1 Sets, set membership, and calibration

Easy reading guide

This book is based on the conviction that the tools of set theory allow for a distinct and fruitful perspective on social science data. In order to develop the argument and to show how the analysis of empirical data works when focusing on set relations, we first clarify how sets refer to concepts (1.1). Then we discuss how set membership scores are derived from empirical and conceptual knowledge. This process is called calibration (1.2). Through calibration of sets, qualitative – and also quantitative, with fuzzy sets – differences between cases are established and expressed by set membership scores that vary between 0 and 1. The usefulness of set-theoretic methods depends on the proper calibration of sets. Beginners should read through the whole chapter with careful attention, while more advanced users might wish to skim through the text if they feel that they are well aware of the principles and practices of good set calibration.

In the Introduction, we have already mentioned that there are two major variants of QCA, namely crisp-set QCA (csQCA, where a case is either a member of a set or it is not) and fuzzy-set QCA (fsQCA, where differences in the degree of set membership can be captured). Both these variants share one fundamental feature: they establish *qualitative* differences between those cases that are (more) in the set and those that are (more) out of the set. Beyond this, both QCA variants have much more in common than is sometimes insinuated in some of the literature. In this book, we therefore emphasize their commonalities. They both aim at identifying subset relations, which, in turn, rest on qualitative differences between cases. Indeed, a crisp set should be seen as the most restrictive form of fuzzy set, one that allows only full membership and full non-membership. Because crisp sets are a special case of fuzzy sets, most of the set operations equally apply to both variants. For all these reasons, we introduce both variants together. Admittedly, crisp sets correspond more to everyday thinking: this is why we introduce all important notions and operations by first explaining their meaning based on crisp sets. The main emphasis of this chapter is on fuzzy sets, though, because they are less intuitive and therefore require more explanation.

1.1 The notion of sets

1.1.1 Sets and concepts

The use of the term "set" is not very broadly diffused in social science methodology. However, a good part of our conceptual reasoning, as Mahoney (2010) shows, is at least based on an implicit idea of sets. According to Mahoney, there are two basic modes of looking at concepts: if we define concepts "as a mental representation of an empirical property" (Mahoney 2010: 2), then we will measure cases "according to whether or the extent to which they are in possession of the represented property" (Mahoney 2010: 2). Measurement theory provides us with many useful techniques for doing this. This ultimately results in the use of variables when defining a concept (Mahoney 2010: 13). If, however, we refer to concepts as sets, defined in terms of "boundaries that define zones of inclusion and exclusion" (Mahoney 2010: 7), then "[c]ases are measured according to their fit within the boundaries of a set" (Mahoney 2010: 2). Sets work as "data containers" (Sartori 1970: 1039). Although this seems to be a subtle and often overlooked differentiation, these two views of concepts are fundamentally different. When we measure a concept by means of traditional measurement theory, it represents a property or a group of properties. The set-theoretic view, instead, uses set membership in order to define whether a case can be described by a concept or not. Therefore, in the framework of set-theoretic methods, issues of concept formation have a somewhat different connotation than in traditional measurement theory, by focussing on whether a case belongs to a concept (i.e., a set) or not. This process of assigning set membership is also called "calibration" (see section 1.2).

1.1.2 The pros and cons of crisp sets

When QCA was first discussed in the 1980s and 1990s, it was limited to crisp sets. This required a decision whether a case is a member of a set or not. As such, this also corresponds to how sets are generally perceived, namely as boxes into which cases can be sorted or not. However, as argued in the Introduction, it is not always easy to make such clear-cut decisions, above all when dealing with more fine-grained social science concepts for which detailed and nuanced information is available. Not surprisingly, the need for "dichotomization" has triggered some serious criticism of crisp-set QCA (Bollen, Entwisle, and Alderson 1993; Goldthorpe 1997; for an overview and

a response, see De Meur, Rihoux, and Yamasaki 2009). This requirement certainly affected the usability and the acceptance of QCA in its early stages. The two major reservations with dichotomies seemed (and still seem) to be that (a) they represent a loss of empirical information and (b) they reduce the robustness of results due to the sensitivity of QCA findings to decisions on where to put the threshold for dichotomization, as the latter is often subject to a relatively large degree of discretion.

At the core of the argument against dichotomization is the belief that the world and large parts of social science phenomena simply do not come in a binary form. Let us take, for example, the notion of democracy again: if we think of cases such as the UK or the USA on the one hand, and North Korea or Zimbabwe on the other, then this might at first glance suggest that a clear-cut dichotomy is appropriate. The former countries are members of the set of democracies, whereas the latter two are clearly not. However, cases often fall in-between these two qualitatively different endpoints. Just think of all the so-called "electoral democracies" or any of the numerous "democracies with adjectives" (Collier and Levitsky 1997) identified in the literature. A closer look at the unquestionably democratic cases in North America and Western Europe also reveals the existence of interesting and analytically relevant differences – both across time and across countries – that defy a straightforward classification as democracies versus non-democracies (for instance, declining trust in the political class or the rise of far-right movements might be said to undermine democracy). We would probably not want to claim that any of these countries has become *un*democratic. Despite sometimes even strong deviations from perfect democracy, they are still *qualitatively* different from non-democracy. As we shall see, fuzzy sets provide the possibility to take both qualitative and quantitative differences into account.

