Measurement

Nancy Cartwright and Rosa Runhardt

1. Is Syria in civil war?

At the time of writing this chapter, it is undecided whether we may classify the violent conflict in Syria that started in 2010 as a civil war. We know US President Barack Obama warned about the 'grave dangers of all-out civil war' (Mason 2012) in the Arab state and that UN secretary general Ban Ki Moon warns of civil war erupting (Black 2012). Who will decide whether it has and on what basis?

When classifications like this are done by social scientists, it is often called *coding*: assigning the conflict to a specific class (civil war/not civil war) or assigning it a specific number (say a level 5 rather than a level 4 civil war), using clear, articulated criteria. Coding is part of the process of measurement. Designing proper measures and carrying them out is one of the key jobs we expect science to accomplish. Properly defined and properly executed scientific measurements provide us with a precise picture of the things we study and give us the kind of information from which we can build scientific laws, models and principles that can help us predict and change the world around us. This chapter is about how this is done in the social sciences. We begin by considering how civil wars are measured.

1.1 Intricacies of categorizing a conflict

Making measures may be one of the central jobs that the social sciences do, but we must from the very start be clear that how it is done can have implications well beyond the confines of the sciences, and scientific measures may for just that reason be hotly politically contested. Take the case of Syria. Both the Syrian government and the opposition activists have till time of writing this denied that the conflict is a civil war, because both are aware that if the conflict is labelled a civil war, the international community may well respond to it in specific ways they don't want. Rather, they'd prefer the conflict to be called an act of terrorism (which would criticize the activists), or on the contrary, for it to be called a massacre (which would shed a positive light on the opposition). By calling the conflict a civil war we are stuck in the middle: The naming implies that there's an appreciable amount of force used on both sides, and neither of the two parties would be happy with that.

1.2 Civil war studies

There are many intricacies when it comes to defining *civil war* and deciding if a conflict falls under the definition we settle on. Just think of the many cases of conflict that are called civil wars in common speech: both the 19th century American War of Secession and the current conflicts in Colombia fall under this label. What connects them?

In the social science civil-war literature, the most common definition focuses on four aspects: A civil war is a war that involves fighting internal to the metropole, the national government participates actively, both sides employ an appreciable amount of force (the 'effective resistance criterion', which may require, for example, that the opposition is responsible for at least 5% of deaths, or at least a hundred government deaths) and a certain number of deaths result from the conflict (the 'death threshold') (Sambanis 2004).

All this can seem ad hoc and especially the choice of threshold levels and percentages. There is nothing in nature that tells us what a civil war is; it's not a category in the same way that atoms or elm trees are. Rather, war scholars use case studies and statistical analysis to find out what kinds of categories help them understand conflicts and accomplish other goals they have in view---like monitoring what is happening, making comparisons across countries, regions, or times, preventing harm or predicting the future, and doing so in either the long or the short run and either globally or locally. On that basis, they've judged definitions requiring these four elements helpful. Importantly, it turns out that internal conflicts that have these characteristics in common have other characteristics in common as well. As a stunning example, scholars have found some evidence that states with a secondary school enrolment of ten percentage points higher than the average have a reduced risk of about three percentage

points to break out in civil war---at least as civil war is characterized by this kind of definition (Collier and Hoeffler 2004).

Even if scholars agree on a definition, that is not the end of the story. Before the definition can be applied, civil war scholars must 'operationalize' the definition---that is, specify just what procedures need to be carried out to decide if the definition is satisfied. This can involve entire new layers of definition. For instance, what exactly is to count as a war death? Deaths of combatants? 'Collateral deaths' of 'civilians' directly due to conflict? Deaths that arise indirectly due to damages to health care and delivery systems, water supplies, food resources, etc.? And then, who counts as a combatant and who counts as a civilian? Must one be armed at the point of death to count as a combatant? More on-the-ground problems arise when it comes to collecting the actual data needed to check whether a state is in civil war. Counting deaths in a conflict is tricky, especially when both sides of the conflict try to hide the number of casualties they have taken and that they have caused to the other party.

2. Three requirements for measurement

Coding for civil war---categorizing a violent conflict as a civil war on clear, articulated criteria---is, as we noted, an example of measurement in social science. You may most

immediately associate measurement with assigning a number to a specific unit (think, for instance, of measuring someone's height). But deciding to put an individual unit in a specific category is as much a measurement as assigning to them a number or value for some quantity like height or income. If we want to be formal about it, we can think of categories as quantities with two values: Individual units (which can be persons, countries, regions, institutions, etc.) take the value 'yes' or 1 for this quantity if they fall within the category and 'no' or zero otherwise.

Measurement, though, isn't just assigning values or numbers or putting things into categories; it is doing so in a systematic and grounded way. This involves three different kinds of activities: *characterization*---laying out clearly and explicitly what the quantity or category is, including any features of it that we intend to make use of in assigning a number or category to a unit; *representation*---providing a way to represent the quantity or category in our scientific work; and *procedures*---describing just what must be done to carry out the measurement successfully. It is important that the three activities mesh. They should not only be consistent but also mutually supporting, as we shall illustrate.

2.1 Characterization

Before we can measure a quantity or specify procedures for deciding if an individual unit fits into a category, we need to be clear what that quantity or category is. We shall illustrate with 'yes-no' quantities for now, pursuing our original example. Categorizing an individual requires a couple of things that have been illustrated by the case of civil war. First, a category must be formulated, a category that is *useful for purpose*---for the specific purposes to be served in measuring the category. For instance, if we are interested in finding ways to lower the probability of armed conflict between government and some opposition groups espousing opposed ideals, we may delineate the category *civil war* one way. On the other hand, if we are more interested in preventing the spread of HIV/AIDS as result of violent conflicts (which may occur by infected members of one side of the conflict purposely raping civilians), the category we have delineated under the label *civil war* for the first purpose may be not so useful. So, delineation depends on the aim of the social science research.

