Chapter 5
Calibrating Sets

Mello, Patrick A. (2021) *Qualitative Comparative Analysis: An Introduction to Research Design and Application,* Washington, DC: Georgetown University Press, Chapter 5.

[I]t is impossible to conduct meaningful fuzzy set-theoretic analysis without attending to issues of calibration.

Charles C. Ragin [1]

Calibration is a precondition for QCA. It is also a much-misunderstood part of the method. Before the analysis can be run, all data must be *calibrated* into crisp or fuzzy sets.[2] This procedure distinguishes QCA from other methods of empirical analysis, since there are vital differences between traditional measurement and set-theoretic calibration (Ragin 2000; 2008; Schneider and Wagemann 2012). Opening with this distinction, this chapter introduces the main calibration techniques, talks about sources of raw data, and introduces the technical routine for the direct method of calibration. Throughout, the chapter also takes into account new perspectives on calibration methodology (Goertz 2020; Ragin and Fiss 2017). The chapter closes with several calibration examples from published studies, a summary of common misconceptions, and good practices of calibration.

Measurement and Calibration

What is the difference between measurement and calibration? In the social sciences, most numerical data is based on *uncalibrated* measures. Examples include economic data, such as the gross domestic product (GDP) and the unemployment rate, or sociological data on education history, health status, or household income. These measures can be compared, aggregated, or placed in relation to an average or some other descriptive statistical indicator. Yet without additional information, we would not know whether a certain household income is "high" or "low" within a given country or region. Nor would we be able to say how a given unemployment rate compares across countries or different parts of the globe.

By contrast, calibrated measures refer to *external standards*, which means that scores can be directly interpreted once these standards are known. For example, temperatures expressed in degrees Celsius can indicate qualitatively different states, because we know that water freezes

at 0 °C and boils at 100 °C. Hence there is a *qualitative* difference between a lake at 10 °C and the same lake at -10 °C, beyond the *quantitative* 20-degree difference in temperature.

Another advantage of calibrated measures is that they allow us to distinguish between meaningful and *less relevant variation* in the uncalibrated raw data. When transforming raw data into sets, we can specify which variation to emphasize. For example, let us suppose we expect the *economic strength* of a country to be a relevant factor in explaining its performance in the implementation of certain policies. We know that there are economic differences between, say Switzerland and Denmark. But on a global scale, both of these countries would be considered economically strong, which means that we may assign them the same fuzzy score (1, or full membership in the set of *strong economies*). However, for countries with weaker economies, even small differences in GDP per capita mean a lot more in terms of economic development. With the calibration procedure, we can emphasize such differences by delineating the *relevant variation* based on our own substantive knowledge of the research area, as will be explained below.

How do we arrive at calibrated measures? Calibration and the assignment of set-theoretic scores requires (1) *plausible and consistent rules* that apply equally to all of the selected cases, (2) *content validity*, as in a close correspondence with the underlying social science concept that the target set shall reflect, and, importantly, (3) the definition of *external criteria* (Ragin 2008, 82; Schneider and Wagemann 2012, 32). In essence, calibration is about "semantic transformations", as in connecting the meaning of a concept to numerical indicators (Goertz 2020, 74).

To delineate external criteria for set-theoretic calibration, three types of knowledge can be distinguished. The first type are *undisputed facts*. These can refer to authoritative statistics or other sources of official information (such as demographic data, economic indicators, or historical timelines of events). The second area contains conceptions that are *generally accepted* or widely used in a given field of research (definitions of terms, classification systems, agreed-upon standards and benchmarks). The final area relates to *individual expertise* – which means that in order to arrive at meaningful calibration criteria, researchers should tap into their own knowledge of specific cases and concepts, and recent developments. In fact, this is a major advantage of QCA because it allows researchers to use their own substantive knowledge of a given field and to devise meaningful sets based on this knowledge.

Some may object that the role of individual knowledge inserts unwarranted subjectivity into the calibration procedure – resulting in what may be seen as arbitrary set membership scores, rather than objective, neutral measurements of the base concepts. This is a valid concern. However, it should be clear that decisions about social science concepts can never be entirely neutral, value-free assessments. Just take concepts like *poverty*, *equality*, or *terrorism* – clearly, there are different understandings of these terms. Depending on our conceptualization, the terms will be

connected to different numerical indicators. This means that any decision made during concept formation will have an impact upon the analysis and eventual results. Yet, this does *not* mean that these decisions are arbitrary. To the contrary, with every concept there will be a corridor of plausible, justifiable conceptualizations that have to resonate with the scholarly literature and prior research on a given topic (Goertz 2020, 77). This means that while calibration decisions are subjective, and typically made by an individual researcher or research group, they are still *constrained* by previous studies and accepted knowledge in a given research area.

Against this backdrop, it is crucial that calibration decisions are made *transparent*, to indicate which part of the calibration rests on external standards, which understandings informed the concept formation, and which individual decisions were made and why. For example, we should highlight where we follow the literature in coding decisions (such as using a democracy threshold of 7 on the combined Polity scale), whether several plausible calibration strategies exist for a given concept (either because different indicators could be used, or due to disagreement in the literature), or whether the data for some of the included cases is not clear-cut (as when there is contradicting or ambiguous information on a given case). The latter two situations would call for *robustness tests*, understood as complementary analyses based on different calibration strategies to compare how such changes affect the analytical results. This resonates with recent efforts to increase research transparency in qualitative and multi-method research (Büthe and Jacobs 2015), including QCA (Wagemann and Schneider 2015).[3] We will return to these points in the good practices section at the end of this chapter.

Calibration Procedures

There are three approaches to calibration. The *manual approach*, quite simply, assigns scores by hand to individual cases. Like all calibration approaches, this requires a prior definition of the target set, as well as consistent coding rules and external criteria. But beyond that, the scores are manually attributed to each case. To be sure, the manual attribution can also be done via software, where values are assigned following specific rules determined by the researcher. The second approach, termed the *direct method* of calibration, uses a software-based routine to transform numerical raw data into crisp or fuzzy sets. Crucially, this involves a prior definition of three "empirical anchors" set by the researcher (Ragin and Fiss 2017, 61). Finally, the *indirect method* of calibration requires an assignment of preliminary scores to individual cases. In the second stage, a statistical estimation technique is then used to calculate predicted fuzzy values based on the raw data and the initially assigned scores (Ragin 2008). This chapter focuses on the manual approach and the direct method of calibration as the most important approaches to calibration. [4]

When calibrating data, the first step is about conceptualizing and naming the target set. What is important here is that nouns should be turned into adjectives. So, instead of investigating

poverty or *income*, we would define a set of *poor people*. Following the *Ragin Adjective Rule*, as Gary Goertz calls it (2020, 85), ensures that the target set entails a direction and that it can be interpreted correctly. Hence, we know what it means to have a case that has full membership in the set of *poor people*, whereas some numerical income level would have to be placed into context first.

The second step in calibration is about deciding whether we want to construct a crisp set or a fuzzy set. For crisp sets, we need criteria to define what constitutes membership as opposed to non-membership in the target set. For fuzzy sets, we have to determine the *cross-over point* of 0.5, which is also known as point of maximum ambiguity. This is the point where we cannot say whether a case is rather inside or outside of a given set. This may be because of missing data, which means that we simply do not know enough about the case, or because the case shows idiosyncrasies, as in characteristics that make it difficult to classify in comparison to other cases.

Fuzzy sets can take on many forms, some of which are shown in Table 5.1. Note that the examples should not preclude other conceivable scales. A simple four-value fuzzy set introduces gradations towards both ends of the scale (*more in than out* and *more out than in*). As we move further to the right, the fuzzy scales become more differentiated, until we reach a continuous scale with fine-grained decimal scores between 0 and 1, calculated with the software-based calibration routine.