The fact that we emphasize qualitative differences and not only quantitative variations is quite important here. In statistics, interval-scale variables are usually considered superior to dichotomous (and ordinal) variables, since their high level of measurement captures more precise quantitative differences. However, the previously mentioned limitations of dichotomous variables should not lead to the conclusion that interval-scale measurements automatically imply a greater level of validity. This is above all doubtful when the underlying concept establishes explicit *qualitative* distinctions between cases, such as, for instance, the concept of "democracy." This implies that, despite the general concerns about the use of dichotomies, not using them at all would go too far. In fact, even in applied quantitative research, where most critiques of the use of dichotomies originate, techniques like logistic

regression, which requires a dichotomous dependent variable, remain widely popular. What is more, the recent shift in the statistical literature towards the experimental design as the gold standard for causal inference has led to a renewed appreciation of dichotomies even among proponents of advanced quantitative methodology.[1]

The second type of critique aimed at using dichotomous data may seem to be a rather technical issue, but it refers back to the critique just mentioned. It is often argued that the decision on where to put the threshold is not only to a considerable extent arbitrary, but also crucially influences the results obtained. What seems to be true is that in research practice, scholars have all too often been using unconvincing criteria as to where to put the threshold for turning their raw data into crisp-set membership scores. As we will explain in section 1.2.2, a very common mistake is to use characteristics of the data at hand, such as the mean or median, as a guide to where to put the threshold.

A central critique says that arbitrariness, or simply a definition that is not perfectly accurate, could cause a case to be on the "wrong" side of the threshold, and that research results could be significantly altered through different case assignments. While true, claims about the manipulability of set-theoretic results through purposeful threshold setting (aka cheating) are largely exaggerated. First, for each concept there is only a certain, often small range where the threshold can plausibly be put. Usually, no huge differences in the results occur due to minor adjustments to the threshold.[2] If the criteria for setting the thresholds are both transparent and plausible, then hardly any chance exists for potential cheating. Finally, the effects of different thresholds on the results obtained are often so intricate that setting thresholds in order to create desired results would be a time-consuming and futile exercise for the researcher.

In sum, working with crisp sets does create some issues. At the same time, when trying to investigate relations between sets, we must establish qualitative differences between cases that are more in a set and those that are more out of the set. So what can we do in order to effectively work with concepts where there is some interesting variation between the qualitative endpoints of implicitly dichotomous social phenomena? In these situations, neither interval scale variables nor dichotomous crisp sets are ideal. The former lack the capacity to establish qualitative differences, and the latter to make differences in degree between cases of the same kind. Thus, an instrument is needed that overcomes the starkly limiting characteristics of dichotomies but which at the

[1] We thank John Gerring for making this point (personal communication, Spring 2010).
[2] See section 11.2 on robustness tests in QCA.

same time continues to possess the potential to show qualitative differences. To this end, Ragin proposed the use of fuzzy sets (Ragin 2000).

1.1.3 Properties of fuzzy sets

The term "fuzzy set" goes back to the writings of Lotfi Zadeh (1965, 1968). The notion of fuzzy sets has triggered volumes of books in disciplines as diverse as mathematics, engineering, and philosophy. Only recently has the tool of fuzzy sets been introduced in the social sciences (Smithson 1987, 2005; Ragin 2000, 2008a, 2008b; Smithson and Verkuilen 2006). Thus, fuzzy-set theory was not invented by social scientists, and the level of complexity of this theoretical and mathematical framework goes well beyond that currently applied in fuzzy-set social sciences.

Because fuzzy-set theory refers to an established body of literature, we stick to the use of the term "fuzzy set" despite its potentially misleading interpretation and negative connotation in everyday language. One could perhaps come up with a less stigmatized adjective for sets that are not crisp, but the use of any other term would contribute to disconnecting the use of fuzzy sets in social sciences from their mathematical and epistemological background. As the extant literature makes clear, "fuzzy" does not mean "unclear" or "wishy-washy." The statement that a given case has a fuzzy-set membership score of, say, 0.8 reflects precise empirical information about that case. The fuzziness stems from imprecise conceptual boundaries. For instance, when we invoke the concept of a "bald person," we all agree that somebody with no hair at all is definitely bald. If, however, we took a person with a lot of hair and started pulling it out one strand after another, it would be difficult to point to a precise and quantifiable amount of remaining hair at which this person would have to suddenly be considered a member of the set of bald people. At the same time, we do see a qualitative difference in terms of baldness between somebody with a lot of hair and somebody with only few hairs. The problem of identifying where exactly the difference is between a bald and a non-bald person is not resolved by knowing the precise number of hairs remaining. Fuzziness, in other words, is due to *conceptual boundaries* that are not sharply defined rather than imprecise *empirical measurement*.