Why not, you may ask, just keep working on the problem till we manage to formulate the 'correct' delineation of civil war? That's the tricky thing about *civil war* as a category, and similarly with a great many of our other concepts in social science. Civil war is not something 'just there' in nature, like a hydrogen atom or a birch tree or the planet Jupiter. It is

not what is called a 'natural kind'. Rather, it is a concept that is socially constructed. It is socially constructed in two senses. First, civil war is an activity that depends on human actions to occur; and second, this concept is only formulated because we care about, and thus wish to focus on, particular kinds of conflicts, like those currently going on in Somalia, Sudan and Colombia.

Second, *civil war* is not something that has definite boundaries nor, it seems, is there some one set of characteristics that all things we label as civil wars have in common, by contrast for instance with electrons, which are all negatively charged or with the number of people whose births are recorded in the local church registry for 1847. Rather, the different violent conflicts that get called civil wars all differ from one another. This reminds us of the way members of a family look alike. They might look like Smiths and each one may look in some significant way like some of the others, but there is probably no set of visible characteristics that all or even most members of the Smith family have in common. The Smith family look alike to us not because they share one common feature, but because there's sufficient overlap among them with respect to the visible features that we take note of. Civil wars seem to be like that. But to make matters more complicated, some civil wars have more elements in common with non-civil wars than they do with other civil wars.

In order to describe this fuzziness we find in concepts like *civil war*, we call them *Ballung* concepts: concepts that are characterized by family resemblance between individuals rather than by a definite property. *Ballung* is a German word for a concentrated cluster; the term *Ballungsgebiet* (*Ballung* region) is used to describe sprawling congested urban complexes like the area from New York to Washington on the East Coast of the US. We take the term from left-wing leader of the Vienna Circle between the Great Wars, Otto Neurath. Neurath worried about the role that Ballung concepts can play in 'proper science' since it seems there can be no strict universal relations of the kind typical in physics (and perhaps other natural sciences) between concepts that have neither strict boundaries nor whose instances share any essential features in common. (Neurath 1936) When we are trying to find an apt categorization to suit the aims of the social science research we are doing, we must find some way to deal with both the Ballung fuzziness of many social concepts as well as the fact that many are socially constructed.

A third problem for characterization is that even if we construct a category or quantity that we are interested in, there is no guarantee that we will be able to do fruitful research using this category or quantity. If we have a category that consists of individuals that have little in common except the fact that they're grouped in this category, we can't do much with our categorization. If instead it turns out that they also have other properties in common---

perhaps all the states we say are in civil war are also very poor, or perhaps they all have a high unemployment rate for young men---then we can begin to formulate some useful claims using these concepts.

One of the important features that social scientists look for in delineating a quantity or category is that there be some shared set of causes or some shared effects from being in the category or possessing specific values of the quantity, though the sets of causes and effects may, like the concepts themselves, have only rough boundaries and no strict criteria of inclusion.

One of the ways in which categorization can go wrong is when we use a category that is so general that all causes become invisible. We can use civil war again as an example. Suppose we take for granted that we have a decent database of all violent conflicts in states since 1945 (such as the Correlates of War (COW) project database founded by political scientist J. David Singer and hosted at Pennsylvania State University since 2003) and that we can do good statistical research on these conflicts. Using the characterization of *civil war* in section 1.2 but considering various different values for the threshold of deaths and the effective resistance criterion, social scientists have looked for a correlation between whether a country is in civil

war and how much ethnic diversity it has. For some values of the threshold and criterion they find a small correlation; for others, none. This suggests that ethnic diversity is not a real cause of the kinds of conflicts we are trying to focus on using the concept *civil war* (Sambanis 2004).

But this is surprising because our intuitions, as well as a large body of literature in international relations and a number of case studies in different countries, paint a different picture. In these other kinds of studies, it seems as if countries that are ethnically diverse are likely to develop conflicts. Why then does it not show up in the statistical research? The reason that some social scientists adduce is that as a category *civil war* as characterized in section 1.2 is too general. By averaging across all conflicts that satisfy that characterization, we lose the information about ethnic diversity. To resolve this problem, a more narrow category of *ethnic civil war* has been devised. For instance, Nicholas Sambanis defines *ethnic civil war* as 'war among communities (ethnicities) that are in conflict over the power relationship that exists between those communities and the state.' Statistical research with this category provides evidence, or at least so it is claimed, that ethnic diversity in a country does indeed increase the probability of ethnic civil war there (Sambanis 2001).

Another familiar way that characterization can go wrong is when different scholars seem to be talking about the same thing but report very different results for it. They seem to be talking about the same thing because they call it by the same name, like 'civil war', or 'degree of poverty', or 'level of inflation'. But often in actuality they are not disagreeing. Instead, they are measuring different concepts, different because they have characterized them in different ways. Some authors might say, for instance, that Somalia is at civil war in a certain year, while others might disagree because they have adopted different thresholds for the number of deaths.

Example: CPI. Moreover, the way we characterize a concept for policy can have different consequences for different groups of people. Consider for example the U.S. Consumer Price Index (CPI), which is meant to be a measure of inflation in the price of consumer goods. To measure the rate of inflation from one year to another, the average price of a basket of goods is compared between the two years. To do that, the CPI procedures specify that a sample be made of the prices of these goods from various kinds of stores across the country. So to provide a detailed characterization of this concept that will be relatively easy to proceduralize, a great many questions need to be answered, amongst which is: At what kinds of stores do individuals do their shopping?

