Calibrating Sets

Patrick A. Mello
Willy Brandt School of Public Policy
University of Erfurt

Version: July 2019

Citation: Mello, Patrick A. (2019) "Calibrating Sets." In: *Qualitative Comparative Analysis: Research Design and Application*. Washington, D.C.: Georgetown University Press (under contract).

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

Charles Ragin [1]

Calibration is a precondition for QCA. It is also a much-misunderstood part of the method.[2] Before the analysis can be run, all data must be *calibrated* into crisp or fuzzy sets. This procedure distinguishes QCA from other methods of empirical analysis since there are vital differences between traditional measurement and set-theoretic calibration. Starting with a discussion of this distinction, this chapter introduces the two main calibration techniques, talks about sources of raw data, and introduces the technical routine for the direct method of calibration. The chapter closes with calibration examples from published studies, a summary of common misconceptions, and good practices of calibration.

---

[1] Ragin (2008b, 8).

[2] The final section of this chapter addresses common misconceptions about calibration.

Measurement and Calibration

What is the difference between measurement and calibration? Most numerical data in the social sciences is based on *uncalibrated* measures. Examples include economic data, such as gross domestic product (GDP) and 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 an unemployment rate of 6% compares across countries or different parts of the globe.

By contrast, calibrated measures refer to *known standards*, which means that scores can be directly interpreted. 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 5 °C and the same lake at -5 °C, beyond the *quantitative* 10-degree difference in temperature.

Another advantage of calibrated measures is that they allow us to distinguish between meaningful and less relevant variation. With fuzzy sets, we can specify which variation to emphasize. For example, let us say we are interested in studying economic development in global comparison. We know that there are differences between, say Switzerland and Denmark. But on a global scale, both countries would be considered economically strong, which means that we may assign the same fuzzy score to these countries (1.0, or full membership in the set of "strong economies"). However, for countries with weaker economies, small improvements in GDP per capita mean a lot more in terms of economic development. With the calibration procedure, we can emphasize these differences, as will be explained below.

How do we arrive at calibrated measures? Calibration and the assignment of set-theoretic scores requires plausible and consistent rules, content validity, as in a close correspondence with the underlying concept, and the definition of "external criteria" (Ragin 2008a; 2008b, 82; Schneider and Wagemann 2012, 32).

Broadly speaking, three sources of knowledge can be distinguished, each of which can help to delineate external criteria for crisp and fuzzy sets. The first area contains *undisputed facts*. These can refer to authoritative statistics or other sources of official information (e.g. demographic data, economic indicators, historical timelines of events, and so forth). The second area contains conceptions that are *generally*

accepted or widely used in a given field (e.g. definitions of terms, classification systems, agreed-upon standards, or thresholds for certain concepts).

The final area relates to *individual expertise* – which means in order to arrive at meaningful calibration criteria, researchers should also utilize 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 substantive knowledge of a given field and to devise meaningful sets based on this knowledge.

Some may caution that such practice introduces unwarranted subjectivity into measurement. Hence it is crucial to be transparent about calibration decisions and to indicate which part of the calibration rests on external standards, and where individual decisions came in (which may of course be justified). To illustrate what this means in practice, the final section of this chapter discusses several examples and also highlights the different sources of knowledge that went into calibration decisions.

Calibration Procedures

There are two main approaches to calibration.[3] In the first approach, scores are assigned "by hand" to individual cases. Like all calibration approaches, this requires a prior definition of consistent coding rules and external criteria. But beyond that, the scores are manually attributed to each individual case. Of course, the "manual attribution" can also be done via software, where values are assigned following specific rules determined by the researcher (e.g. values determined in a codebook). The second approach, termed the "direct" method of calibration, uses a software-based routine to assign scores to cases based on numerical raw data. This requires a prior definition of three qualitive break-points or "anchors" set by the researcher.