Note that there is no right or wrong with regards to the calibration scale. Crisp sets can be useful, especially for concepts with binary distinctions (a case is either inside or outside the set). Likewise, if the data allows for a more fine-grained calibration, then a continuous fuzzy set has its advantages. But there can also be substantive reasons to use other calibration scales, different from the ones suggested in Table 5.1. Note also that crisp sets and fuzzy sets both evolve around the same fundamental distinction between a case's membership or non-membership in a given set. Regardless of how nuanced a fuzzy set is, the essential question remains: what distinguishes cases that are *rather inside* (scores above 0.5) from those that are *rather outside* a given set (scores below 0.5)?

Table 5.1 Calibration Scales

Verbal label	Crisp set	Fuzzy sets			
	Binary	Four values	Five values	Eleven values	Continuous scale
Fully in	1	1	1	1	1
				0.9	$0.5 < X_i < 1$
More in than out		0.7	0.7	0.8	
				0.7	
				0.6	
Neither in nor out			0.5	0.5	0.5
				0.4	$0.5 > X_i > 0$
More out than in	0	0.3	0.3	0.3	
				0.2	
				0.1	
Fully out		0	0	0	0

To illustrate the calibration process and differences based on the choice of the approach, let us take an example using unemployment data. Suppose we wanted to construct the set *high unemployment* for the 28 European Union (EU) member states, just before the departure of the United Kingdom in January 2020. Using official reports on the EU's progress towards it stated goals for the labor market as our benchmark (EPRS 2019), [5] we may say that 10% unemployment and higher can be considered *fully in* the set high unemployment, whereas 6% shall be the *cross-over*, where a country falls right in between being inside and outside the set. Finally, 3% and lower is considered *fully out* of the set. For the direct method of calibration, these empirical anchors would let the software transform the raw data (given in percentages) into decimal fuzzy scores, ranging from 0 to 1. Note that we will look into the technical steps of the transformation procedure in the next section.

Table 5.2 shows the raw data and the calibration results for a crisp set and two fuzzy sets. The raw data is the unemployment rate among EU countries (Eurostat 2020). The right-hand columns show a crisp set calibration, and two fuzzy set calibrations. These are based on the same criteria, using a cross-over of 6% unemployment to distinguish between being inside or outside the set. For the crisp set, this results in the upper ten countries receiving scores of 1 (reflecting unemployment rates between 6.3% and 16.4%) and the lower 18 countries receiving

scores of 0 (with unemployment rates between 2.0% and 5.9%). The five-value fuzzy set uses the same cross-over, but further distinguishes between *fully in* (1) for raw data scores equal to or above 10, *almost fully in* (0.7) for those below 10 and above 6, *almost fully out* (0.3) for those below 6 and above 3, and *fully out* (0) for those with scores of 3 and less. Note that there is no country that receives a set-membership of 0.5, which would be assigned to a case with exactly 6% unemployment. Finally, on the right-hand side of Table 5.2 is the continuous fuzzy set, which is based on the direct method of calibration that is processed with the software. We can see that this calibration approach yields the most fine-grained values and that it introduces further distinctions between some of the cases that still received the same values in the five-value fuzzy set.

Does it matter which calibration approach is chosen? Is there a best approach to calibration? Table 5.2 shows that irrespective of which procedure is adopted, the resulting scores across the three types of sets remain fairly similar. This can also be seen from Figure 5.1, which visualizes the relationship between the raw data and the three different calibration scales. With crisp sets there is a qualitative jump around the cross-over. With the five-value fuzzy set there are several such jumps, while the continuous fuzzy set approximates a smooth s-shaped curve. This is so because the transformation is based on a logarithmic scale (see the discussion in the next section), which means that the default pattern is the s-shape. Certainly, the continuous fuzzy set yields the most nuanced set membership scores, but this should not be taken to imply that the other approaches are subordinate to the direct method of calibration.

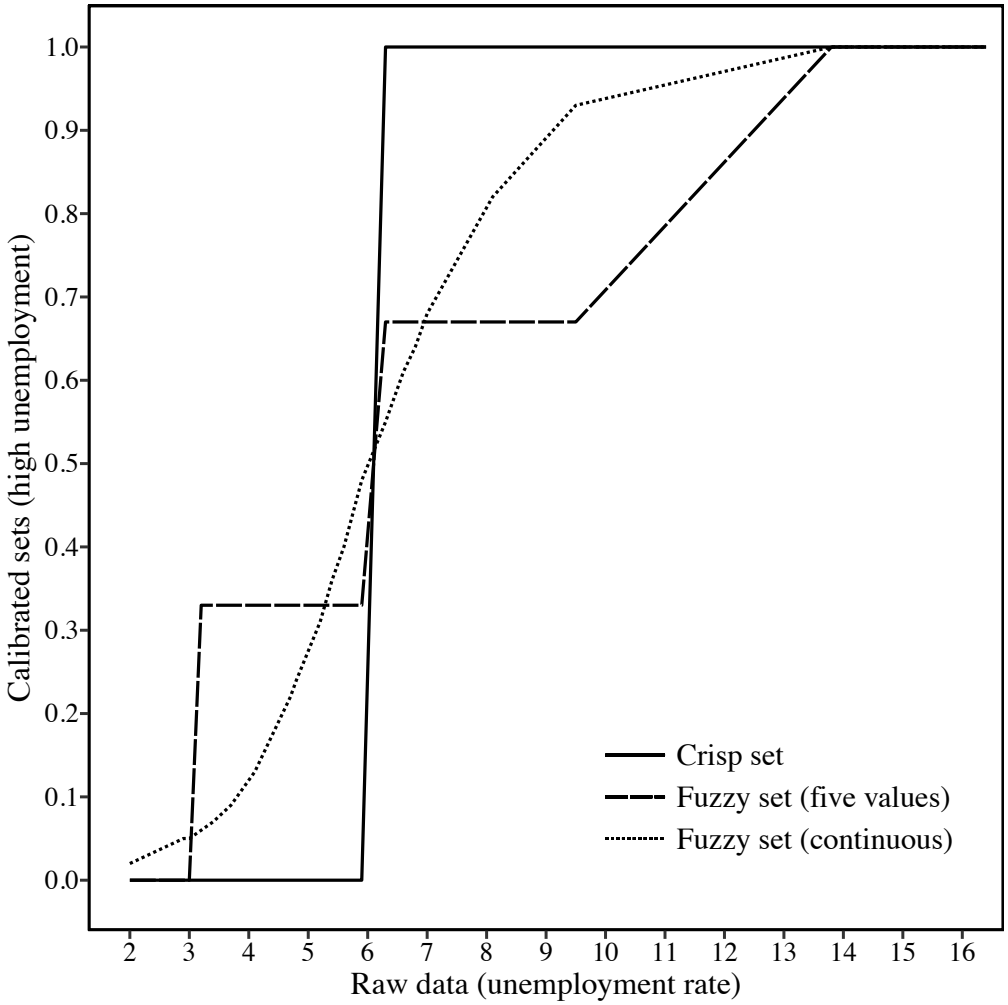
The most consequential decision in calibration is the assignment of the empirical anchors, and of these, the cross-over has the largest *qualitative* impact because it can determine whether a case is inside or outside the target set. In practice, it is often helpful to begin the calibration process by thinking about the cross-over as the distinguishing criterion. This mirrors the discussion in Chapter 2 on the selection of positive and negative cases. As a preliminary step during calibration, one can create a crisp set for an initial analysis to find out whether the selected condition provides inferential leverage that helps towards explaining the phenomenon under study. Throughout this process, it may become clear that the cross-over should be set differently, or that further distinctions and thus fuzzy sets are warranted.