One of the issues in more recent years is that people have started shopping at discount stores rather than through traditional distribution channels such as grocers. The Boskin commission, a panel of experts appointed by the US Senate Finance Committee to study the accuracy of the CPI, observed this gain in market share of discount stores and decided to adjust the CPI accordingly. This made a difference to the CPI value and correlatively to the welfare of a great many people whose income---like social security and veterans' benefits---is pegged to the CPI. However, groups differ in crucial respects when it comes to the stores they go to. There are large groups of people in the US, like the elderly and poor veterans dependent on their benefits, who are unable to go to discount stores, which are often located far away from town centres and thus difficult to get to. The elderly and poor veterans are therefore disadvantaged by the new categorization of inflation compared to the people who do shop at discount stores.

Note that this example also shows that the consequences of using a specific measure rather than another can require a very specific form of expertise. Not only do those appraising the consequences of using one CPI rather than another need to be experts in economics, they also need knowledge of the shopping patterns of, for example, the elderly.

2.2 Representation

Once social scientists have come to a specific characterization of a social science concept, they need to devise a method of representing the concept in their scientific work---in their claims, their calculations, and their studies. It is important that the representations adopted do justice to the characterization. This is true for all kinds of social science concepts. Think, for instance, of *poverty*. We can characterize poverty in various ways that make it a yes-no matter: Either you're poor, or you're not. For instance, Do you earn less than $2 a day, or less than 2/3 the median income in your country? In this case poverty should be represented as a 2-valued variable. Alternatively the concept of interest may be not *poverty* but *degree of poverty*. This might be characterized in terms of which quartile of the population your income falls in, in which case it would be represented by a variable that takes four values. Alternatively one could simply characterize *degree of poverty* as how much income one earns. Then it would be represented as a continuous variable. Or we may have a more nuanced concept characterized in terms of various aspects of being poor that concern us. For instance we may define a concept where not only income matters but also amount of access to clean water, electricity, education, and housing. In this case the concept should be represented by a table of different indicators, where the indicators themselves may be yes-no variables or multivalued.

Similarly, there are several ways in which we can represent *civil war*. First, we may represent civil war with a two-valued variable. In that case we judge that either there is no civil war in a state, and we assign a 0 to that state, or there is, and we assign a 1 to it. This is done by many social scientists interested in civil war, particularly by those who wish to undertake purely quantitative, statistical research. Recently, however, this two-valued representation has come into question. For one, some commentators feel it is wrong to suggest that civil war is the same no matter where it takes place. The civil war in Sudan is different from the one in Colombia. But the two-valued representation can make it seem as if all civil wars are the same by lumping the countries in which they occur together under the label '1---engaged in civil war'. This means that the two-valued representation does not do justice to the characterizations of *civil war* that these commentators have in view (even if they have not provided a characterization explicitly).

If we look at other characterizations of civil war, like some of those we describe below, it is more reasonable to argue against the two-valued indicator. This kind of back-and-forth process of mutual adjustment is typical in devising measurers in both the natural and the social sciences. Characterization and representation must get changed in tandem. They must also change in step with the procedures for measurement (which we turn to in section 2.3) and vice versa. A change or improvement in any one of the three typically produces the need for change in the other two.

It is useful to note here Stanley Smith Stevens' distinctions to summarize some familiar types of representation that social scientists make use of. Stevens (1951) describes four ways of representing a concept. We have already seen the first type. We may represent a concept using a *numeral* scale by assigning different numbers or letters or names to the different units that fall under the concept. This is the kind of representation we employ when we use a two-valued variable, as in the case of *civil war* where countries are divided into two groups, conventionally labelled '1' and '0'. But the numbers are just labels---we don't treat them as numbers.

Stevens' second type of representation is the *ordinal* scale. Using an ordinal scale means *ranking* the units that fall under the concept. Here the numbers do mean something. The higher the number assigned an individual unit, the more of the quantity it possesses. So we might rank *degree of poverty* on a scale from 1 to 10 . With a merely ordinal scale the size of the differences between any two numbers doesn't mean anything. There is no implication with an ordinal scale that, for instance, the difference in the degree of poverty between individuals assigned '4'and those assigned '6' is the same as that between those assigned '8' and those assigned '10', nor that there is twice the difference in the degree of poverty for individuals falling in either of these groups as the differences between individuals assigned '1'

versus '2' or '7' versus '8'. This contrasts with Stevens' third type of scale, the *interval* scale, which both orders individual units and has equal intervals between points with equal numerical separation. (You see a similar distinction to that between interval and ordinal scales discussed in Chapter 10 between cardinal and ordinal scales.)

Fourth, social scientists may rank the units under study on a *ratio* scale---an interval scale with a natural zero point. We would for instance be using a ratio scale if we assigned the label *degree of civil war* according to the number of deaths, e.g. this civil war claimed 5,000 deaths, whilst that one claimed 23,000.

A different way of representing civil war is with a probability distribution. In line with the idea that a different representation generally presupposes a different concept---a different characterization---the representation in this case is not of *civil war* as characterized in section 1.2, but of 'probability of civil war onset in a state'. Nicholas Sambanis, for instance, uses this representation in order to explore and represent how strongly various different social and economic factors are conducive to civil war. He wishes to know what other aspects of a state influence the chances of civil war breaking out. He then uses the conditional probability for civil war onset, conditional on these other aspects, as an indicator that these aspects influence

the onset of civil war. We have already seen an example of this research in the case of the conditional probability of civil war onset given a specific unemployment rate and/or school enrolment rate for young men. Another conditional probability we encountered was the probability of civil war breaking out given a certain amount of ethnic diversity.

In general, this probability research goes as follows. Faced with a certain state, say Angola, the civil war scholar tries to find the value of a set of parameters that are meant to represent properties of that country. These parameters might include the school enrolment rate and amount of ethnic diversity already mentioned, but also how poor the country is (represented perhaps by the GDP---Gross Domestic Product), the percentage of the country that has a mountainous area (which is of interest because it is thought that rebels might hide there and thus be able to put up more effective resistance against the government), how many valuable goods (like diamonds or oil) are exported, how weak the government is, etc. Then, the scholar also looks at whether there is civil war in the country. Data on these parameters is collected across as many countries and as many different kinds of countries as possible. The data is then subjected to various kinds of statistical analysis to see if any significant correlations can be found. Of course, as Chapter 16 in this book stresses, this is not a proof of genuine influence since 'correlation $\neq$ causation'. But it can be helpful in suggesting causal hypotheses to subject to further tests.