Fuzzy sets preserve the capability of establishing difference-in-kind between cases (qualitative difference) and add to this the ability to establish difference-in-degree (quantitative difference) between qualitatively identical cases. The term fuzzy set implies a different usage of the term "set" than we are used to from traditional set theory, which defines sets through strict membership criteria (Klir, Clair, and Yuan 1997: 48). Individual members either clearly belong to sets, or else they do not. Fuzzy sets, by contrast, allow for cases to have

partial membership in the set (Klir *et al.* 1997: 73ff.). Cases can be more in than out of a set without being full members of the set, and they can be more out than in the set without being full non-members of the set. For instance, two countries might have a fuzzy-set membership score of 0.7 and 0.8 in the fuzzy set of democracies, respectively. This indicates that both are rather more democratic than non-democratic (a qualitative property), but also that one of the two countries is slightly more democratic than the other one (a quantitative difference). Fuzzy sets are thus characterized by the fact that the boundaries between membership and non-membership are blurred. This also implies that a case – unless it has full (non-)membership in the set – is actually a partial member of both the set *and* its negation. In our example, each state is not only a member to some degree in the fuzzy set of democracies, but also of the opposite fuzzy set, that of non-democracies. The principle of the "excluded middle" whereby an element can be only a member of a set *or* of its complementary set (a fundamental rule of crisp sets) does not hold for fuzzy sets.

Fuzzy sets allow for degrees of membership, thus differentiating between different levels of belonging anchored by two extreme membership scores at 1 and 0 (Ragin 2000: 154; Ragin 2008b). In addition, a membership score at 0.5 locates the so-called point of indifference where we do not know whether a case should be considered more a member or a non-member of the set (Ragin 2000: 157). It constitutes the threshold between membership and non-membership in a set – the qualitative distinction that is maintained in fuzzy sets – and represents the point of maximum ambiguity with regard to a case's membership in the concept. Fuzzy sets explicitly require that the definition of set-membership values is based on three qualitative anchors: full set membership (1), full non-membership (0), and indifference (0.5). In crisp sets, these three anchors are all collapsed into one – the distinction between full membership and full non-membership.

Defining the precise location of the 0.5 qualitative anchor is crucial. Assigning cases a 0.5 fuzzy set membership score, however, should be avoided. It means that we are unable to say for an individual case whether it is more a member of the set or more a non-member. Because we avoid a decision on the qualitative status of the case in question, assigning the 0.5 score has important consequences for the analysis of fuzzy data that we explain in detail in Chapters 4 and 7. For all other degrees of membership and non-membership so-called fuzzy values are used to quantify the levels of membership of a case in a set. As Table 1.1 exemplifies, for each fuzzy value, linguistic qualifiers can be assigned (Ragin 2000: 156).

It is not necessary for there to be actual empirical elements corresponding to every fuzzy value, i.e., even if a fuzzy set allows for a membership of, say, 0.8 it might well be that it is not assigned to any empirical case. In particular,

Table 1.1 Verbal description of fuzzy-set membership scores

Fuzzy value	The element is …
1	Fully in
0.9	Almost fully in
0.8	Mostly in
0.6	More in than out
0.5	Crossover: neither in nor out
0.4	More out than in
0.2	Mostly out
0.1	Almost fully out
0	Fully out

Adapted from Ragin (2000: 156)

this also applies to the membership values of 1 and 0. Also, different intervals between the fuzzy-set membership scores are possible: it is perfectly fine if a fuzzy set shows membership scores of, say, 0.1, 0.4, 0.6, and 1, if theoretical considerations warrant it.

We can also imagine fuzzy scales that are differentiated even further than this. However, with increasing levels of differentiation it becomes ever more difficult to come up with theory-based and empirically observed distinctions between the values, not to mention the need to assign verbal descriptions to each value. Any such representation suggests a level of precision that is unlikely to be grounded in empirical information or theory. One should therefore not over-interpret the substantive meaning of marginal differences in set-membership scores, such as the difference between 0.62 and 0.63. Such small differences also have only a negligible impact on the analytic results.

Note that frequently some of the variation in the raw data is conceptually irrelevant. When translating raw data into corresponding fuzzy-set membership scores, this must be taken into account. Imagine that we want to assign membership scores of all countries in the fuzzy set "rich countries." If we take GDP per capita as an indicator for richness, then we find a large variation among the four countries with the highest GDP per capita (IMF data for 2010): Qatar ($88,500), Luxembourg ($81,400), Singapore ($56,500), and Norway ($52,000). Under many (if not most) definitions of "rich country," all four would be considered rich and would thus receive a membership score of 1 in the set of rich countries. The fact that Qatar is quantitatively about 1.7 times richer than Norway is deemed qualitatively irrelevant for research purposes (Ragin 2008a: 77ff.).