Table 5.2 Raw Data and Calibrated Crisp and Fuzzy Sets

Case/ country	Raw data	Calibrated set *High Unemployment*		
	Unemployment rate (%)	Crisp set	Fuzzy set (4-values)	Fuzzy set (continuous)
Greece	16.4	1	1	1
Spain	13.8	1	1	1
Italy	9.5	1	0.7	0.93
France	8.1	1	0.7	0.82
Sweden	7.0	1	0.7	0.68
Portugal	6.8	1	0.7	0.64
Lithuania	6.6	1	0.7	0.61
Finland	6.6	1	0.7	0.61
Latvia	6.5	1	0.7	0.59
Croatia	6.3	1	0.7	0.55
Cyprus	5.9	0	0.3	0.48
Luxembourg	5.6	0	0.3	0.40
Slovakia	5.4	0	0.3	0.36
Belgium	5.2	0	0.3	0.31
Denmark	4.8	0	0.3	0.24
Ireland	4.8	0	0.3	0.24
Estonia	4.7	0	0.3	0.22
Austria	4.3	0	0.3	0.16
Bulgaria	4.1	0	0.3	0.13
Romania	3.9	0	0.3	0.11
United Kingdom	3.8	0	0.3	0.10
Slovenia	3.7	0	0.3	0.09
Hungary	3.4	0	0.3	0.07
Malta	3.4	0	0.3	0.07
Germany	3.2	0	0.3	0.06
Netherlands	3.0	0	0	0.05
Poland	2.9	0	0	0.05
Czechia	2.0	0	0	0.02

Data source: Eurostat (January 2020).

Figure 5.1 Raw Data and Different Calibration Scales



Types of Data

Which types of data can be used for calibration? And what is needed in order to construct crisp and fuzzy sets? Clearly, a major advantage of QCA is the method's openness to both *qualitative* and *quantitative* data (for an applied example, see Box 5.1 by Marij Swinkels, at the end of this section). Although emphasis is often placed on the transformation of quantitative data (e.g. Ragin 2008; Schneider and Wagemann 2012), calibration can equally draw on qualitative information gained from interviews, documents, field observations, or historical archives, just as it can be based on quantitative indicators, taken from existing or newly created datasets or surveys.[6] Importantly, different approaches can be merged in the same QCA study, as in having qualitative conditions alongside quantitative ones. There is no need to use the same calibration approach across all of the conditions in a single study. Along the same lines, crisp set and fuzzy set conditions can be combined, as in having a fuzzy-set outcome and crisp-set conditions, or vice versa, if that suits the research aim.

Finally, depending on the context, it may be helpful to draw on several different data sources to inform the calibration procedure, for instance by first using a generally accepted statistic to determine whether a case is inside or outside a set, and then using interviews and in-depth research to determine fine-grained scores above and below the 0.5 cross-over. For such an approach, one could either devise a codebook and apply the rules manually to each case, or one could use a stepwise procedure in R to conduct the calibration. That being said, to use the direct method of calibration one would usually work with interval- or ratio-level data, while all levels of measurement can be used for the qualitative calibration procedure that assigns values by hand.

Table 5.3 summarizes the levels of measurement commonly used in the social sciences, with the addition of fuzzy sets. The *nominal* level means that numbers or categories are used to classify observations or cases.[7] All cases in a category are treated as equal to each other and there is no ranking between categories. The *ordinal* level of measurement means that observations can be placed in relation to each other This is common in survey responses like "more happy" versus "less happy", or "stronger agreement" versus "moderate agreement", and so forth. Measurement at the *interval* level means that the exact distance between the observations is known and that it can be expressed in numerical terms, whereas the *ratio* level is typically considered the highest form of measurement because it entails a natural zero point (Frankfort-Nachmias and Nachmias 2008).[8] However, as Ragin (2000, 155) suggests, fuzzy sets could be considered an even higher form of measurement because these entail a natural zero and a natural maximum (see also Goertz 2020, Ch. 5):

> [I]t could be argued that fuzzy-set membership is a higher form of measurement than the conventional ratio scale – it is a ratio scale with a fixed and meaningful minimum *and maximum*. Still, the purpose of a fuzzy set is parallel to that of the nominal scale – to indicate set membership.
>
> (Ragin 2000, 155, original emphasis)

With *qualitative data*, the main challenge is to find a consistent and systematic way of linking information to numbers, while staying as close as feasible to the underlying concept. As Debora De Block and Barbara Vis (2018) note, most of the methodological literature on calibration has focused on the transformation of quantitative data. This is sometimes taken as if QCA could only work with quantitative data sets, which is not the case. Yet, the bottom line is that there are no firm rules about how to transform qualitative raw data into crisp or fuzzy sets. However, as with general aspects of calibration, it is useful to start by thinking about the cross-over as the central criterion. So, regardless of how much qualitative in-depth information you may have gathered – you should think about this part first. This is easier illustrated with examples, which are provided after the technical discussion of the calibration routine in the next section.

Table 5.3 Five Levels of Measurement

Level	Equi-valence	Greater than	Fixed interval	Natural zero	Natural maximum	Examples
Nominal	✓					Marital status, continent
Ordinal	✓	✓				Happiness, agreement
Interval	✓	✓	✓			Temperature (°C), intelligence (IQ)
Ratio	✓	✓	✓	✓		Years of education, GDP per capita
Fuzzy	✓	✓	✓	✓	✓	Educated person, developed country

The Direct Method of Calibration

What is known as the direct method of calibration is a software-based routine to transform numerical raw data into fuzzy-set scores between 0 and 1 (Ragin 2008). For the researcher, the key step is setting the three *empirical anchors* that guide this transformation (Ragin and Fiss 2017, 61). From there, the software applies a procedure that happens under the hood until we see the resulting calibrated scores. To shed light on this somewhat opaque process, I will describe the calculations that are entailed in the calibration routine.[9] Clearly, readers without an inclination for mathematics may safely skip this part and will still be able to apply the calibration procedure. Yet, users often wonder how, exactly, the calibrated scores are calculated in QCA – a question that is answered in this section.[10]

How is the raw data transformed into fuzzy values? The calibration uses a logarithmic function, which is useful because the function is symmetric, and the transformed values will stay within the boundaries of 1 and 0. In fact, the values will never exactly reach the end points because the s-shaped curve flattens out near these values, towards positive and negative infinity. Hence, by convention values of 0.95 and 0.05 are interpreted as *full membership* and *full non-membership* in a given set (Ragin 2008, 87). One should thus not be surprised to see cases with values just below 1 even though their uncalibrated value is above the threshold for full set membership. This also means that small numerical differences in fuzzy sets should not be over-interpreted as they give a false impression of preciseness. This is one of the reasons why I recommend against reporting more than two decimal points for set-membership scores in publications (another reason being poor readability).

Table 5.4 shows the verbal labels for the three empirical anchors, the corresponding degree of membership, and two additional metrics that are needed for the transformation, as I will illustrate shortly: *associated odds* and *log odds*. Associated odds are calculated with the following formula:

$$\text{associated odds} = \frac{\text{degree of membership}}{(1 - \text{degree of membership})}$$

In turn, log odds are calculated by taking the natural logarithm (ln) of the associated odds.[11] In essence, the three numerical columns in Table 5.4 are merely different representations of the same values, starting with degree of membership.

Table 5.4 Metrics for the Direct Method of Calibration

Empirical anchors	Degree of set membership	Associated odds	Log odds
Full set membership	0.95	19.000	2.944
Cross-over	0.50	1.000	0.000
Full set non-membership	0.05	0.053	-2.944

With these three metrics in place, we can now turn to the actual calibration procedure. Table 5.5 shows raw data from the Human Development Index (HDI) of the United Nations (UN 2018) for a sample of 16 countries out of a population of 189 countries for which HDI data is available. HDI scores closer to 1 indicate very high development, whereas scores closer to 0 refer to low development (these should not be confused with fuzzy scores). Conveniently, the methodology of the HDI entails a classification system according to which "high human development" is reflected in scores of 0.70 and higher, and "very high human development" relates to scores of 0.80 and above. At the low end of the scale, scores below 0.55 are considered "low human development" (UN 2018, 17). Hence, for our example, we can use these external standards as empirical anchors for the calibration of the fuzzy set *high human development*.