Yet another way to represent social science concepts is with a table of indicators. This is generally a good representation for Ballung-type concepts where a number of features matter to the concept but no one or two can be singled out as essential and where it is not clear which combinations of features are better or worse than which others. The table simply lists what values the individual to be measured takes for each of the features that matter.

For instance, the European Union has adopted a common set of social indicators in order to represent social exclusion in a three-layer table. The first layer contains seven lead indicators that are supposed to be important aspects of social inclusion/exclusion throughout the European Union, such as the proportion of 18 to 24 year olds who have only lower secondary education and are not in education or training leading to such a qualification, or the proportion of people living in households that lack specified housing amenities or have specified housing faults. The second layer contains additional indicators taken to be important to the concept but less central, such as the proportion of the population aged 18 to 64 with only lower secondary education, and an indicator for the proportion of people living in overcrowded housing. The third layer contains factors that matter only more locally and that member states decide for themselves to include, so these can differ from country to country. Member states should amongst others include non-monetary indicators of

deprivation that are particular to their context. Italy for instance measures the percentage of elderly persons living alone without living siblings or children; Belgium measures the relation between education of parents and children; and Finland measures the number of people on the waiting list for their State Housing Board rental accommodation.

As with all representations there are advantages and disadvantages to tables of indicators. Two related advantages stand out. First, often this kind of representation is the only appropriate one for the concept we have in mind. When the concept we care about is a Ballung-type concept made of many aspects and with fuzzy edges, simpler measures end up omitting aspects and drawing boundaries that can leave individuals on the wrong side. Second, a table of indicators provides a far more detailed picture that allows us to survey the range of plusses and minuses that contribute to assigning individuals into or out of the category. We can, for instance measure the poverty or welfare of a country on a ratio scale via GDP per capita. But the fact that many low GDP countries have had higher life expectancy at various times than would be predicted from such a measure---like Sri Lanka, Costa Rica, the Indian state of Kerala, China, some of the Soviet socialist states or Jamaica--- suggests that the individuals in these countries are not as 'poor' as this measure would make them out to be. One standard account is that the societies provide things that individuals cannot buy for themselves, like an educational system, clean water, and health care facilities,

and many individuals in the country have access to them. This is one of the reasons we might think of, say, lack of access to clean water for its citizens as part of what we mean by categorizing a society as poor.

The chief disadvantage of a table of indicators is that it does not allow for comparisons, either across time or across different units falling under the concept, except for rare cases where one unit performs better on every indicator in the table than does another. Yet we do want to compare, both to see how things are changing in time and to see which social systems work better for reducing or enhancing the characteristics of societies that we care about. So it is not unusual to turn a table of indicators into an index number by weighting the different aspects in some way to come up with a single number. The *HDI---Human Development Index*---is a good example. The HDI, a number between 0 and 1, is the geometric mean of normalized measures of life expectancy, education level, and national income. Norway and Australia, for instance, have HDIs above .920, and Niger and the Democratic Republic of Congo, below .300.

As usual there is a trade-off. For Ballung-type concepts there is generally no appropriate way to weigh the various items from a table that makes good sense across all applications. The

weightings are, in that sense, arbitrary. This matters because different ways of weighing will give rise to different rankings, both for the same units across time---are they improving or getting worse---and in comparing units to each other.

This provides yet another illustration of the fact that different methods of representation generally mean that it is really different concepts---different characterizations---being represented. For instance, one of the aspects of civil wars that some scholars believe are not currently well represented in civil war studies is the variation in the conflict in a state over time. By giving only a 0 for a year in which a state is not in civil war and a 1 for a year in which it is, we ignore a large amount of information about the development of the conflict over time. A 1 in one of the first years of conflict may hide a very different kind of situation than a 1 in one of the last years of conflict. Similarly, by focusing on a 0 or 1 for an entire state rather than for a location in the country, the social scientist ignores information on local armed conflicts.

Which of these representation/characterization pairs is best? As we have stressed, for the most part there is no correct answer. It depends on purposes. If we want a very accurate account of just what the poverty (or lack of it) of the inhabitants is like in a country, a

Ballung-type concept with its associated table of indicators would generally be best. If we must have rankings for some reason, we will need an index. And again, which is the best index, weighting which factors in which way, depends on purpose.

This may suggest that to serve all our different purposes we should construct more and more different measures surrounding the same basic idea like *civil war* or *poverty* or *human development*. This will naturally make for more accuracy. On the other hand this has its disadvantages.

First, too many measures are confusing, not only for the general public but for social scientists themselves. Having a great many measures available also makes it easier for those who want to bend the results to fit their own ends to cherry pick the measures, choosing to use or report just those that make them look good. Generally it will take a real expert in the various measures and what they really mean to spot that this is going on. A second important disadvantage is that it is difficult to accumulate knowledge when social science studies use different measures and thus turn out to be studying different things. Third, it becomes difficult to make genuine comparisons since different measures can give different verdicts. What's better on one measure is worse on another. But it may be important to be able to

make comparisons. Consider the HDI indicators for example. It seems important to identify the developing countries that have performed better than others in terms of these indicators so that we can study the relation of this progress to the nature of these economies and to public policies pursued. In civil war studies it is also common to compare different states to see what causes some of them to break out in civil war (say, Burundi) whilst others (say, Kenya) do not. Finally, we know that data collection is very difficult and expensive and we simply are unable to collect the right kind of data for each measure when measures multiply.