Fuzzy scales, with their well-defined starting- and end-points, the cross-over point, and the combination of both qualitative and quantitative differentiations, seem to defy standard classifications of measurement levels (Ragin 2008b). Both

the idea of seeing them as continuous scales (since every possible grading between 0 and 1 can be obtained) and seeing them as ordinal scales (since they display an ordered list of empirical representations of a given concept) could seem reasonable. However, the argument against interpreting fuzzy sets as continuous scales is that it downplays the establishment of *qualitative* differences between cases above and below the 0.5 anchor, which remains the essential principle of fuzzy sets. The step from a fuzzy value of 0.4 to 0.6 is something different from the step from 0.1 and 0.3. Although the quantitative difference in the degree of membership is 0.2 in both situations, there is a qualitatively different situation: in moving from 0.4 to 0.6, the qualitative anchor of 0.5 is crossed. While 0.6 indicates that the case is more like a member of the set, 0.4 tells us that it is more of a non-member of the set. The fuzzy values 0.1 and 0.3 indicate, instead, that both cases are on the same side of the point of indifference and thus both indicate non-membership, although to different degrees. This distinction does not, however, also mean that a fuzzy set will be reinterpreted as a dichotomy in the analysis: although the qualitative difference is maintained, the quantitative gradings also count. A fuzzy value of 0.3 describes something different from the fuzzy value of 0.1, although both values indicate the absence of the concept rather than its presence. Hence, fuzzy scales are neither continuous nor ordinal, since their "continuity" and their "rank order," respectively, are interrupted at the point of indifference, and since the inherent qualitative difference is dominant in the definition of the values.

Ragin (2008b) points out that this combination of qualitative anchors and quantitative gradings, which sits uneasily with mainstream social science classifications of measurement levels, is standard in disciplines that are usually regarded as more "scientific" than the social sciences, such as physics, chemistry, and astronomy. Ragin gives the example of "temperature" and the measurement "degrees Celsius." There are senses in which a temperature can be qualitatively interpreted. When falling below 0° or rising above 100°, the state of water qualitatively changes: it turns into ice and vapor, respectively. Hence, a 10-degree change from 95° to 105° implies a qualitative difference, whereas a change from 30° to 40° does not. Just using temperature at face value, without anchors that establish qualitative differences, one would miss this important information about the state of water. So far, in the social sciences it is rare to use knowledge ("the temperature at which water freezes or boils") that is external to the raw data ("mercury expanding and contracting with heat") to decide how to calibrate a scale.

1.1.4 What fuzzy sets are not

Fuzzy sets express a specific kind of uncertainty and take on values between 0 and 1. It is perhaps because of these two characteristics that fuzzy set membership

scores are sometimes interpreted as probabilities (e.g., Altman and Perez-Linan 2002: 91; Eliason and Stryker 2009). We side with those scholars who reject that view, among them Zadeh (1995) himself, whose article's title captures the essence of the argument: "Probability Theory and Fuzzy Logic are Complementary Rather than Competitive." A similar point is made by McNeill and Freiberger (1993: 185ff.), who argue that uncertainty has various aspects and that probability and fuzziness capture different forms of uncertainty. The following example helps to illustrate the difference between probability and fuzzy values.

Imagine two water glasses, each containing a different liquid, and about which the following is known. Glass A contains a liquid that has a 1 percent probability (0.01) of being poisonous. Glass B, on the other hand, contains a liquid that has a fuzzy-set membership score of 0.01 in the set of poisonous liquids. When forced to choose between the two (and assuming that we do not have suicidal tendencies), which glass is safer to drink? The answer is glass B. We know exactly what is in this glass – a liquid that is all but fully out of the set of poisonous drinks. This applies, for example, to energy drinks of the kind that are popular among college students; they are certainly not poisonous, but also not completely free of toxins as is, say, a glass of pure spring water. In contrast, we do not know what is in glass A. It is either extremely poisonous or completely non-toxic. All we know is that it comes from a population of other glasses, of which 1 out of 100 is deadly poisonous. There is a 99 percent chance that drinking from glass A is completely safe, but a 1 percent chance it will turn out to be lethal. In contrast, glass B will cause us to feel, at best, slightly bloated and a little twitchy but does not present any risk of dying.

At-a-glance: the notion of sets

The use of set theory in the social sciences requires a different perspective on concepts: cases are assessed with regard to their membership in previously defined sets.

Crisp sets are restricted to the membership values 1 (full membership of a case in a set) and 0 (full non-membership). This ultimately requires the definition of all concepts as dichotomies.

Fuzzy scales possess three qualitative anchors – the complete presence of a concept (1), its complete absence (0), and the point of indifference (0.5) – with quantitative gradings representing the degree of presence of the concept. Verbal descriptions ("linguistic qualifiers") help to connect the quantitative assessment to natural language.

Crisp sets can be seen as special cases of fuzzy sets. Thus, the rules for fuzzy sets are more general and subsume those for crisp sets.

A fuzzy-set membership score does *not* express the probability of a case's membership in a set. Fuzzy scores and probabilities express different aspects of uncertainty. The uncertainty expressed in fuzzy sets stems from conceptual rather than empirical imprecision, which, in turn, is inherent to most verbally defined concepts – especially those in the social sciences.