Accordingly, a raw data value of 0.80 and higher is taken to indicate that a country is fully in the set, 0.70 is considered the cross-over, and 0.55 marks the threshold for being fully outside the set. The three horizontal lines between the countries in Table 5.5 reflect these thresholds.

Table 5.5 Direct Method of Calibration, Human Development Example

Country	Raw data (HDI, 2017)	Deviation	Scalars	Product	Fuzzy set *High Human Development*	Empirical anchors
Switzerland	0.944	0.244	29.44	7.18	1	
Germany	0.936	0.236	29.44	6.95	1	
Lithuania	0.858	0.158	29.44	4.65	0.99	
Romania	0.811	0.111	29.44	3.27	0.96	0.80 (Fully in)
Turkey	0.791	0.091	29.44	2.68	0.94	
Brazil	0.759	0.059	29.44	1.74	0.85	
China	0.752	0.052	29.44	1.53	0.82	
Uzbekistan	0.710	0.010	29.44	0.29	0.57	0.70 (Cross-over)
El Salvador	0.674	-0.026	19.63	-0.51	0.38	
India	0.640	-0.060	19.63	-1.18	0.24	
Kenya	0.590	-0.110	19.63	-2.16	0.10	
Pakistan	0.562	-0.138	19.63	-2.71	0.06	0.55 (Fully out)
Togo	0.503	-0.197	19.63	-3.87	0.02	
Ethiopia	0.463	-0.237	19.63	-4.65	0.01	
South Sudan	0.388	-0.312	19.63	-6.12	0	
Niger	0.354	-0.346	19.63	-6.79	0	

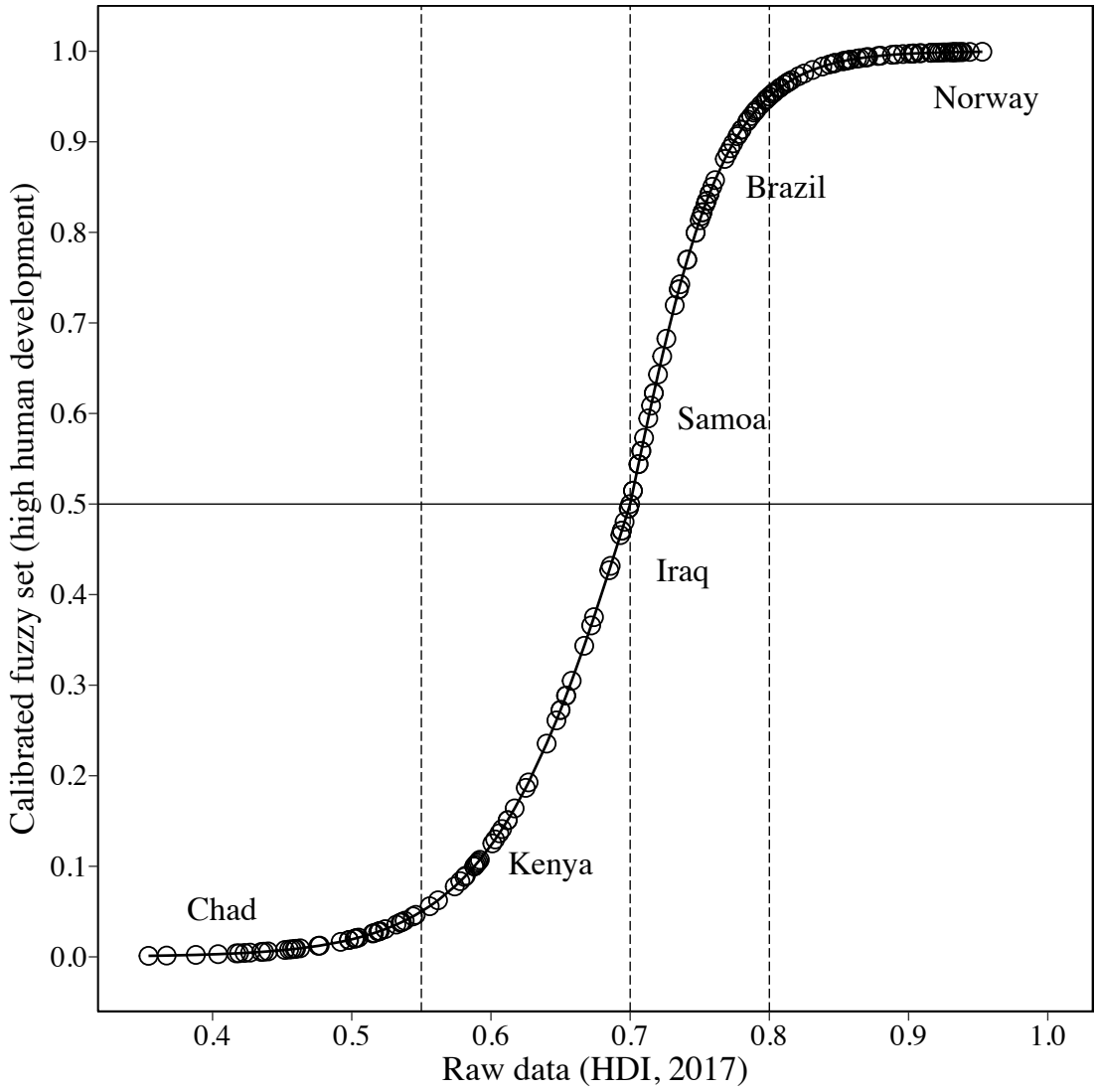
Once the empirical anchors are in place, we can calculate each cases' *deviation* (or numerical distance) from the cross-over of 0.70. This results in positive scores for the cases above the cross-over and negative scores for those below. In the next step, we take the *log odds* for full set membership (+2.944, see Table 5.4) and full set non-membership (-2.944) and divide them by the deviation of the threshold for fully in (0.10) and fully out (-0.15), which gives us *scalars* of 29.44 and 19.63, respectively. For each case, these scalars are then multiplied by the cases' deviation, which yields a *product*, as shown in the respective column of Table 5.5. Finally, the product is transformed into scores between 0 and 1, by taking the *exponent* of the product and dividing it by itself plus one, as in the following example. The result is rounded to two digits, which yields a fuzzy set membership of 1 in the case of Switzerland. In the same way, we calculate the fuzzy values for all of the countries, as shown in the second-to-last column in Table 5.5:

$$\text{Switzerland} = \frac{\exp(7.18)}{(1 + \exp(7.18))} = \frac{1312.91}{1313.91} = 0.9992 = 1$$

The effects of the calibration procedure and the choice of the empirical anchors can be visualized with an XY plot that displays the raw data against the calibrated data. Figure 5.2 shows the s-shaped curve that is typical of a logarithmic function. The x-axis displays the complete raw data from the Human Development Index (2018), which ranges from 0.354 (Niger) to 0.953 (Norway), covering 189 countries, as opposed to the selected 16 countries show in Table 5.5. The y-axis displays the calibrated fuzzy values. The figure also includes three dashed vertical lines for the empirical anchors and a solid horizontal line for the 0.5 cut-off that separates cases that are rather inside the fuzzy set (above the horizontal line) from those that are rather outside the fuzzy set (below the horizontal line). Finally, the XY plot shows the location of selected cases across the range of raw data shown in Table 5.5 (from low to high: Chad, Kenya, Iraq, Samoa, Brazil, and Norway). As can be seen in the figure, the calibrated values rise steeply around the cut-off and between the anchors for full exclusion and full inclusion. Beyond those, the line flattens out on both ends, closing in on 0 and 1, respectively. This means that any empirical variation beyond the thresholds is de-emphasized. Countries receive similar fuzzy values beyond the empirical anchor chosen for full exclusion (0.55) and full inclusion (0.8), whereas small increases in the raw data in the center (around the cross-over, between these thresholds) lead to substantially higher fuzzy values as visualized in the steep ascent of the curve. Note that, for presentational purposes, the x-axis is truncated because there are no countries with less than 0.354 HDI (Niger's score).