All this drives us to try to devise *common metrics* for central social science concepts---one way of characterizing, representing, and proceduralizing that is widely used, researched and reported. But, as we noted, that can distort what we mean, fail to be fit for purpose, and lack nuance, detail, and accuracy. Finding a good balance among all these different aims and needs is one of the central problems that social scientists continuously confront in their efforts to measure society.

2.3 Procedures

In this section, we will discuss what on-the-ground procedures entail, and some problems we face in devising them for measurement of social science concepts. To that end, we will give

23

several examples: *civil war*, *human well-being*, and the *disability adjusted life year (DALY)*. These examples show that often coming to correct procedures means we have to get back to earlier stages in the process of developing measurements (characterization and representation) in order fully to do the concepts justice. Social science measurement is thus a process that involves continuous feedback and refinement, even at the procedures stage.

Let us start by giving a brief account of what we mean by *on-the-ground procedures*. Assume for this purpose that we have found a satisfactory characterization and representation of a concept, e.g. *poverty*. Let's say we have characterized poverty as living on less than 2 dollars a day and represent it as a two-valued variable (i.e. you're poor, or you're not). Now we wish to measure, say for every person over 18 in Cameroon, who is poor and who is not. The methods we use to find out who is poor are called the *procedures*. We may, in this example, look at civil service records already in place; we may conduct a census; or we may gather a representative sample of the population and extrapolate from the data we collect from them. Which procedure is right for any given characterization depends on which one---if any---is most accurate: in this case, which is most accurate in grouping the adults in Cameroon in the poor and the not-poor categories according to the criterion of 'living on less than $2 a day'. However, generally there are also other considerations we must take into account. Some procedures may be more costly or time-consuming to undertake, and some procedures may

be unfair or unethical. So it is not just the degree of accuracy that helps us choose a procedure. It is very much a matter of prioritizing (e.g. we ask: Is the extra accuracy worth the extra time and money?), just as we have seen with choosing characterizations and representations.

This mock example is an easy case, one in which we do not encounter any further difficulties with the concept itself when settling on and carrying out the procedure for conducting our measurement. Unfortunately, social scientists are rarely so lucky. Often, scholars do not measure the parameters they need for their study independently of other scholars' work. Data collection is expensive and difficult. So most social scientists need to depend on shared databases for their figures. In the civil war case there are several databases they can use, such as that collected by the COW or the Uppsala Conflict Data Programme (UCDP) at Uppsala University in Sweden. Each of these databases uses different measures to come to, for instance, the number of fatalities. For instance, the COW database started out by using a death count that refers to military deaths only. In later years this has been changed for civil wars to include civilian deaths, although the matter is still at time of writing under discussion in the COW community. Thus, although social scientists may use the figures in these databases for their own purposes, they have to rely on the procedures someone else has devised and these may not fit well with the concept they need for the purposes of their own

study. It is a constant challenge to social scientists, on the one hand, to devise procedures for data collection that can allow for wide usage and, on the other hand, to try to figure out information needed about their specific concepts from data that measures something different.

Even given that a set of specific procedures has been settled on, there can still be problems on the ground in interpreting them. Measuring fatalities on the ground is difficult. Not only do both sides of the conflict tend to give a skewed image of the number of deaths to favour their own side, there are also difficulties in classifying a death. Say we find a young adult male in civilian clothing who has been fatally shot. We then tend to classify him as a civilian death. What, however, if he was carrying a weapon that he threw on the ground just before he was shot? Is this still a civilian death, or is it a battle death? Social scientists need to be both sure of how their procedures dictate the death should be classified in this situation and sure of what the database collectors they are relying on have done. These two need to match up, but whether they do requires both further specification on the part of the social scientists and clear communication on the part of database collectors.

So far, we've seen that coming to correct procedures involves setting priorities (e.g. accuracy versus costliness) and questioning whether the data we use from others was collected with our aims in mind. Further difficulties in settling on the right procedures come from issues involving specification and de-abstraction. Imagine, for example, that our characterization of *civil war* calls for over 1000 deaths each year, and we correlatively represent civil war as a 2-valued variable. Now someone asks, 'Is Syria at civil war or not?' We then realize that, though they use the same characterization and representation, some scholars classify Syria as being at civil war, while we do not. We look for the differences in the procedures of these social scientists and find to our great surprise that we are only measuring deaths of combatants whereas the other scholars are also measuring deaths of civilians. This then indicates to us that we need to find a more detailed characterization of civil war that specifies which deaths to take into account. Problems with our procedures force us to refine our characterization.

We have already stressed one reason that this problem may come up for social science concepts much more than it does for natural science concepts, like *electrons* and *oranges*: There is no right or wrong characterization of these Ballung, socially constructed concepts. We have a different concept of *civil war* if we count different kinds of deaths, but this is neither a right nor a wrong concept. By contrast, although some thinkers disagree, it is widely

maintained that if we characterized electrons or oranges differently we would simply get the concept wrong.

Of course, there is an upside to the civil war example. Though measuring the number of deaths in a population requires sophisticated procedures, in the end we know that either someone is dead or they're not. Proceduralizing is generally more problematic than this in social science. In the remainder of this section, we will consider two examples that require more sophisticated specification.

Capabilities. Suppose we want to measure well-being, perhaps to see if the well-being of the inhabitants of a country or an ethnic group is improving, or to study how on average well-being might be affected by wealth, education, employment, health, etc., or to compare the well-being of different groups. We can see from Chapter 1 how complex and controversial the concept *human well-being* is. Here we introduce it to illustrate some of the problems raised by abstract concepts and some of the problems solved by using them.

The Human Development Index (HDI), which we described above, is used by the United Nations as a measure of well-being. The HDI represents well-being as the geometric mean of three features, which we can take to be the core of the characterization of this particular concept: education, life expectancy, and income. Let us not for now worry about why to choose the geometric mean but concentrate instead on these three factors. Do these constitute *well-being*? Is that really what we mean by this concept? Probably not. Still they may be relevant to measuring well-being. For instance they might be tools that are relatively necessary to secure well-being, or alternatively they might be features that fairly regularly accompany well-being. In either of these cases they might serve as reasonably good indicators of the degree of well-being in a country. Both these, however, depend on having some other concept of well-being which these features are good indictors of or good tools for securing. Whatever that concept is, it needs a proper characterization if it is to play a serious role in social science. What can it be?