When calibrating data, apart from conceptualizing the target set, the first question is whether we want to construct a crisp set or a fuzzy set. For crisp sets, we need to define what constitutes membership as opposed to non-membership in the target

[3] There is also a third approach: the "indirect" method of calibration, which is a variant of the direct method of calibration. However, there are few substantive advantages of applying its more complicated procedure. This may be the reason why, to my knowledge, there has not yet been an empirical application that uses the indirect method of calibration. The technical procedure is discussed in Ragin (2008b, 94-97). See also Duşa (2018, 92-94).

set. For fuzzy sets, we further 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 that make it difficult to classify when compared to other cases.

Fuzzy sets can take on many forms, some of which are shown in Table 1.[4] Starting with a rudimentary three-value fuzzy set that still bears semblance to a crisp set, over to a five-value fuzzy set that allows for distinctions near the ends of the scale ("more in than out" and "more out than in"), up to a continuous fuzzy set that takes on many different values between 0 and 1, calculated with the software-based routine.

Note that there is no "right" or "wrong" with regards to the coding scheme. Crisp sets can be useful, especially for dichotomous concepts. Likewise, if the data allows a fine-grained calibration, then a continuous fuzzy set has its advantages. But there can also be good reasons to devise other coding schemes, based on theoretical and empirical grounds. 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" (scores below 0.5) a given set.

[4] Note that Table 1 is not meant to be exhaustive: many other fuzzy-scales are conceivable (with four or six values, for example).

Table 1: Crisp and Fuzzy Calibration

Crisp Set	*Three-Value Fuzzy Set*	*Five-Value Fuzzy Set*	*Continuous Fuzzy Set*
1 = fully in	1 = fully in	1 = fully in	1 = fully in
		0.7 = more in than out	more in than out: $0.5 < X_i < 1$
	0.5 = neither in nor out	0.5 = neither in nor out	0.5 = neither in nor out
		0.3 = more out than in	more out than in: $0.5 > X_i > 0$
0 = fully out	0 = fully out	0 = fully out	0 = fully out

To illustrate the calibration process and differences based on the choice of set and calibration approach, let us take an example using unemployment data. Suppose we wanted to construct the fuzzy set "high unemployment" for Eurozone countries. We may say that 10% unemployment and higher would be considered "fully in" the set of high unemployment, whereas 6% may be the cross-over, and 3% and lower would be considered "fully out" of the set. Based on these three anchors, the software would transform the raw data (given in percentages) into decimal fuzzy scores, ranging from 0 to 1. (A technical description of the mathematical transformation that is entailed in the direct method of calibration is given in the next section).

Table 2 shows the raw data and the calibrated crisp set and two fuzzy sets. The raw data is the unemployment rate in eight Eurozone countries (Economist 2019, 80). The right-hand columns show a crisp set calibration, and two fuzzy set calibrations. These are based on the same criteria or thresholds – using a cross-over of 6% unemployment between being inside or outside the set "high unemployment".

For the crisp set, this results in the upper four countries receiving scores of 1 (reflecting their high unemployment rates between 18.5% and 8.8%) and the lower four countries receiving scores of 0 (with unemployment rates between 5.7% and 3.2%). The five-value fuzzy set uses the same cross-over but further distinguishes between "fully in" (1.0), "almost fully in" (0.67), "fully out" (0), and "almost fully out" (0.33).

Note that Belgium receives a set-membership of 0.50, because when rounded to single-digits, the country's score is exactly on the cross-over point. With a five-value

fuzzy set, this is arguably a more plausible coding decision than assigning a value of 0.33, which would have been the alternative with a two-digit value. Finally, the continuous fuzzy set "high unemployment" is based on the direct method of calibration that is processed with the software. For the calculation, the qualitative anchors were 3% for fully out, 6% as the cross-over, and 10% for fully in.

How much does the choice of the calibration approach matter for the resulting values? Table 2 shows that irrespective of which procedure is chosen, the resulting set values across the three types of sets are fairly similar. But surely, the continuous fuzzy set yields more fine-grained scores than the five-value fuzzy set and the crisp set. The most striking *qualitative* difference is the assignment of 0.50 set membership to Belgium in the five-value fuzzy-set, whereas Belgium is considered rather outside the set in the continuous fuzzy set and fully outside in the crisp set. Note that a score of 0.5 has implications for the truth table analysis, as we will see in later chapters, so this decision should not be taken lightly (it may also be the starting point for a robustness test, using alternative calibrations for an ambiguous case like Belgium).