What fuzzy set calibration does, and what can be gleaned from the XY plot, is that it essentially *stretches* the raw data that is located in the center, above and below the cross-over, and it *flattens* the raw data that is located around the edges, below full exclusion and above full inclusion in the set. This procedure emphasizes small differences in the area that is deemed most relevant (above and below the cross-over) and it de-emphasizes variation that is held to be less relevant for the phenomenon under study. Plotting the raw data against the calibrated fuzzy set makes this immediately visible, which is one reason why such plots should be included in online appendices or supplementary material for empirical applications of QCA (see guidelines in the final section of this chapter).

Figure 5.2 XY Plot of Raw Data and Calibrated Fuzzy Set



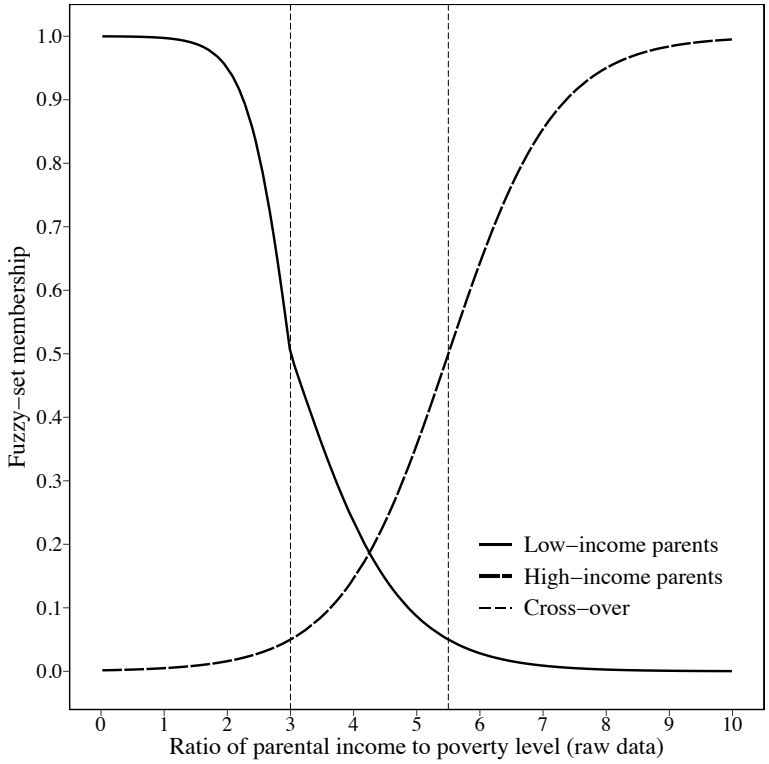
Calibration: Applied Examples

This section provides several calibration examples from published studies (Cacciatore et al. 2015; Fagerholm 2014; Kuehn et al. 2017; Mello 2020; Ragin and Fiss 2017; Schmitt 2018). These were selected to cover a range of different calibration scenarios – including procedures of qualitative calibration, the transformation of quantitative raw data, the combination of several conditions into a macro-condition, the weighting of indicators, and asymmetric concepts. The examples presented are illustrative in nature – without going into the context and details of the respective studies, nor discussing their QCA results. Here it must suffice to introduce the concept of the respective condition or outcome, the underlying data, calibration thresholds, and the results of the calibration process.

Example 1: "Parental Income" (Ragin and Fiss 2017)

In their book-length study *Intersectional Inequality*, Charles Ragin and Peer Fiss (2017) explore life chances and social inequality based on various configurations of race, gender, educational achievement, and family background. Two of their fuzzy-set conditions are based on parental income in relation to the household-adjusted poverty level: *low-income parents* and *high-income parents*. What makes this interesting for our purposes is that these are examples of *asymmetric concepts*. As Ragin and Fiss note, there is a difference between individuals with high-income parents and those with *not* low-income parents (2017, 69). The former group of people is certainly smaller than the latter and hence it makes sense to analyze these as separate sets. This is visualized in Figure 5.3, where we can see the area between the two vertical cross-over lines. Individuals in this area are rather outside of *both* sets (fuzzy scores below 0.5). Ragin and Fiss (2017, 65-70) base their calibration criteria on prior studies, official government thresholds, and national surveys, noting that there are "value judgments" involved because "there is no consensus within the scientific community on exactly where the poverty line should be drawn" (2017, 67). However. their own criteria are substantiated by recommendations from the National Research Council. Against this backdrop, the authors use the direct method of calibration on raw data that reflects multiples of the income to poverty ratio to construct the fuzzy sets *low-income parents* (fully out: 5.5, cross-over: 3, fully in: 2) and *high-income parents* (fully out: 3, cross-over: 5.5, fully in: 8).

Figure 5.3 Low-Income and High-Income Parents (Ragin and Fiss 2017)



Example 2: "Path Dependence" (Cacciatore et al. 2015)

In their study of Europeanization and the implementation of national reform programs among EU member states, Federica Cacciatore, Alessandro Natalini, and Claudius Wagemann (2015) include a *path dependence* condition ("ACCESS") that builds on theoretical expectations derived from historical institutionalism, namely that a country's familiarity with EU institutions and practices "will determine the domestic implementation success" of new European legislation (2015, 1191). The authors use a straightforward metric to operationalize the path dependence logic by measuring the number of years that passed since a member state joined the European institutions. The fuzzy set is calibrated using the direct method of calibration with the empirical anchors 5 years (fully out), 20 years (cross-over), and 40 years (fully in), as summarized in Table 5.6. The table shows only a selection of cases to illustrate the calibration procedure, but it should be noted that the study includes three observations per country (in the years 2011-2013). As can be seen from the table, the empirical anchors effectively distinguish the founding and early members of the EU from those that joined in later accession rounds of the organization. The authors discuss their calibration strategies in a supplementary document where they argue that their choice for the calibration thresholds for the path dependence condition ACCESS "seems to be an appropriate time span for a country to fully achieve and assimilate the EU's political culture and mechanisms" (Cacciatore et al. 2015, online appendix, p. 5). Arguably, with numerical indicators such as these, there will always be some leeway and others might opt for a higher or lower threshold. What the authors highlight, however, is that during the three observed years, none of the cases passed the cross-over from being considered less accustomed to more accustomed with EU institutions (ibid.).