Consider the proposal from Amartya Sen (1999), already discussed in Chapter 1. For Sen, human well-being consists in having a good amount of *substantive freedom*. *Freedom* for Sen involves having choice; *substantive* freedom, in having choice among things we have reason to value. *Substantive freedom* for Sen consists then in having the choice of many *lives worth living*. We can only live one of them, but we have many that we can freely choose among.

Sen represents these ideas using notions of *functioning* and *capability*. A *functioning* is anything you can do or be, like being well or poorly nourished, a doctor, a runner, trapped in an unhappy marriage, or a loving parent. Because of our abilities and the constraints of our positions only certain functionings are available to us and only certain combinations can be taken up in one life. We cannot simultaneously be an Artic explorer and also work daily in your local homeless shelter. Your *capability set* is the set of all the consistent sets of functionings available to you. As with an ordinal scale, this representation provides an ordering of substantive freedoms; but unlike ordinal scales, it is only a *partial order* (which is also the case with tables of indicators). If your capability set contains all the lives worth living that mine does and then some more, you have more substantive freedom than I do. If the sets are just different, no order is decreed. Judgements can be made from case to case but they come from independent sources not from the notion of *substantive freedom* characterized and represented as we have just described.

This characterization is very abstract, using concepts like *substantive freedom* and *lives worth living*. This gives it certain advantages. Although it may be a controversial proposal, in part by virtue of its abstractness it can get far closer to what one might actually intend with a concept of *human well-being* than a more concrete concept like *life expectancy* or *income*

can. It also puts us more squarely in the realm of moral discourse, which, as you see in Chapter 9 and we will discuss briefly in the next section, many would argue cannot be avoided with central social science concepts. As Chapter 1 explains, this is especially true with *human well-being*: Surely what constitutes human well-being depends on what we think constitutes a good for humans and a good way of life.

These two advantages are the flip side of its two central disadvantages. First, moral debate cannot in the end be avoided. Definitions cannot be agreed on just by conducting 'proper science'. To apply Sen's concept of *well-being*, decisions must be made about which combinations of functionings make for *lives we have reason to value* and that is not a purely scientific matter. Second, this kind of characterization and its correlative representation is extremely difficult to operationalize. Just what procedures shall we follow to decide which and how many lives worth living are available to individuals? The problem is not just that it is difficult to come up with procedures or that the procedures might be difficult to carry out. It is that the characterization does not contain enough information to help. For any procedures we settle on, the fit between the concept and the procedure carried out to measure it will be extremely loose. This can result in dramatically different verdicts being given by different procedures for what is supposed to be the same concept. By contrast a far more concrete

concept like HDI will have a far closer fit with the procedures used to carry out measurements of it on the ground.

This is yet another problem that social science must continuously deal with, and without any specified prescriptions for how to do so. There is no easy way to balance the advantages and disadvantages of adopting more abstract versus more concrete concepts or of using concepts that engage moral issues directly versus ones that appear to duck them but may therefore be less true to our intended meanings.

Disability-adjusted life year. As a second briefer example, consider the disability-adjusted life year (DALY). This measure for health liability combines the years of life lost with the years lived with disability in one single measure. In order to formulate a specific procedure for measuring DALYs, social scientists have to specify the concept further. For instance, to find a DALY figure for the death of a 25 year old versus that of a 70 year old, a 'social weighting' is specified in order to compare the value of life depending on age. The life of a 70 year old may be discounted at a specific rate so that it is worth less than that of a 25 year old. As we have stressed in discussing other concepts, these specifications should be sensitive to the aims and values the social science project has in measuring the DALY. Discounting the

lives of the elderly could, for instance, indicate that the project values the potential for further productivity of an individual.

As a second example, to come to a procedure for measuring DALYs we must also specify a trade-off between keeping people alive and enhancing their standard of living. Are five years lived in full health worth the same as ten years lived with a serious disability? So, as with *human well-being,* we see that the DALY too is a case where a very abstract, value-laden characterization needs to be further specified and de-abstracted in order to formulate reasonable procedures for measurement.

3. Values in social science measurement

In the last section, we will give an analysis of how values influence social science measurement. We will not here argue in favour or against the use of values in different kinds of scientific practice, but we will highlight several areas of social science work where the interaction between values and social science measurement is more complex than it may seem at first. We will start by briefly touching on the existing general literature of values in both natural and social science, before applying this literature to the social science measurement case.

3.1 General issues of values in science

As you see in Chapter 9, there is wide agreement within philosophy as well as in both the natural and social sciences that several areas of a scientist's job are influenced by values. First, when scientists decide to study a particular topic rather than another, most agree that they do this for reasons that may not have anything to do with science. It is also widely agreed that this may well have no negative impact on science's search for truth. For example, if a scientist decides to study the global distribution of a particular kind of blue algae because her favourite professor did his doctoral work on this topic, that does not suggest that her results will be biased. Or, if a scholar wants to study the violent conflict in Rwanda rather than Colombia because she is more familiar with African internal politics than with that of South-America, this is not judged as an illicit intrusion of personal interest or values into her scientific work.

Further, once we've measured a concept in a certain way, e.g. once we've measured that 24% of Cameroon's population lives under the $2 per day poverty line, then what we subsequently do with that information depends on value judgements. For instance, we may decide to recommend that our government should increase the amount of development aid we send to Cameroon, or we may decide that this situation is not bad enough to deserve this

recommendation. Similarly, once we've discovered nuclear fission, applying that research to build nuclear weapons is based on a value-judgement. We may disagree with building weapons because we do not share these values, but that is not grounds to argue that the application is not going to work, i.e. that the bombs are not going to go off.

