances not be adequate to the research question (why, otherwise, choose a more specific concept in the first place?). Relying on inference in this context means to use a different set of indicators which, based on qualitative background knowledge, can be assumed to actually measure the same concept. This strategy, needless to say, is demanding (Spector, 1981, p. 26). The shifting of focus from direct observation to inference is consequential, and it requires that attempts to assure validity are increased. Beyond this very general discussion, there is little which can be recommended irrespective of a concrete research design to deal with the problem of comparability (Przeworski and Teune, 1970, pp. 11–12). When drawing up a research design, we need to nevertheless keep comparability in mind.

Reliability

Validity and comparability, however, do not suffice to test a measure; Reliability is crucial. It indicates that the *same method* is supposed to arrive at the *same results* for the *same phenomenon*. Therefore reliability cannot be tested if the conditions for data gathering have changed (King, Keohane and Verba, 1994, p. 26). As every social researcher knows, there are plenty of research areas for which conditions are not the same (Rohwer and Pötter, 2002). It is here that research on institutions is privileged as changes are both rare and well documented. Therefore, data based on secondary sources is more amenable to reliability testing, as coding procedures can relatively easily be reproduced. The trade-off involved is, of course, that it is more difficult to assess the validity of the sources the analysis is based on. It is crucial, therefore, that researchers document the instrument used in obtaining data – and ideally base their information on more than one source. Returning to the point of expert information I mentioned in the last section: Comparing expert estimates (in analogy to standard deviation) will yield an estimate of how reliable a measure is.

For small-n research there is the concern that 'thick description' might sacrifice reliability for validity (King, Keohane and Verba, 1994, p. 152). King and colleagues' recommendation is not to rely on subjective data which could be influenced by the researcher's own hypotheses. Yet external sources are often not a viable solution. Other means to assure reliability therefore deserve attention. Most importantly, all coding needs to proceed along precise and unambiguous criteria – documented in a codebook (Geddes, 2003, p. 147). In more qualitative work – but also when using quantitative indicators based on qualitative data – the sources a decision is based on are necessary to make the data construction transparent. Döring (1995) or Franchino (2007) illustrate, however, that it is often feasible to base one's coding on qualitative expert data which are not in danger of being influenced on behalf of the researcher. Finally, when using secondary sources, the researcher should alert the reader to the fact that the sources diverge, and justify why one was chosen over another. This, however, points to a reliability problem of the secondary sources and cannot be controlled on the part of the researcher herself.

Summary

Theory, all authors agree, is central to measurement. This at the same time is probably one of the most significant problems with most of the more technical literature on measurement, but also with much of the literature on research design in general. Unless theory is put first and measurement second, it is hard to see how correspondence between concepts and measures is to be achieved. Operationalization needs to be tied closely to theory and should follow clear guidelines. In due course, the measure needs to be validated with recourse to qualitative evidence – even if there are additional quantitative tests of validity. Reliability, as I argued, requires first and foremost a well documented data gathering process.

Practical guidelines

Whole books have been dedicated to measurement and the discussion here necessarily needs to be limited in scope. Moreover, it is hard to offer generalized, practical advice on issues like operationalization. I therefore focus on a topic which can be helpful for many applications. Measurement entails the problem that with increasing complexity a measure is more difficult to validate. Indices are an attractive, yet often neglected, way to put this recommendation to use. They provide, as I shall argue in the third part, a means to tackle several problems or uncertainties researchers will often be confronted with. A second part briefly outlines trade-offs in practical research and recommends solutions. A final part summarizes all suggestions discussed in this chapter.

Indices

As I have argued in the last section, validity tests should be based on qualitative evidence or expert judgments. While experts might be able to validate more complex measures, we have seen that there might be a trade-off with reliability. The data generated might be unduly influenced by the expert. Asking multiple experts can alleviate the problem – but this will often not be feasible. More complex measures also render data collection more demanding. It is therefore attractive to resort to simple and parsimonious indicators (Geddes, 2003, p. 157). On the other hand, the simpler the indicator, the easier to validate. This leaves us with a seemingly contradictory recommendation – find simple indicators for complex concepts.