Table 2: Raw Data and Calibrated Crisp and Fuzzy Sets

Case/ Country	*Raw Data*	*Calibrated Set "High Unemployment"*		
	Unemployment Rate	*Crisp Set*	*Fuzzy Set (5-values)*	*Fuzzy Set (continuous)*
Greece	18.5%	1	1.00	1.00
Spain	14.0%	1	1.00	1.00
Italy	10.2%	1	1.00	0.96
France	8.8%	1	0.67	0.89
Belgium	5.7%	0	0.50	0.43
Austria	4.8%	0	0.33	0.24
Netherlands	4.2%	0	0.33	0.15
Germany	3.2%	0	0.00	0.06

Data: Economist (2019: 80)

Data Sources

What kind of data can be used for calibration? 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. Although emphasis is often placed on the transformation of quantitative data (cf. Ragin 2008b; Schneider and Wagemann 2012), calibration can equally draw on qualitative information gained from interviews, field observations, or historical archives, just as it can be based on quantitative indicators, taken from existing or newly created datasets or surveys.[5] Importantly, different approaches can be combined 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. We can also use a crisp-set outcome and fuzzy-set conditions, or vice versa, if that suits our research aim.

Finally, we can integrate and combine data sources, for instance by using a generally accepted statistic to determine whether a case is inside or outside a given set and then to use interviews and in-depth research to determine fine-grained scores above and below 0.5. That being said, to use the direct method of calibration one needs interval- or ratio-level data, while all levels of measurement can be used to inform the qualitative calibration procedure that assigns values "by hand".

Table 3 summarizes the four levels of measurement commonly used in the social sciences.[6] The nominal level means that numbers or categories are used to classify observations or cases. All cases in a category are treated as equal to each other but there is no ranking between categories. The ordinal level of measurement means that observations can be placed in relation to each other (as in "more happy" versus "less happy", or "stronger agreement" versus "moderate agreement"). Measurement at the interval level means that the exact distance between observations is known and

[5] 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).

[6] The table draws on Frankfort-Nachmias and Nachmias (2008, 148).

be expressed in numerical terms, whereas ratio-level data further entails a natural or absolute zero point.[7]

Table 3: Four Levels of Measurement

Level	*Equivalence*	*Greater than*	*Fixed Interval*	*Natural Zero*	*Examples*
Nominal	✓				Marital Status, Continent
Ordinal	✓	✓			Happiness, Agreement
Interval	✓	✓	✓		Temperature (°C), Intelligence (IQ)
Ratio	✓	✓	✓	✓	Temperature (°K), Height

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 values between 0 and 1. For the researcher, the key step is setting the three qualitative anchors that guide this transformation. From there, the software applies a procedure, which happens under the "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.[8]

[7] 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.

[8] This section draws on Ragin's (2008b, 85-105) detailed account of the calibration process. See also Duşa (2018, 74-92), who describes how the procedure happens inside the "QCA" package for R and who also introduces an alternative to the logarithmic function.

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 2008b, 87).

Table 4 shows the verbal labels for the three qualitative anchors, the corresponding degree of membership, and two additional metrics that are needed for the transformation: 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.[9] In essence, the three numerical columns in Table 4 are merely different representations of the same values, starting with degree of membership.

Table 4: Metrics for the Direct Method of Calibration

Verbal Label	*Degree of Set Membership*	*Associated Odds*	*Log Odds*
Threshold for full set membership	0.95	19.000	2.944
Cross-over point	0.50	1.000	0.000
Threshold for full set non-membership	0.05	0.053	-2.944

[9] 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, the results may occasionally differ (Duşa 2018, 87). Note that Ragin (2008b, 88) rounds up the values in Table 5.1 of his book, which is why a manual re-calculation with his data does not yield the same results.

With these three metrics in place, we can now turn to the actual calibration procedure. Table 5 shows raw data from the Human Development Index of the United Nations (HDI, UN 2018) for a sample of 16 countries. 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). 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 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). For our example, we can use these *external standards* as qualitative break-points for our calibration of the fuzzy set "high human development".