1.2 The calibration of set membership

Assigning set membership scores to cases is crucial for any set-theoretic method. The process of using empirical information on cases for assigning set membership to them is called "calibration." In order to be analytically fruitful, calibration requires the following: (a) a careful definition of the relevant population of cases; (b) a precise definition of the meaning of all concepts (both the conditions and the outcome) used in the analysis; (c) a decision on where the point of maximum indifference about membership versus non-membership is located (signified by the 0.5 anchor in fuzzy sets and the threshold in crisp sets); (d) a decision on the definition of full membership (1) and full non-membership (0); (e) a decision about the graded membership in between the qualitative anchors.

1.2.1 Principles of calibration

The first (and very simple) answer to the question of how to assign set-membership values is to base the calibration on the combination of theoretical knowledge and empirical evidence (Ragin 2000: 150). It is the responsibility of the researcher to find valid rules for assigning set-membership values to cases. The top priorities of this process are to make the calibration process transparent and to make it lead to a set that has high content validity for the concept of interest. When turning raw data into set-membership scores, researchers make use of knowledge that is external to the data at hand (Ragin 2008a, 2008b). Such knowledge comes in different forms and from different sources. There are, for instance, *obvious facts*. For example, it is generally true that completing the twelfth grade in the United States leads to receiving a high school diploma. If we are trying to calibrate the set "high school-educated citizens," there is a qualitative difference between completing the eleventh grade and completing the twelfth grade. There are also some *generally accepted notions in the social sciences*. In addition, there is *the knowledge of the researcher accumulated in a specific field of study or specific cases*. This requires extensive fieldwork and a very careful analysis of primary and secondary sources before proceeding to the actual calibration. As such, interviews, questionnaires, data obtained with participant observation or focus groups, and organizational analysis, quantitative and qualitative content analysis, etc., can all provide useful information sources in the process of set calibration.

1.2.2 The use of quantitative scales for calibration

Multiple non-quantitative data sources are often used for calibration. Sometimes, however, we do have one data source and it is an interval-scale measure. For instance, if we want to calibrate the set "rich countries," then a GDP per capita indicator might provide a reasonably good source of information.[3] When interval-scale data are at hand, researchers have several calibration options. In this section, we first describe what one should *not* do when calibrating sets based on interval scales. We then provide a good example of how to combine case knowledge and empirical distribution for meaningful set calibration. Then, in a separate section, we describe the direct and indirect methods of calibration (Ragin 2008a, 2008b).

When calibrating fuzzy sets, it might be tempting to simply transform the GDP per capita scale into the 0–1 interval while preserving each case's relative distances to each other.[4] When calibrating a crisp set, we might even simply want to use the arithmetic mean or the median and to define all cases above the mean or median as "in the set" and the others as "out of the set." Such purely data-driven calibration strategies are fundamentally flawed, though. Measures like the mean or median are properties of the data at hand and, as such, void of any substantive meaning vis-à-vis the concept that one aims to capture with a set. Just dropping or adding a case with an extreme value on the GDP per capita scale will change the mean. Using parameters such as the mean therefore implies that the classification of a case does not only depend on its own absolute value, but on its relative value with regard to other cases. Why, however, should the presence or absence of specific cases in the data influence the set-membership score of other cases in the set of rich countries? It should not.

This is why calibration must also make use of criteria for set membership that are *external* to the data. Certainly this does not mean that the distribution of cases on our raw data should be disregarded. It is simply another piece of evidence, but certainly not the sole guidance when calibrating. Along these lines, also consider that depending on the research context, one and the same raw data translate into different set-membership scores. This is so because the meaning of concepts, and therefore their respective sets, is highly dependent on the research context (Ragin 2008a: 72ff.). For example, in research on EU member states, a GDP per capita of, say, $19,000 (roughly the value for

[3] Here we sidestep the substantive arguments against using GDP as a proxy for "richness" (see, e.g., Dogan 1994).

[4] The easiest method here would be to simply divide the GDP of each state by the highest value of GDP in the sample.

Table 1.2 Calibration of condition "many institutional veto points"

Country	Federalism, 1945–96	Bicameralism, 1945–96	Combined indicator	Fuzzy-set membership in "many institutional veto points"
Australia	5	4	10.00	1.00
Austria	4.5	2	7.00	0.67
Belgium	3.1	3	6.85	0.67
Canada	5	3	8.75	1.00
Denmark	2	1.3	3.63	0.00
Finland	2	1	3.25	0.00
France	1.2	3	4.95	0.33
Germany	5	4	10.00	1.00
Ireland	1	2	3.50	0.00
Italy	1.3	3	5.05	0.33
Netherlands	3	3	6.75	0.67
New Zealand	1	1.1	2.38	0.00
Norway	2	1.5	3.88	0.00
Portugal	1	1	2.25	0.00
Spain	3	3	6.75	0.67
Sweden	2	2	4.50	0.33
Switzerland	5	4	10.00	1.00
UK	1	2.5	4.13	0.00
USA	5	4	10.00	1.00

Source: Emmenegger (2011)

Hungary) would not translate into full membership in the set of rich countries. In the context of a global study, in contrast, Hungary would be a member of the set of rich countries. Set-membership values are intrinsic to the research in which they are used. They are not universal indicators of a concept (Collier 1998: 5), but directly depend on the definition of a concept, which in turn is closely linked to the research context.