Application

The solution I suggest may seem somewhat old-fashioned, but indices have some substantial arguments in their favor. Indices are composite measures which combine two or more indicators on the basis of predefined rules. There are three particularly important areas in which indices are useful. First, many theoretical concepts require a look at more than one variable – indices reduce multiple indicators to one.[13] Second, researchers might have a number of different operationalizations for the index at their disposal, but little theoretical reason to favor one over the other. Indices can combine such different options, and different combinations can be compared. The reader might object that this is the kind of a-theoretical testing against which contributors to this volume strongly argue. Indeed *only* if a concise concept specification is still insufficient to yield an unambiguous operationalization – that is, in cases where we cannot make conclusive assumptions about the relation between latent and observed variables – should a more empirical solution be employed. The third reason is related: In some cases it might be possible to directly measure a concept, but the causal connection between latent and observed variable might be hard to trace. In order to avoid arbitrarily opting for one measure, an index comprised of different indicators of the same concept might be used as a basis to test validity.

Construction of indices

In this section I discuss the difficulties of constructing indices both in theory and in practical applications. As the term 'constructed variable' indicates, these variables are based on existing measures which are combined to be theoretically useful. It is therefore imperative that measurement problems be solved before an index is computed (Duncan, 1984, p. 231) and that the measures upon which an index is based are reliable and valid. The researcher's task is then in justifying the aggregation and testing the validity of the construct itself. If the researcher suspects specific index elements might involve problems, she should assess the robustness of the index by testing the impact of removing (or exchanging) the critical elements (for an example see Kaiser's [2004] alternation indices discussed below).

To construct an index, the researcher has to specify aggregation rules and justify why individual components are to be combined in a specific fashion. Many indices are additive and attach equal weight to their elements. To add up elements furthermore requires the assumption that all components affect the index in the same direction (that latent and observed variables correlate with the same sign). This unidirectional relation is important as indicators otherwise cancel each other out. In variation, weighted indices attribute more impact to some elements. Weights can be endogenous – party positions can, for example, be weighted by the seat percentage – or exogenous. An index of influence might, for instance, consider money to be twice as important as other lobbying efforts. Additive and weighted indices are particularly useful for combining several conceptually related (but distinct) elements into one measure as I will demonstrate in the next section. An index may also be based on any other mathematical transformation. Whenever an index is constructed, the analyst needs to make sure that its elements are indeed related to the latent variable and might not reflect some other construct. The logic corresponds to controlling for alternative explanations. It has to be certain that it is not some other concept, 'hidden' in the index that enters the analysis.

There are some theoretical and some more operational criteria to be observed when constructing an index:

- *Indices based on ideal types.* Indices can be built around ideal types (Lehnert, Chapter 4). Shugart and Carey (1992), for example, construct their presidentialism-index on the basis of an ideal-typical concept of presidentialism. The ideal type is defined by theoretically derived states of all variables in the index. Deviations from the defined extreme then constitute changes in index values. Indeed, a typology can form the basis of an index – the theoretical challenge is to align the types identified on one dimension.

- *Theoretically justified index values.* Taagepera and Grofman (2003) review 19 indices of disproportionality. They demand that for each index the minimal and the maximal values of the index should be defined and that there needs to be a theoretical rule to decide which units of observation should enter the index.
- *Weighting.* Weighting is crucial, as minimal changes to weights might alter the nature of the whole index. At this point indices probably are most prone to manipulation. Therefore weighting criteria need to be justified, and, if appropriate, different weights need to be discussed with explicit reference to the theory.
- *Discussion of empirical effects.* The researcher should contemplate counterfactually how extreme values (e.g., in the case of Taagepera and Grofman [2003] a large number of very small parties) would affect the index results and whether such effects are theoretically desirable.
- *Multiple possible operationalizations.* A more practical recommendation is that indices can be used to incorporate more than one way to operationalize a concept. Kaiser and colleagues (2002) argue that alternation is an element of democratic quality. In developing a measure to test the argument, Kaiser (2004) suggests three indices each using slightly different interpretations of alternation. This allows him to document the implications and trace them back to shades in the theoretical argument.
- *Test of index robustness.* Kaiser (2004) also demonstrates how the robustness of an index can be tested by removing data generated in a potentially problematic way.[14] Robustness can be tested for two scenarios. One, a whole variable for which measurement is problematic might be removed. Alternatively single cases (outliers) can be removed. If the change barely affects the index values, the measure can be considered robust with regard to the potential measurement problem.