Accordingly, a raw data value of 0.80 and higher is taken to indicate being 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 reflect these thresholds.

Once the qualitative 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.

Table 5: Direct Method of Calibration, Human Development Example

Country	*Raw Data (HDI, 2017)*	*Deviation*	*Scalars*	*Product*	*Calibrated Fuzzy Values*	*Qualitative Anchors*
Switzerland	0.944	0.24	29.44	7.18	1.00	
Germany	0.936	0.24	29.44	6.95	1.00	
Lithuania	0.858	0.16	29.44	4.65	0.99	
Romania	0.811	0.11	29.44	3.27	0.96	0.80 (fully in)
Turkey	0.791	0.09	29.44	2.68	0.94	
Brazil	0.759	0.06	29.44	1.74	0.85	
China	0.752	0.05	29.44	1.53	0.82	
Uzbekistan	0.710	0.01	29.44	0.29	0.57	0.70 (cross-over)
El Salvador	0.674	-0.03	19.63	-0.51	0.38	
India	0.640	-0.06	19.63	-1.18	0.24	
Kenya	0.590	-0.11	19.63	-2.16	0.10	
Pakistan	0.562	-0.14	19.63	-2.71	0.06	0.55 (fully out)
Togo	0.503	-0.20	19.63	-3.87	0.02	
Ethiopia	0.463	-0.24	19.63	-4.65	0.01	
South Sudan	0.388	-0.31	19.63	-6.12	0.00	
Niger	0.354	-0.35	19.63	-6.79	0.00	

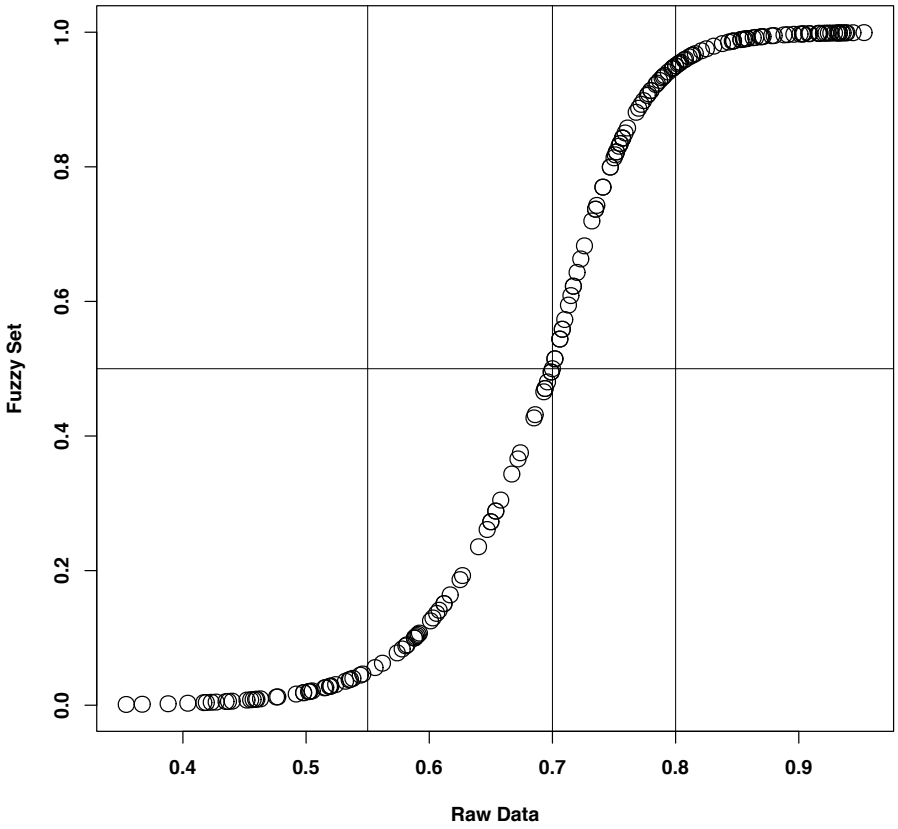
In the next step, we take the log odds for full set membership (+2.944, see Table 4) and full set non-membership (-2.944) and divide them by the deviation of the threshold for "fully in" (0.10) and "fully out" (-0.15), which gives us scalars of 29.44 and 19.63, respectively. For each case, these scalars are multiplied by the cases' deviation, which yields a product, as shown in the respective column of Table 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 1.00 in the case of Switzerland:

$$\text{Switzerland} = \frac{\exp(7.18)}{(1 + \exp(7.18))} = \frac{1312.91}{1313.91} = 0.9992 = 1.00$$