A good example to illustrate the calibration of fuzzy sets based on quantitative data is Emmenegger's (2011) work on job security regulations in selected OECD countries. One of his conditions is the fuzzy set "many institutional veto points." The raw data consists of an additive index based on Lijphart's (1999) data on federalism and bicameralism (Table 1.2). Emmenegger opts for a four-value fuzzy scale (0, 0.33, 0.67, and 1). The location of the qualitative

anchors – the most important decisions to be made when calibrating sets – is derived in the following manner. All countries achieving a score lower than or equal to that of the UK (4.13 in Emmenegger's combined indicator) receive a fuzzy membership score of 0 in the set of "many institutional veto points."

Case knowledge is used in an exemplary manner in order to identify and justify meaningful qualitative anchors on the composite index that separates cases with full non-membership and partial non-membership. A prominent gap in the combined indicator between the raw values of 5.05 and 6.75 is then used to establish the point of indifference. All countries below that gap, but above the UK, are assigned a fuzzy value of 0.33. Finally, another gap in the combined indicator between 7.00 and 8.75 is used to define full set membership: countries higher than 8.75 are deemed full members of the set of "many institutional veto points."

While there might be room for debate about specific decisions in Emmenegger's strategy (e.g., the choice of the indicators or the way of aggregating them), the level of transparency and the combined use of conceptual and case knowledge for imposing qualitative anchors represent a good standard of calibration practice. It allows readers to follow the reasoning behind calibration decisions and to either agree or to disagree and, if the latter, to make specific suggestions for change in the calibration.

1.2.3 The "direct" and "indirect" methods of calibration

Ragin (2008a: 85–105) proposes the so-called "direct" and "indirect" methods of calibration. Both apply only to fuzzy and not crisp sets. Unlike in the previous calibration example, these two techniques are more formalized and rely partially on statistical models. The direct method uses a logistic function to fit the raw data in-between the three qualitative anchors at 1 (full membership), 0.5 (point of indifference), and 0 (full non-membership).[5] The location of these qualitative anchors is established by the researcher using criteria external to the data at hand. The "indirect method," by contrast, requires an initial grouping of cases into set-membership scores. The researcher has to indicate which cases could be roughly classified with, say, a 0.8 membership in the set; with 0.6; 0.4; and 0.2 and so on. Using a fractional logit model, these preliminary set-membership scores are then regressed on the raw data. The predicted

[5] Because a logistic function is used, the actual anchors are at 0.95, 0.5, and 0.05.

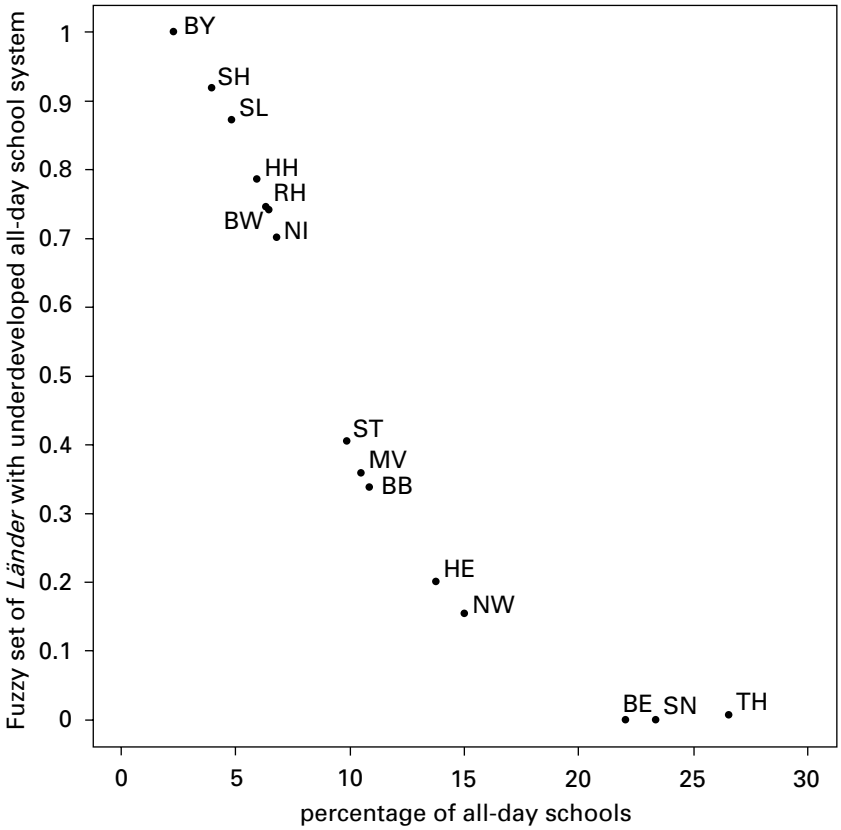

Figure 1.1 Membership in fuzzy set of *Länder* with underdeveloped all-day schools plotted against percentage of pupils enrolled in all-day schools