Table 5.6 Condition "Path Dependence" (Cacciatore et al. 2015)

Country (*selection*)	Year of EU/EC accession	Years in EU/EC	Fuzzy set ACCESS	Empirical anchors
Belgium	1952	51	0.99	
Germany	1952	51	0.99	
Italy	1952	51	0.99	
Denmark	1973	40	0.95	
United Kingdom	1973	40	0.95	Fully in ≥ 40 years
Greece	1981	32	0.86	
Portugal	1986	27	0.74	
Spain	1986	26	0.71	
Sweden	1995	18	0.40	Cross-over = 20 years
Finland	1995	18	0.40	
Hungary	2004	9	0.10	
Slovenia	2004	8	0.08	Fully out ≤ 5 years

Data source: Cacciatore et al. (2015, online appendix)

Example 3: "Civilian Control of the Military" (Kuehn et al. 2017)

In their study of civil-military relations in 28 new democracies, David Kuehn, Aurel Croissant, Jil Kamerling, Hans Lueders, and André Strecker (2017) operationalize the fuzzy-set outcome *civilian control of the military* by taking weighted averages across five fuzzy-set dimensions: elite recruitment (ER), public policy (PP), internal security (IS), national defense (ND), and military organization (MO). Due to its centrality for the functioning of democratic institutions, elite recruitment receives a five-fold weight, whereas public policy and internal security are weighted two-fold and the other conditions receive single weights (Kuehn et al. 2017, 432). The authors justify this weighting scheme with a body of theoretical work on this topic. This yields the following formula for the calculation of the outcome:

$$\text{Civilian Control} = \frac{\text{ER} \times 5 + \text{PP} \times 2 + \text{IS} \times 2 + \text{ND} + \text{MO}}{11}$$

The study by Kuehn et al. (2017) illustrates how a complex condition can be modelled with fuzzy sets and how individual indicators can be assigned different weights. Here, the authors use the *average* across the included indicators, but one could have also constructed a fuzzy set from combinations of conditions that are joined by Boolean operators, as introduced in Chapter 2 of this book. For instance, Goertz (2020) provides some illustrative examples of how multiple-level theories and concepts can be modelled in this way. Table 5.7 shows the calibrated data for the outcome and its constituent conditions for five of the 28 cases of the study. The use of higher-order concepts, such as *civilian control* employed by Kuehn et al. (2017), has the advantage that a broad array of indicators can be taken into account to inform the calibration of a single condition. One downside of this is that the interpretation of the eventual scores is complicated and it may be necessary to re-examine a case's scores in the constituent conditions to make sense of empirical patterns (for example, Nepal 2 and Taiwan 1 receive similar scores in the outcome, but they differ substantively on *internal security* and *military organization*). Another concern is the weighting scheme, as one could plausibly justify different weights, but these have a large impact upon the outcome scores. Hence, when considering an *index condition,* it is vital to justify the constituent conditions and their weighting (especially when these are not weighted equally), and to consider alternative analyses as robustness tests.

Table 5.7 Outcome "Civilian Control" (Kuehn et al. 2017)

Country (*selection*)	Civilian control (Outcome)	Elite Recruitment	Public policy	Internal security	National defence	Military organization
Brazil 1, 1985-87	0.13	0.2	0.2	0	0	0
Brazil 2, 1988-98	0.65	0.85	0.85	0.4	0	0.4
Nepal 1, 1999-01	0.74	1	0.85	0.6	0	0.2
Nepal 2, 2006-10	0.86	1	1	0.55	0.4	1
Taiwan 1, 1992-01	0.89	1	1	1	0.4	0.4
Taiwan 2, 2002-10	1	1	1	1	1	1

Data source: Kuehn et al. (2017, online appendix).

Example 4: "Utility of Junior Partners in Coalition Warfare" (Schmitt 2018)

The book-length study *Allies that Count* by Olivier Schmitt (2018) investigates the role of junior partners in coalition warfare, such as the multinational military operations in Iraq and Afghanistan. The book primarily comprises intensive case studies based on archives, interviews, participatory observation, and secondary sources. These are complemented by a crisp-set QCA for a systematic exploration of the qualities that make a junior partner useful in coalition warfare. Schmitt defines the outcome *utility* as a junior partner's "capacity to positively assist in the achievement of the desired result" (2018, 202). The calibration involves in-depth qualitative assessments for each of the observed coalition members. As Schmitt explains, "The question asked in assessing the utility was, would the campaign planning and conduct have been substantially different in the absence of this ally?" (2018, 203). This question is addressed for each of the observed cases and conflicts. For example, as for the United Kingdom and France as coalition partners in Afghanistan, Schmitt argues:

In Afghanistan, the nonparticipation of France or the United Kingdom would have forced the United States to devote key resources to strategically important areas of operations such as the Helmand and Kapisa Provinces. While the performance of the British and French troops can certainly be criticized, whether another ally would have done better is unclear. In any case, the absence of these two states would have negatively impacted the legitimacy of the operation and the strategic planning by draining US resources. The French and British participation were coded as [1]. (Schmitt 2018, 202).

The study by Schmitt (2018) is an example of a *qualitative* calibration procedure which boils down to value judgments, the results of which cannot easily be transformed into numbers.[12] Against this backdrop, Schmitt chooses the crisp-set variant of QCA, which emphasizes qualitative differences rather than differences in degree.

Example 5: "Ecological Change" (Fagerholm 2014)

In his study of ecological change within social democratic parties, Andreas Fagerholm (2014) examines the ecological orientation of parties' election programs. His study makes use of data from the Comparative Manifesto Project, CMP (Klingemann et al. 2006; Volkens et al. 2013), which contains several variables that are related to ecological preferences. Specifically, Fagerholm focuses on mentions of "anti-growth economy: positive" (per416) and "environmental protection: positive" (per501) and subtracts mentions of "productivity: positive" (per410), using the following formula to calculate an indicator for social democratic parties' ecologism:

$$\text{Ecologism} = [(\text{per416} + \text{per501}) - \text{per410}] \geq 5.0$$

Given this formula, ecologism is seen as present when "ecological issues outnumber non-ecological issues by a difference of at least 5 per cent", as Fagerholm explains (2014, 5). Table 5.8 shows selected raw data for 7 of Fagerholm's 19 social democratic parties (abbreviated for purposes of illustration). The study analyzed party programs over three time periods, between 1970 and 1999. The ecologism indicator was used as a basis to calibrate the crisp-set outcome *ecological change*. This was coded positively when ecologism scored above 5 per cent during at least one of the observed time periods. From Table 5.8, we can see that this is the case for social democratic parties in the Netherlands, Switzerland, Luxembourg, and Germany but not in Spain, the United Kingdom, and France. The study by Fagerholm (2014) is another example where several indicators inform the calibration of a single condition. Here, the same indicator (ecologism) is taken at three different points in time to determine what is conceptualized as *ecological change*. This way, the study also takes into account *temporality* – an aspect that is typically neglected by the static comparisons of QCA. Why did the study use a crisp-set outcome? Fagerholm argues that emphasis was placed on "case-based knowledge, since *differences in kind* are deemed to be more important than fine-grained *differences in degree*" (Fagerholm 2014, 11, emphasis added). This is a plausible argument but given that the study already has fine-grained CMP data at hand, it might have been worthwhile to calibrate a fuzzy-set outcome from this.

Table 5.8 Outcome "Ecological Change" (Fagerholm 2014)

Social democratic party (*selection*)	Ecological change (Outcome)	Ecologism		
		1970–79	1980–89	1990–99
PvdA (Netherlands)	1	6.40	5.72	8.12
SPS (Switzerland)	1	5.24	15.74	7.31
SPÖ (Luxembourg)	1	2.97	8.10	7.10
SPD (Germany)	1	–0.84	3.30	15.27
PSOE (Spain)	0	–0.75	–0.82	3.40
LP (United Kingdom)	0	0.50	1.91	3.10
PS (France)	0	0.60	–3.03	0.95

Data source: Fagerholm (2014).

Example 6: "Fatalities" (Mello 2020)

In my own study of coalition defection during the Iraq War (Mello 2020), I included a condition that takes into account civilian and military deaths that resulted from terrorist attacks – as a potential influence on democratic leaders' decision-making. The study drew on two databases: the Iraq Casualties Project for military casualties and the Rand Database of Worldwide Terrorism Incidents, for civilian casualties. I calibrated this data starting with a *qualitative* criterion – distinguishing "primarily between leaders who experienced casualties and those who did not" (Mello 2020, 15). This means that government leaders without any casualties received fuzzy values of 0, while those with casualties received fuzzy values between 0.51 and 1.0.