In the same way, we can calculate the fuzzy values for all of the countries, as shown in the second-to-last column in Table 5. The effects of the calibration procedure and the choice of the qualitative anchors can be visualized with an XY plot that displays the raw data against the calibrated data. Figure 1 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. The y-axis displays the calibrated fuzzy values. The figure also includes vertical lines for the three qualitative anchors and a horizontal line for the 0.5 cut-off that separates cases that are "rather inside" the set (above the horizontal line) from those that are "rather outside" (below).

As can be seen from 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 1.0, respectively. This means that any empirical variation beyond the thresholds is de-emphasized. Countries receive similar fuzzy values beyond the thresholds 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.

Figure 1: XY Plot of Raw Data and Calibrated Data



Calibration: Applied Examples

Example 1: "Path Dependence" (Cacciatore et al. 2015)

In their study of Europeanization and the implementation of national reform programs, Cacciatore et al. (2015) include a *path dependence* condition ("ACCESS") that builds on expectations derived from historical institutionalism, namely that familiarity with EU institutions and practices "will determine the domestic implementation success" of new European legislation. The authors use a straight-forward 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 qualitative anchors 40 years (fully in), 20 years (cross-over), and 5 years (fully out), as summarized in Table 6 (abbreviated table). As the table shows, this effectively distinguishes the founding and early members of the EU from those that joined in later periods. Yet, the 20-year-threshold as the cross-over means that countries like Sweden and Finland remain just

below the threshold and stay in the same qualitative group as those countries that joined during the large EU accession rounds of the 2000s.

Table 6: Condition "Path Dependence" (Cacciatore et al. 2015)

Country	*Accession Year*	*Years in EU/EC*	*Fuzzy Set ACCESS*	*Qualitative 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	Cross-over = 20 years
Portugal	1986	27	0.74	Fully out ≤ 5 years
Spain	1986	26	0.71	
Sweden	1995	18	0.40	
Finland	1995	18	0.40	
Hungary	2004	9	0.10	
Slovenia	2004	8	0.08	

Source: Cacciatore et al. (2017, online appendix), abbreviated table.

Example 2: "Civilian Control of the Military" (Kuehn et al. 2017)

In their study of civil-military relations in 28 new democracies, Kuehn et al. (2017) operationalize the 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:

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

The study by Kuehn et al. (2017) illustrates how a complex condition can be modelled with fuzzy sets and how indicators can be emphasized differently. Here, the authors use the average across the included indicators, but one could have also constructed a fuzzy set from the combination of various other conditions that are joined by Boolean operators.[10] Table 7 shows raw data and calibrated data for five of the 28 cases from Kuehn et al. (2017):

Table 7: Outcome "Civilian Control" and Indicators (Kuehn et al. 2017)

Country	*Civilian Control (Outcome)*	*Elite Recruit.*	*Public Policy*	*Internal Security*	*National Defence*	*Military Organization*
Brazil 1, 1985-87	0.13	0.20	0.20	0.00	0.00	0.00
Brazil 2, 1988-98	0.65	0.85	0.85	0.40	0.00	0.40
Nepal 1, 1999-01	0.74	1.00	0.85	0.60	0.00	0.20
Nepal 2, 2006-10	0.86	1.00	1.00	0.55	0.40	1.00
Taiwan 1, 1992-01	0.89	1.00	1.00	1.00	0.40	0.40
Taiwan 2, 2002-10	1.00	1.00	1.00	1.00	1.00	1.00

Source: Kuehn et al. (2017, online appendix), abbreviated table.