Trade-offs

At the end of this section on practical guidelines let me briefly consider some trade-offs which so often occur in everyday research practice.

Some trade-offs might seem less important than others, but all deserve consideration:

- First and most importantly, there is *no* inherent trade-off between theoretically desirable measures and empirically available indicators. Despite limited time and money, effort should be invested in identifying the most suitable indicator. Picking off-the-shelf indicators is a legitimate alternative but might lead to a trade-off of quality for availability. Therefore, extra efforts need to be made for testing the validity of such indicators.
- In many cases available empirical information will capture theoretical concepts reasonably well. If not, there is a likely trade-off between the resources required for gathering new data and the quality of the measures. There is little to be said about this trade-off except for noting that researchers should try to find creative ways of digging up useful data.
- A trade-off exists between the complexity of measures and their reliability. The more discretionary decisions involved, the harder to test the reliability. Indices, as I have argued, might be a way of solving the problem as they consist of several elements which can each be of reduced complexity.

A summary of practical recommendations

The following list translates points from the discussion above into hands-on advice:

1. The scale of a measure should not deviate from the scale suggested by the concept.
2. In your operationalization, always explicate relations of the indicator to the concept, and clearly delineate different categories or subunits.
3. Be careful not to assume validity all too easily. Make sure that there is qualitative evidence which supports correspondence to your concept.
4. If there are multiple ways to operationalize, discuss the implications of the different choices.
5. Break down complex constructs into simple and parsimonious elements and aggregate into a composite index (explicating the construction logic). This simplifies validity and reliability tests.
6. When building indices, pay attention to justifying the aggregation rules.
7. If multiple measures exist, attempt to select based on theory. Whenever this is not possible, compare and discuss potential differences.

8. As a minimum, use (or refer to) qualitative validity tests – not only for outliers, but systematically. Other methods can be used in a complementary fashion.
9. Use outliers to support quantitative measures with qualitative evidence. They are potential sources for uncovering errors in measurement.
10. To ensure reliability, follow a codebook and document all coding decisions with the respective sources.

Application

This section illustrates the practical application of some of the recommendations above. I discuss the steps from concept to measure on the basis of my own work on informal institutions, and illustrate index construction based on research on the ombudsman.