To determine the exact value for the condition, the direct method of calibration was used on a quantitative measure of the ratio between casualties and the number of deployed troops, using the thresholds 0.005 (fully out), 0.006 (cross-over), and 0.50 (fully in). The cross-over effectively means that countries with *at least one fatality* receive scores above 0.51 and above (and are thus considered rather inside the set). Table 5.9 summarizes the raw data and calibrated fuzzy set for 12 out of the 51 leaders included in the study. For example, we can see that El Salvador received a fuzzy set value of 1 whereas Japan only received a value of 0.98. This is because the fatalities per deployment ratio is higher for El Salvador under President Flores (2.5%) than for Japan under Prime Minister Koizumi (0.67%). Figure 5.4 visualizes the calibration.

Hence, the fuzzy set *fatalities* is another example of an *asymmetric condition* (see Example 1). Moreover, the condition is conceptualized in a way that allows for scores between 0.51 and 1, but not for scores larger than 0 and less than 0.51 (since either there are civilian or military

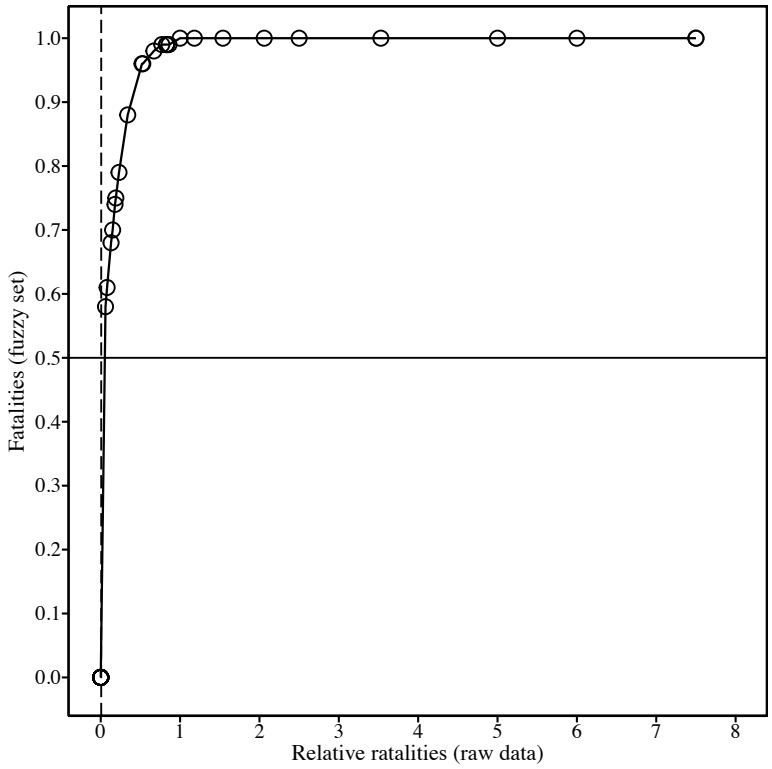
deaths or there are none). Alternatively, a crisp-set condition could have been used, but this could not have taken into account gradations based on fatalities per deployment.

Table 5.9 Condition "Fatalities" (Mello 2020)

Country (*selection*)	Government leader (*selection*)	Raw data			Fuzzy set
		Fatalities per deployment	Fatalities (nominal)	Deployed number of troops	Fatalities
Albania	F. Nano	0.00%	0	80	0.00
Australia	J.W. Howard	0.19%	2	1,048	0.75
Denmark	A.F. Rasmussen	0.83%	4	480	0.99
El Salvador	F. Flores	2.50%	1	40	1.00
Estonia	A. Ansip	0.00%	0	40	0.00
Hungary	F. Gyurcsány	0.00%	0	300	0.00
Japan	J. Koizumi	0.67%	4	600	0.98
Netherlands	J.P. Balkenende	0.23%	3	1,288	0.79
Poland	L. Miller	0.06%	1	1,667	0.58
Portugal	J.M.D. Barroso	0.00%	0	120	0.00
Spain	J.L.R. Zapatero	0.08%	1	1,300	0.61
United Kingdom	G. Brown	0.34%	16	4,770	0.88

Data source: Mello (2020).

Figure 5.4 Raw Data and Calibrated Fuzzy Set Fatalities



Common Misconceptions about Calibration

At the outset of this chapter, I mentioned that calibration is a much-misunderstood part of QCA. What do I mean by that? First, sometimes you encounter a view that holds calibration to be an essentially "arbitrary" process of "making up" crisp and fuzzy values. Clearly, that is not the case. As this chapter has shown, there are *well-defined standards* for calibration. Additionally, many published studies go to great lengths to clarify what data they used and how this was transformed into set-membership scores. But it is understandable that the process may seem opaque when one is not used to working with calibrated measures. And, surely, there are also QCA articles where it is difficult to decipher the calibration process, much less to replicate the results.[13] In sum, the best way to address concerns about calibration is with conceptual clarity, data transparency, and plausible calibration criteria (see also the good practices of calibration in the next section).

Second, calibration is occasionally approached as a mechanical exercise of transforming numerical raw data into crisp and fuzzy scores. As mentioned in this chapter, it makes no sense to simply take descriptive statistics like the average or median and use these as cut-offs for the calibration. Nor does it make sense to take maximum and minimum scores in the empirical data and use them as upper and lower bounds. Such approaches miss the fundamental advantage of QCA, namely that *meaningful variation* can be separated from irrelevant variation (Ragin 2000, 161). This can only happen on a substantive basis and therefore requires conceptual work.

Third, there are also misconceptions about crisp and fuzzy sets. Regardless of the choice, users should justify *why* they opted for crisp or fuzzy sets. Some argue that crisp sets are the more conservative, or safer choice when there is not enough data or when it is difficult, and thus potentially arbitrary, where to introduce gradations in set membership. This may be true for some research endeavors, and there are certainly concepts that have a greater affinity towards crisp sets. But often it appears that users apply crisp sets simply because it is more convenient and less research intensive to work with binary values. Moreover, a widely-held view is that it is easier to identify consistent set-theoretic relationships with crisp sets than with fuzzy sets (Ragin 2009, 114; Schneider and Wagemann 2012, 69). Hence, a weak set-theoretic relationship can look stronger because crisp sets were used. This seems to resonate with the experience of many applied QCA researchers, who find that fuzzy sets provide for more demanding tests of consistency and coverage (we will look into the calculation of these metrics in Chapter 6). However, in a recent contribution Ingo Rohlfing (2020) reaches the conclusion that the relationship between crisp sets, fuzzy sets, and consistency scores is "ambiguous" and, as such, can go in both directions, depending on the structure of the empirical data. Hence it may be unwarranted to claim that crisp sets *generally* lead towards higher consistency scores.

Finally, it is sometimes implied that the choice of the calibration approach can greatly affect the results. This leaves new users wondering about the "correct" choice when calibrating their data. However, as shown in the example using unemployment data (Table 5.2), the differences in the results are rarely substantial. What matters are the positions of the empirical anchors, especially the cross-over, but it is of secondary concern whether the direct method of calibration was used or whether the values were assigned manually, based on previously defined calibration rules. We will return to these issues in later chapters.

Good Practices of Calibration

The discussion of common misunderstandings lends itself to the following list of good practices of calibration. This goes in line with efforts to formulate coherent methodological standards for QCA.[14]

- Studies should clearly document the conditions that were used throughout the analysis. Ideally, the publication includes at least a summary table with this kind of information. (This is particularly relevant for studies that use various models of conditions, see Chapter 2).
- Data sources must be made transparent. The raw and calibrated data should be made available, either on the journal website or in an openly accessible data repository. For qualitative data, this can be more demanding, and even raise ethical issues (as in protecting sources and sharing sensitive information). In that case, researchers should use discretion to decide what information can be shared and what needs to be aggregated or anonymized before publication (on transparency, see the contributions in Büthe and Jacobs 2015).
- For analyses conducted within the R software environment, the R script should be made available on an openly accessible data repository.
- The method of calibration and calibration thresholds should be reported. For conditions calibrated with the direct method, histograms and plots for raw and calibrated data should be provided in supplementary documents.
- Calibration thresholds must be verbally justified, and their impact should be discussed (why cases were considered below/above the threshold, what their inclusion/exclusion means for the results). For critical cases, there should be an alternative analysis based on different calibration thresholds.
- Set labels should be as concise and unambiguous as possible (acronyms do not work particularly well if there is no legend provided in the table).
- Set names should adhere to the adjective rule to indicate the directionality of a given set (for instance, *generous* welfare state, *supportive* public, *high* unemployment, and so forth).