values of this model are then used as the fuzzy-set membership scores. Thus, if interval-scale data are at hand, the direct and indirect method of calibration can be fruitfully applied and represent progress in one of the core issues of set-theoretic methods: the creating and calibration of sets. The technical details are explained in detail by Ragin (2008a, 2008b). Conceptually, the important message is, however, that despite the complexity of the underlying statistical model, the calibration and thus set-membership scores of cases is predominantly driven by the location of the qualitative anchors. These locations, in turn, are determined by the researcher, who uses external knowledge rather than properties of the data at hand.

Freitag and Schlicht (2009) provide an example of the direct method of calibration. In their comparative work on the differences in schooling

systems in the 16 German *Länder*, they calibrate the set "*Länder* with underdeveloped all-day school system." The raw data for calibration consist of the percentage of pupils enrolled in all-day schools in a *Land*. These values vary between 2.4% (Bavaria) and 26.6% (Thuringia). Because the fuzzy set is labeled *under*development, high values in the raw data convert into low fuzzy-set membership scores and vice versa. The 0.5 qualitative anchors is located at 8.3%, which is exactly the middle of a notable gap in the raw data between 6.8% (Lower Saxony) and 9.8% (Saxony-Anhalt); the 1 anchor is located at 3% (leaving only Bavaria with full membership); and the 0 anchor at 20% (assigning 0 to Berlin, Saxony, and Thuringia).

If we plot the fuzzy-set membership scores that result from applying the direct method of calibration (for details, see Ragin 2008a: 84–94) with the qualitative anchors just described against the raw data, we clearly see the logistic nature of the transformation (Figure 1.1). We also see that despite the use of a (complex) mathematical procedure in the background, the qualitative differences between cases' set membership is clearly driven by decisions that the researcher makes based on theoretical considerations and knowledge that exist outside the raw data.

Some critiques of the direct and indirect methods of calibration have been formulated. First, partly because these calibration techniques can be performed by using the relevant software packages (fsQCA 2.5, Stata, or R), the temptation might be high to apply them in a mechanistic manner and to thus under-appreciate the importance of standards for imposing thresholds external to the data. Second, both procedures lead to very fine-grained fuzzy scales, thus suggesting a level of precision that usually goes well beyond the available empirical information and the conceptual level of differentiation that is possible. Put differently, these calibration techniques might create an impression of false precision. Another issue is the use of the logistic function for assigning set-membership scores, a choice that is not sufficiently justified. Calibration procedures using different functional forms are equally plausible and, as Thiem (2010) shows, do have a measurable impact on the set-membership scores. In other words, to some degree, the set membership of cases depends on the arbitrary choice of the functional form employed in the calibration procedure. We agree that the logistic function is arbitrary and that other functions are equally (im)plausible. Yet, as long as the 0.5 anchor remains unchanged – and its location should be determined by theoretical arguments and never by the functional form – then the effect of different functional forms on the set-membership scores remains only marginal in virtually all scenarios. The only empirical situation in which differences in

the functional form of calibration can produce differences in set membership even if the qualitative anchor remains the same is when set membership is highly skewed, i.e., when most cases are located either above or below the 0.5 qualitative anchor.

1.2.4 Does the choice of calibration strategy matter much?

Both Emmenegger and Freitag and Schlicht have (quasi-)interval-level data at hand. Yet, the first opts for a qualitative calibration while the latter apply the direct method of calibration. Does the choice of calibration strategy lead to substantively different membership scores? The general answer to this question is this: as long as the locations of the qualitative anchors are carefully chosen and thus not subject to changes in the calibration strategy (theory-guided, direct, indirect, etc.) or the functional form used in the semi-automated procedures (logistic, quadratic, linear, etc.), then the differences in set-membership scores will not be of major substantive importance.

In order to illustrate this, let us compare Emmenegger's qualitative calibration of the set of many institutional veto points with the fuzzy scores that result from applying the direct calibration method to the same data. In both procedures, we use the same qualitative anchors for full non-membership (values below 4.13) and full membership (values above 8.75). For the qualitative anchor at 0.5, it is impossible to choose the same value, though. In the qualitative calibration, Emmenegger locates it anywhere between the values of 5.05 and 6.75. The direct method of calibration, however, requires a precise location for the 0.5 cut-off. Here we encounter a major difference in calibration strategies: while in qualitative calibration no precise location for the 0.5 anchor is required, in the direct method a precise value is required. What is perhaps even more problematic is that different choices about that precise location influence the set membership scores of *all* cases, even those far above and below the point of indifference. Graphically speaking, the exact shape of the S-curve as shown in Figure 1.1 crucially depends on the location of the 0.5 anchor. Because some discretion is often exercised on the exact location of this anchor, this introduces at least some level of arbitrariness that is not found in the qualitative calibration strategy.