Coalition Committees (CoC) are informal institutions in the sense that they have no basis in either legal statutes or constitutional law. Thus their existence, procedural rules, and decisions are beyond enforcement by state institutions. CoCs are most often seen as conflict-management mechanisms (see contributions to Müller and Strøm, 2000; Andeweg and Timmermans, 2007).[15] Given the prevalence of coalition governments in Europe, the question of how coalition partners maintain their cooperation through CoCs is both theoretically and empirically interesting: Theoretically, because the literature essentially assumes that coalitions, once established, are stable unless terminated by (rare) exogenous events; empirically, because there is no information on how these informal institutions operate. I address this topic for the German case.[16] Conceptually, I focus on 'reliance on informal institutions' defined as the recourse of political actors to informal venues while at the same time a functioning set of formal institutions exists. Some problems emerge: Given the informality, it is difficult to observe reliance. First, it is a data problem. Second, it is a problem of correspondence between latent and observed variable. The former problem is ameliorated by two factors. Many coalitions fix up the frequency of CoC meetings in their coalition agreements (the 2005 coalition agreement of the CDU/CSU, SPD coalition in Germany, for instance, mandates one meeting of the CoC per month). In addition, the German media report on some CoC meetings – at least in instances where important issues are on the agenda. Tackling the second question: How closely do the observed variables – frequency as agreed in the coalition agreement and reports in the media – represent the latent variable? The first indicator (frequency in coalition agreement) is unsuitable to uncover the flexibility in the use of informal institutions the literature leads us to expect. The second (frequency according to media reports) compensates for the limitations of the first, but might be biased (see Thiem, Chapter 7) since in some instances the media will be preoccupied with other issues and decide not to cover a CoC meeting. Still, given the limitations, 'reliance' can be operationalized as the frequency of publicly visible meetings, and achieve close correspondence based on the assumption (which needs qualitative backup) that in publicly visible cases the reasons to resort to informal structures are particularly important. A precise indicator is still required, and the research design needs to make up for this deficit through qualitative evidence. Given this operationalization the measurement model assumes a true score *and* an unknown but systematic error (the media bias) as well as the standard error term. Securing reliability for these indicators is not a problem: I ran a full text search with a set of terms the substantial meaning of which is stable over time. Validity is harder to assure. For one thing, there are few sources for comparing my results. However, tests were confirmatory for the cases where it was possible. More important is a qualitative analysis on the basis of interviews a) of meetings not covered in the press (were they to some degree different?) but b) also of the meetings in the measure (were they really more important so as to support the assumption above?). Such tests obviously can only be conducted for some meetings, and the test itself is potentially biased as the participating actors might not remember all details from the CoC after some time. In this case, there is no quantitative test that could even potentially be used to assess validity.

In Miller (2006), I analyze reasons for the institutional design of national ombudspersons responsible for overseeing the administration in 25 democracies.[17] The theory leads me to expect variance with respect to both the competencies the institution enjoys during investigations, and also with respect to the degree to which it is free from the influence of other institutional actors. The two concepts I specify are (1) 'investigative competencies' and (2) 'independence'. While competencies directly imply measurement along a catalogue of statutory competencies, independence could be defined either behaviorally or institutionally. Since my theory models the establishment of the ombudsman as a principal-agent relation, however, I need to look at the formal basis of the institution and not at potential deviations from it in actual practice. As both concepts are highly abstract, it is not possible to find one indicator which would capture them to any satisfying degree.[18] The literature, however, helps to identify a set of indicators, most of which are nominal. Since all indicators have a unidirectional relation to the latent variables and there is no compelling reason why some indicators should be more important than others, I have developed an unweighted additive index for each concept. The scaling of the indices (between 0 and 1) in order to achieve theoretically sensible extreme values was not as straightforward, though. For the independence dimension there are ten indicators which, if they all apply, can indeed be taken to indicate a maximum of independence.[19] For 'investigative competencies', it does not make sense to define abstract, extreme values: First, some competencies always exist (limiting variation) and, second because there is no theoretical reason to demand that all indicators be positive in order to speak of an institution with 'complete' competencies. I therefore scaled the second dimension empirically, using the countries with the highest and lowest number of existing competencies as extreme values. The example shows how two rather abstract concepts can be measured in a useful manner with the help of indices. Both can be interpreted as a continuous variable ranging from a totally independent / resourceful ombudsman to an institution which is tied to its political principals and / or has limited means to conduct investigations.

The remainder of the discussion only looks at the independence index. To assess its validity, the index was tested based on qualitative evidence in the literature. I identified outliers and extreme values and compared these values to the literature which usually comments on independence and the competencies of ombudspersons. Paired comparisons (is *a* really x units more independent than *b*?), however, was not a feasible option because there is no naturally observable equivalent to what the index as a whole measures. Since the index elements had been qualitatively established to be important for independence, however, this is not disadvantageous but rather shows the strength of indices. The reliability was fostered through the use of a rigorous coding scheme and an objectively identifiable basis for the data (statutes and constitutional provisions).[20]

These examples demonstrate that measurement design is to be clearly and explicitly connected to theoretical constructs. While in many cases this is probably done already, spelling it out allows for more transparency and renders this aspect of research design easier.