There are also aspects of science that may be negatively impacted if value judgements or personal interests play a crucial role. For instance, the traditional view in philosophy of science is that values ought not to play a role in gathering evidence and in choosing what hypothesis or theory to accept. This normative standpoint that scientists may not make value judgements in carrying out these practices is, as we see in Chapter 9, called the *value-free ideal*. For instance, when we are gathering evidence, or when we are categorizing data into groups, or when we are checking if certain evidence supports a hypothesis, the value-free ideal dictates that the scientist not let values supplant the evidence. This means that we do not want scientists to say the evidence supports a hypothesis simply because they agree with the hypothesis or on whim or for personal or political reasons. Nor do we want scientists to consider only evidence that supports their initial beliefs; nor to claim that one theory is more likely to be true than another because they are paid to get that outcome or because having that theory adopted advantages some group they support.

But Amartya Sen points out a closely related social science practice that, he argues is necessarily value-laden: the action of giving a certain account (Sen 1983). Say a social scientist has done a statistical study into the causes of ethnic civil war and finds that it could be prevented by isolating different ethnic minorities in a country from one another by mass displacement. The question is: Would the action of publishing this result be right? Sen maintains that this is a kind of question the social scientist cannot avoid (nor, in fact, can any other type of scientist). Publishing is an action one performs and one's actions can always be subject to moral scrutiny, whether one does them wearing the hat of a social scientist or as an ordinary citizen or as a parent, a teacher, a grocer, etc. Whether it is or isn't ethically wrong is not our topic. What we want is to alert you to Sen's claim that *the question needs to be asked by the social scientist*---that giving an account, publishing a statement, choosing a research topic, and all else that social scientists do, these are all actions; and as actions, they subject to moral scrutiny. So their evaluation cannot avoid value judgements.

One may argue against Sen's view, thinking: Doesn't truth (and therefore also giving a true account) trump all other considerations in science? Is it not the task of the scientist to uncover truth at all costs, that is, regardless of the consequences that outing the truth will have? Perhaps this is universally correct. Or perhaps it is correct over some questions and not others---like carrying out research on biological warfare or on the Manhattan project to build

the atom bomb, both of which have a number of arguments pro and con. Perhaps it depends on how immediate and easy to predict the consequences are. Perhaps not. The insight we get from Sen's arguments is that the answers to these questions themselves require value-judgements, including judgements about what the aims of science should be.

3.2 Values in social science measurement

With this backdrop in mind, we can revisit the case of social science measurement. What we find is that most of the issues that arise there may be interpreted as issues of values influencing science, but in a unique social science way that does not always match up with the general problems of values in science just described.

First, however, let's look at the overlap. Just as happens in science in general, the motivation behind choosing a certain field to study may be personal. A scholar might decide to measure ethnic diversity rather than poverty because she is part of an ethnic minority. Similarly, applying theory in a certain way is a value judgement. Using a statistical study's result to try to influence NGO policy rather than governmental policy is our own choice. There are, however, many more areas of social science measurement where values necessarily play a role.

We've already come across one aspect of social science measurement that is value-laden in section 2.1, where we saw that in the characterization stage of measurement, social scientists work with an uneasy tension between two aims. First, there is the aim of studying those concepts that they and the society they work in are concerned about. The father of sociology, Max Weber, argued that this is essential to social science (Weber 1949). It is the job of social science to study the things that our societies care about---which in large part depends on what we value and disvalue. Second, there is the aim of finding concepts that aid in explanation, prediction and control. As we've noted, it is not always easy to bring these aims into accordance. Sometimes the concepts we're interested in (like *civil war* or *poverty* or *well-being*) cannot be brought into the kinds of systematic relations with other concepts that allow prediction and control. There might, for instance, be no specific features that can help us to predict or explain those conflicts we want to label as civil wars, nor any interventions that would have a good chance of preventing even a reasonable percentage of conflicts we want to label that way. Similarly there may not be any significant systematic effects we can expect from conflicts falling into this category. Pushed by the second aim, we may be forced to abandon the concepts we care about and to study others that are not really the same. We may refine our concept of civil war, for instance, by looking at ethnic civil wars, even though we were interested in the more general concept of civil war in the first instance.

A special case of this is the refinement of Ballung concepts. As we have seen, a Ballung concept is characterized by family resemblance and can be usefully represented by a table of indicators. This makes a reasonable level of prediction, explanation and control unlikely since some features in the table may have one set of causes, other features very different causes, and some may themselves not be the kinds of features that have stable systematic relations with others. For predicting and control we generally need more narrow concepts and especially ones chosen because they do have definable causes and effects. This leads us to refine our Ballung concepts, substituting for them other concepts that do not have the same meaning but are better for prediction and explanation. The way in which we refine a Ballung concept is, however, not uniquely determined. There are many ways of doing it and what concepts we end up with will, therefore, depend on value judgements. As we see from Chapter 9 and in the discussion of *human well-being* here and in Chapter 1, one reason that social science measures are often value-laden is that social science concepts often do what Weber says they should: They refer to things we care about, things we value and disvalue.

A second reason that social science measures are value-laden, more so probably than measures of natural science concepts like *neutrino stars*, *finches* or *natrium chloride*, is one we have mentioned repeatedly, that many social science concepts are not already there in

nature for us to find out about. They are socially constructed concepts. This makes them susceptible to value judgement, because there is in principle no right or wrong characterization over and above the demand they help us achieve the aims of our social science projects. This situation then is similar to the initial choice of the social scientist to study ethnic diversity rather than poverty; her characterization of ethnic diversity will be inherently value laden as well. There is no unique right way of doing it.