Not

[1] Ragin (2008, 8).

[2] The chapter focuses on crisp and fuzzy sets. On multi-value conditions, see Chapter 8.

[3] In 2015, the American Political Science Association (APSA) initiated a process that led to wide-ranging "Qualitative Transparency Deliberations" across four thematic clusters (the discussion forums can still be accessed at: https://www.qualtd.net). This also led to a report with specific transparency recommendations for QCA (Schneider et al. 2019).

[4] There are few empirical applications of the indirect method of calibration, which is probably due to its more complicated, two-fold procedure (when compared to the direct method). For a technical discussion, see Ragin (2008, 94-97) and Duşa (2019, 92-94).

[5] It should be highlighted that whether the three percentage scores are plausible empirical anchors would have to be justified within the research context of a given study.

[6] On challenges specific to the transformation of qualitative data, see De Block and Vis (2018). Qualitative calibration examples are also provided in Kahwati and Kane (2020, 76-86).

[7] Goertz (2020, 138) argues that it is "fundamentally misleading" to perceive of the nominal level of measurement as a scale, because a nominal condition is really is "two (or more) concepts".

[8] This distinguishes temperature measured in degrees Kelvin from temperature measured in degrees Celsius. While Celsius has a zero point, it is arbitrary, whereas Kelvin has a meaningful zero.

[9] This section is based on Ragin's (2008, 85-105) detailed account of the calibration process. See also Duşa (2019, 74-92), who describes how the procedure happens inside the QCA package for R and who also introduces an alternative to the logarithmic function on which the standard calibration procedure is based.

[10] This is complicated by the fact that the numbers for the metrics given in Ragin (2008, 88) are rounded up, which is why a manual re-calculation with his data does not produce identical results.

[11] These metrics are geared towards the thresholds of 0.95, 0.50, and 0.05. The three-digit values for associated odds and log odds are approximations that are rounded for convenience. Because the software works with the exact numbers, a manual re-calculation may yield slightly different results (Duşa 2019, 87).

[12] On qualitative calibration, see also De Block and Vis (2018).

[13] For my QCA courses at the University of Erfurt, one of the tasks is for students to replicate published studies. This exercise showed that few studies could be replicated with matching results, as many articles missed information on raw data, calibration thresholds, and analytical decisions. Clearly, this issue applies more broadly and is not limited to set-theoretic methods.

[14] On standards of good practice, see Schneider and Wagemann (2010). For a survey of empirical applications and their alignment with formulated standards, see Mello (2013). For guidelines, see also Oana et al. (2021).

References

Büthe, Tim, and Alan M. Jacobs. 2015. "Introduction to the Symposium: Transparency in Qualitative and Multi-Method Research." *Qualitative & Multi-Method Research* 13 (1): 2-8.

Cacciatore, Federica, Alessandro Natalini, and Claudius Wagemann. 2015. "Clustered Europeanization and National Reform Programmes: A Qualitative Comparative Analysis." *Journal of European Public Policy* 22 (8): 1186-211.

De Block, Debora, and Barbara Vis. 2018. "Addressing the Challenges Related to Transforming Qualitative Into Quantitative Data in Qualitative Comparative Analysis." *Journal of Mixed Methods Research* (8 May): https://doi.org/10.1177%2F1558689818770061.

Duşa, Adrian. 2019. *QCA with R. A Comprehensive Resource*. Cham: Springer.

European Parliamentary Research Service (2019) "The Fight against Unemployment." Briefing, PE 630.274.

Eurostat. 2020. "Euro area unemployment at 7.4%, EU at 6.6%." Eurostat News Release: 30 April 2020.

Fagerholm, Andreas. 2014. "Social Democratic Parties and the Rise of Ecologism: A Comparative Analysis of Western Europe." *Comparative European Politics* 14 (5): 1-25.

Frankfort-Nachmias, Chava, and David Nachmias. 2008. *Research Methods in the Social Sciences*. 7th ed. New York, NY: Worth Publishers.

Goertz, Gary. 2020. *Social Science Concepts and Measurement*. Princeton: Princeton University Press.

Kahwati, Leila C., and Heather L. Kane. 2020. *Qualitative Comparative Analysis in Mixed Methods Research and Evaluation*. Los Angeles Sage.

Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge, and Michael McDonald, ed. 2006. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990-2003.* Oxford: Oxford University Press.

Kuehn, David, Aurel Croissant, Jil Kamerling, Hans Lueders, and André Strecker. 2017. "Conditions of Civilian Control in New Democracies: An Empirical Analysis of 28 'Third Wave' Democracies." *European Political Science Review* 9 (3): 425-47.

Mello, Patrick A. 2013. "From Prospect to Practice: A Critical Review of Applications in Fuzzy-Set Qualitative Comparative Analysis." 8th Pan-European Conference on International Relations, Warsaw, 18-21 September.

———. 2020. "Paths towards Coalition Defection: Democracies and Withdrawal from the Iraq War." *European Journal of International Security* 5 (1): 45-76.

Oana, Ioana-Elena, Carsten Q. Schneider, and Eva Thomann. 2021. *Qualitative Comparative Analysis Using R: A Beginner's Guide*. New York: Cambridge University Press.

Ragin, Charles C. 2000. *Fuzzy-Set Social Science*. Chicago, IL: University of Chicago Press.

———. 2008. *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago, IL: University of Chicago Press.

———. 2009. "Qualitative Comparative Analysis Using Fuzzy Sets (fsQCA)." In *Configurational Comparative Methods*, edited by Rihoux, Benoît and Charles C. Ragin. Thousand Oaks: Sage, 87–121.

Ragin, Charles C., and Peer C. Fiss. 2017. *Intersectional Inequality: Race, Class, Test Scores, and Poverty*. Chicago: University of Chicago Press.

Rohlfing, Ingo. 2020. "The Choice between Crisp and Fuzzy Sets in Qualitative Comparative Analysis and the Ambiguous Consequences for Finding Consistent Set Relations." *Field Methods* 32 (1): 75-88.

Schmitt, Olivier. 2018. *Allies that Count: Junior Partners in Coalition Warfare*. Washington, DC: Georgetown University Press.

Schneider, Carsten Q., Barbara Vis, and Kendra Koivu. 2019. "Set-Analytic Approaches, Especially Qualitative Comparative Analysis (QCA)." APSA Section for Qualitative and Multi-Method Research, Final Report of the Qualitative Transparency Deliberations Working Group III.4.

Schneider, Carsten Q., and Claudius Wagemann. 2010. "Standards of Good Practice in Qualitative Comparative Analysis (QCA) and Fuzzy-Sets." *Comparative Sociology* 9 (3): 397–418.

———. 2012. *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. New York, NY: Cambridge University Press.

United Nations. 2018. *Human Development Indices and Indicators*. New York, United Nations Development Programme.

Volkens, Andrea, Judith Bara, Ian Budge, and Michael D. McDonald, ed. 2013. *Mapping Policy Preferences from Texts III: Statistical Solutions for Manifesto Analysts.* Oxford: Oxford University Press.

Wagemann, Claudius, and Carsten Q. Schneider. 2015. "Transparency Standards in Qualitative Comparative Analysis." *Qualitative & Multi-Method Research* 13 (1): 38-42.