Conclusion

This paper has provided an overview of measurement – and the design of measures in particular – in research design. I have stressed the role of theory and concepts as crucial parts of any measurement process. Linking measures to theory is central both for design and for testing the correspondence of a measure with the underlying latent variables. I have argued that empirical tests of measures such as correlation analysis cannot and should not substitute more qualitative approaches for assessing validity. Researchers applying these techniques should bear their limitations in mind. The contrast between the qualitative small-n and quantitative large-n approach, portrayed by some as profound (Thomas, 2005), has not played much of a role in the discussion. This is due to the similar requirements each of the two approaches suggest for treating measurement. The overall challenge is – to reiterate Geddes's (2003) call – to provide theoretically convincing empirical tests for our hypotheses. If this purpose is served, then measurement contributes to research in a meaningful way.

Notes

1. There are, of course, other definitions of measurement – all of which, however, are covered by a definition of measurement as a process (Brady, 2004; Duncan, 1984; Schmidt, 1994, p. 257).
2. While the steps described here are, of course, applicable to any kind of social science research, methods available in survey research (especially when data are directly collected) exceed those available to others.
3. Those of us who thought that measurement in the natural sciences was straightforward would be surprised by just how substantial difficulties are (De Bièvre, 2006).
4. The term 'latent variable' is often used synonymously with 'concept'. I maintain the distinction to indicate that they are used in different literatures but also to accentuate the difference between the theoretical task of specifying a concept and the empirical aspect of testing it. Moreover, some concepts such as policy space can only be operationalized with two (or more) latent variables.
5. While the error term here is assumed to have a mean of zero, the very fact of the existence of error is important to take into account.
6. Nominal measures consist of different and distinct categories (e.g., gender), ordinal scales allow for an ordering of different states (degree of citizen participation in dictatorships, feudal states, and democracies), interval measures allow for a comparison of distances (GDP) and, finally, ratio scales have a defined zero-value (temperature).
7. This should not be seen as a weakness. Measurement error – read imprecise classification – will often render quantitative measures inaccurate – a fate qualitative measures do not share (Brady, 2004).
8. There are further types of validity tests. The ones presented here, however, are broadly representative.
9. Surveys allow for controlled and different operationalizations of concepts which then facilitate tests unavailable to most other research designs.

10. These indices assess the correspondence between voteshare and seatshare of a party in a given system (Taagepera and Grofman, 2003).
11. They argue that length is a valid proxy for the amount of detail in legislation and is thus inversely proportional to the amount of discretion.
12. Bollen (1989) suggests convergent and discriminate validity as one further set of tests. The term refers to a multi-method design where indicators of two or more concepts are measured by two or more methods each. Correlations are employed to estimate the validity. Correlations of different measures of the same concept need to be higher than correlations between concepts. Also, correlations between the same measure of different concepts need to be higher than correlations between different measures of different concepts. If applicable, this method would indeed provide a quantifiable measure of validity of the respective measures. While the objection that correlations cannot prove correspondence to the latent variable is still valid, this test makes it much more plausible that the measures actually measure the same thing. However, in research designs which cannot rely on surveys, it will not be possible to even devise such a validity test.
13. Multidimensional concepts cannot, however, be combined in an index.
14. Kaiser (2004), for example, argues that to measure the end of a cabinet on the basis of the exchange of a prime minister might be misleading as the basis for his alternation variable – as policy might depend more on parties than on its personnel.
15. Helmke and Levitsky (2004) provide an outstanding overview on the conceptualization and measurement of informal institutions.
16. For a collection of essays on the German CoC, see Rudzio (2005). Kropp (2004) provides some of the same arguments in an English contribution.
17. The data are based on bills or articles of the counties' constitutions; that is, they are prescriptive. Therefore the indicators need not be adjusted to the specific context.
18. The institution's annual budget might serve as a proxy for independence. It could, however, also measure all sorts of organizational peculiarities and would misinform our judgment in countries where the ombudsman-institution need not cover all expenditures from its own budget.
19. Since all indicators are for the most part explicitly related to specified aspects of ombudsman independence, interference of another – unspecified – latent variable is unlikely.
20. Inter-coder reliability was not an issue as all data were coded by one person only – on the basis of coding instructions.