By this point, it will be clear that a similar situation to the one for concept characterization holds for representation and procedures; there, also, there is no unique right way of doing things. Thus, social science measurement is value-laden through and through. In general, we can say that *the aims we have with social science measurement decisively influence the results we will find.* If we aim to find a causal mechanism that tells us what properties of a country cause ethnic civil war, we will characterize, represent and proceduralize the concept *civil war* differently than when we aim to give an explanation for the development of the conflict in Syria between 2010 and 2012 and different again if we aim to count how many civil wars there are in the decade 2000-2010 as opposed to 1950-1960 using some broad Ballung-type concept of *civil war* that captures something that we care about far better than any of the more refined concepts. These aims are value-laden. Thus, social science measurement is value-laden.

Suppose though that characterization, representation and procedures have been settled on for some purposes. What of the scientific work that comes after this, the categorization of individuals? We may want to insist that once the characterization and procedures have been specified, the decision of what individuals fit the concept should be based on facts alone, and not on values. Consider for example the measurement of which individuals have Asperger syndrome. This measurement is important for social policy as it may determine whether or not an individual qualifies for benefits or special arrangements in the workplace. Now assume that a characterization, representation and procedure for measuring whether a person has Asperger have been settled on. The procedures might, for instance, require individuals to score a certain number on the autism spectrum disorder severity scale. Then we may want to ensure that the classification of an individual as having or not having Asperger should not be influenced by values, for instance by the judgement that this particular individual is especially deserving of help.

This is one reason we see such attention to the details of the procedures and the recent drive to 'mechanize' classification as far as possible. There is a concern that when judgements enter, values may play a role they should not, whether consciously or unconsciously. On the other hand, it is difficult in a fuzzy world to design systems that cover all cases and our best

efforts can end up making absurd rulings in many real applications. Also, as we have seen, there is often equally good reason to choose different measures and sometimes we feel we are confronted with a case that should fit the label, and does under a different measure than the one that has been chosen, perhaps chosen arbitrarily. How should we weigh up the dangers of over mechanization versus the dangers of conscious and unconscious bias? Here is yet another area where good decisions require delicate case-by-case balancing.

4. Conclusion

This chapter has discussed measurement in the social sciences. In it we have seen that both assigning a number to an individual unit (e.g. GDP to a country) and assigning the unit to a specific category (e.g. calling the Syrian conflict a civil war) are instances of measurement. Measurement involves three mutually interlinked stages: characterization, representation, and on-the-ground procedures. We have seen that how social science should fill these in depends on the aims of the research. So, a social science project that aims to prevent the mass displacement of civilians due to internal armed conflict may choose different ways of characterizing and representing and different on-the-ground procedures than one that aims to find a causal mechanism linking some economic variables of a country with civil war duration. We have also seen how designing good measures requires trade-offs and balancing among a great variety of different demands and aims.

Asking whether Syria is at civil war is not sensible unless we say to what end we would like to classify Syria as at civil war or not. If we want to know whether the conflict will have certain effects, so that we can act to prevent these, then we will most likely give a different answer than if we wanted to explain the development of the conflict since 2010. Neither of these two answers will be simply right or wrong; they will only be right for a certain purpose.

References

Black, I. (2012). "Syria: Ban Ki-moon warns of catastrophic civil war," *The Guardian*, May 31. http://www.guardian.co.uk/world/2012/may/31/syria-ban-ki-moon-civil-war. Accessed September 29, 2013.

Collier, P. and Hoeffler, A. (2004). "Greed and Grievance in Civil War," *Oxford Economic Papers*, 56(4): 563--595. Figure on 581.

Mason, J. (2012). "Obama says China, Russia 'not signed on' for Assad's removal," *Reuters*, July 19. http://www.reuters.com/article/2012/06/20/us-g20-syria-obama-idUSBRE85J01620120620. Accessed September 29, 2013.

Neurath, O. (1936). "Physikalismus und Erkenntnisforschung," *Theoria,* 2: 97--105 and 234--7. Translated as "Physicalism and Investigation of Knowledge," in R. Cohen and M. Neurath (eds.), *Philosophical Papers 1913-1946.* Dordrecht: Reidel. 1983. 159--167.

Sambanis, N. (2001). Do Ethnic and Nonethnic Civil Wars Have the Same Causes?: A Theoretical and Empirical Enquiry (Part 1)," *The Journal of Conflict Resolution,* 45(3): 259--282.

Sambanis, N. (2004). "What Is Civil War?," *Journal of Conflict Resolution,* 48(6): 814--858.

Sen, A. (1983). "Accounts, Actions and Values: Objectivity in Social Science," in C. Lloyd (ed.), *Social Theory and Political Practice: Wolfson College Lectures 1981.* Oxford: Clarendon Press. 87--107.

Sen, A. (1999). *Development as Freedom.* Oxford: Oxford University Press.

Weber, M. (1949). *The Methodology of the Social Sciences.* Glencoe: Free Press.

Stevens, S. (1951). "Mathematics, measurement, and psychophysics," in S. Stevens (ed.), *Handbook of Experimental Psychology.* New York: Wiley.

Further readings

Sarkees, M. (2000). "The Correlates of War Data on War: An Update to 1997," *Conflict Management and Peace Science,* 18(1): 123--144.

Cartwright, N., and Bradburn, N. (2011). "A Theory of Measurement," in R. M. Li (ed.), *The Importance of Common Metrics for Advancing Social Science Theory and Research: A Workshop Summary*. Washington: The National Academies Press. 53--70.

Cartwright, N. and Chang, H. (2008). "Measurement," in S. Psillos, and M. Curd (eds.), *The Routledge Companion to Philosophy of Science.* New York: Routledge.

Reiss, J. (2008). *Error in Economics: Towards a More Evidence-Based Methodology*. London: Routledge. Chapters 2--4.

Reiss, J. (2013). *The Philosophy of Economics: A Contemporary Introduction*. New York: Routledge. Chapter 8, "Measurement".