Table 1.3 compares Emmenegger's original fuzzy set scores with the ones obtained by such a use of the direct method of calibration. As the values in the last column indicate, the majority of cases display identical membership

Table 1.3 QUALITATIVE versus direct method of calibration for set "many institutional veto points"

		Membership in set "many institutional veto points"		
	Raw data	Qualitative calibration	Direct method of calibration	Difference
Australia	10	1	1	0
Austria	7	0.67	0.76	−0.09
Belgium	6.85	0.67	0.73	−0.06
Canada	8.75	1	1	0
Denmark	3.63	0	0	0
Finland	3.25	0	0	0
France	4.95	0.33	0.17	0.16
Germany	10	1	1	0
Ireland	3.5	0	0	0
Italy	5.05	0.33	0.19	0.14
Netherlands	6.75	0.67		−0.04
New Zealand	2.38	0	0	0
Norway	3.88	0	0	0
Portugal	2.25	0	0	0
Spain	6.75	0.67	0.71	−0.04
Sweden	4.5	0.33	0.09	0.24
Switzerland	10	1	1	0
UK	4.13	0	0	0
USA	10	1	1	0

Adapted from Emmenegger (2011)

scores. This is true for those located at the two extreme ends of the fuzzy scale. In addition, no case crosses the crucial qualitative anchor at 0.5 from one calibration strategy to the other. Only the cases with fuzzy set membership scores of 0.33 or 0.67 in Emmenegger's original calibration see a change in membership score when using the direct calibration approach. However, the difference in membership is usually too small to warrant a meaningful substantive distinction. The biggest difference occurs for Sweden, which according to the direct method of calibration is almost fully out of the set of "many institutional veto points," whereas the qualitative calibration assigns it a fuzzy value of 0.33. The reason for this is simple: Sweden's value in the raw data is just slightly higher than the UK's. This results in a marginal difference using the

direct method. However, if we just use four categories, such as Emmenegger does, then Sweden is part of the next higher category, which is described by the fuzzy value of 0.33.

When discussing the usefulness of a purely qualitative approach and of semi-automatic procedures such as the direct method, we should not forget that Emmenegger's original data (i.e., Lijphart's raw data) are not perfectly quantitative, whereas Freitag and Schlicht, for example, work with empirical quantities. Emmenegger's values are close to qualitative assessments themselves so that a complicated mathematical transformation, such as a logit function, might be a less appropriate way of reflecting the (partial) presence of a concept in given cases.

1.2.5 Assessing calibration

We have presented different ways of data calibration: starting off from theory-based, or qualitative, calibration strategies, we discussed the use of quantitative underlying scales, arriving finally at the semi-automatic direct and indirect methods. Of course, we might feel tempted to automatically resort to the latter strategies as soon as underlying quantitative measures exist. The hope of higher reliability and validity might motivate such a choice. By contrast, qualitative forms of calibration are often disregarded as being less transparent and less "scientific." However, this criticism is put in a different light if we consider that comparative research often relies on indicators generated from quantitative data of questionable quality due to issues such as low intercoder reliability; opaque aggregation strategies; or unclear content validity. For illustration, just think of the Freedom House Index as one of the most frequently used indicators of democracy used in research (see Munck and Verkuilen 2002 for a detailed critique).

Yet another reason why the critique against more theory-guided methods of calibration is somewhat misleading lies in the fact that, in practice, analytical results derived from QCA are generally robust to slight changes in the calibration method. That is to say, most results rarely vary in important ways if a case's membership value is altered slightly. We will come back to this in Chapter 11 (section 2).

In sum, it is not the principles underlying the assignment of fuzzy values which are problematic, but rather it is the temptation to disregard the central principles of calibration that causes trouble.

At-a-glance: the calibration of set membership

The **calibration** of **fuzzy-set membership scores** has to be based on theoretical knowledge and empirical evidence. Obvious facts, accepted social scientific knowledge, and the researchers' own data collection process all inform the calibration process.

Statistical distributions and parameters of underlying quantitative data can provide useful information for calibration. However, an automatic transformation of quantitative scales or the default use of statistical parameters in the calibration process is strongly discouraged, as this does not fulfill the requirement of using calibration criteria that are *external* to the data and is thus unlikely to lead to set-membership scores that reflect the meaning of the concept that is meant to be captured. A number of mathematical problems further discourage such procedures.

The **direct and indirect methods of calibration** can be applied when interval-scale data are at hand and when **fuzzy sets** (as opposed to **crisp sets**) are calibrated. These semi-automatic ways of transposing quantitative data into set-membership values are a valuable addition to the set-theoretic method toolset. Set-membership scores hinge upon the definition of the precise location of the qualitative anchors, which, in turn, are determined based on knowledge outside of the data. Thus, conceptual and theoretical knowledge remains the most important feature in these semi-automated calibration techniques.