Example 3: "Public Support" (Mello 2014)

In his study of democracies' war involvement in Iraq, Mello (2014) includes a fuzzy-set condition based on public opinion surveys across 30 democracies. The study used data from various international polls that asked whether respondents would support their country's participation in a military intervention against Iraq, two months before the eventual invasion by a US-led coalition in March 2003. Mello used the direct method of calibration and the three qualitative anchors 75% public support (fully in), 45% (cross-over), and 15% (fully out). Why a cross-over of 45% and not 50%? Mello argues that "Since roughly 10 per cent of the respondents gave no answer or were undecided, the point of maximum ambiguity is at 45 per cent public support – a point at which an equal share of respondents were opposed to military involvement" (Mello 2014, 167).

[10] Goertz (2006) provides illustrative examples of how multiple-level theories and concepts can be modelled.

Table 8 shows raw and calibrated data for 10 countries from Mello (2014, 168). What is striking about the public opinion data is that no country reached the threshold for full inclusion and not even the cross-over. In effect, there was no *qualitative* variation as all countries remained more or less outside the set "public support". Normally, there would be little inferential value in keeping such a condition (which is by definition necessary for the outcome and the non-outcome, as the author acknowledges (Mello 2014, 179), but the author justifies the inclusion of public opinion because the factor plays a prominent role in prevalent accounts of the Iraq War (Mello 2014, 148-55).

Table 8: "Public Support" (Mello 2014)

Country	*Raw Data (Public Opinion)*	*Calibrated Set "Public Support"*
Slovakia	41.0%	0.40
United States	33.0%	0.23
Czech Republic	30.0%	0.18
United Kingdom	27.0%	0.14
Poland	21.0%	0.08
Netherlands	13.0%	0.04
Australia	12.0%	0.04
Spain	12.0%	0.04
Germany	10.0%	0.03
Finland	7.0%	0.02

Source: Mello (2014: 168), abbreviated table.

Example 4: "Ecological Orientation" (Fagerholm 2014)

In his study of ecological change within social democratic parties, Fagerholm (2014) examines the ecological orientation of parties' election programs. His study makes use of data from the Comparative Manifesto Project (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, for his outcome social democratic parties' "ecologism":

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

This means that ecologism is seen as present when "ecological issues outnumber non-ecological issues by a difference of at least 5 per cent"(Fagerholm 2014, 5). Table 9 shows the selected raw data for 7 of Fagerholm's 19 social democratic parties. The study analyzed party programs over three time periods, between 1970 and 1999. The crisp-set outcome "ecological change" was coded positively, if ecologism scored above 5 per cent during at least one time period. 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). This is a plausible argument, but given that the study already has the fine-grained CMP data at hand, it may have been worthwhile to calibrate a fuzzy-set outcome from this.

Table 9: "Ecologism" (Fagerholm 2014)

Social Democratic Party	*Ecologism*			*Ecological Change*
	1970–79	*1980–89*	*1990–99*	
PvdA (NED)	6.40	5.72	8.12	1
SPS (SUI)	5.24	15.74	7.31	1
SPÖ (LUX)	2.97	8.10	7.10	1
SPD (GER)	–0.84	3.30	15.27	1
PSOE (ESP)	–0.75	–0.82	3.40	0
LP (UK)	0.50	1.91	3.10	0
PS (FRA)	0.60	–3.03	0.95	0

Source: Fagerholm (2014: 8; 12), abbreviated table.

Example 5: "Fatalities" (Mello 2019)

In his study of coalition defection during the Iraq War, Mello (2019) includes 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 draws on two databases: the Iraq Casualties Project for military casualties and the Rand Database of Worldwide Terrorism Incidents, for civilian casualties. Mello calibrates this data starting with a *qualitative* criterion – distinguishing "primarily between leaders who experienced casualties and those who did not" (Mello 2019, 15). This means that 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, 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 one at least one fatality receive scores above 0.51 and above. Table 10 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 1.0 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%).

Table 10: "Fatalities" Mello (2019)

Country	Leader	Raw Data			Fuzzy Set
		Fatalities per Deployment	Fatalities (Nominal)	Deployed 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

Source : Mello (2019: 27-32), abbreviated table.

Common Misconceptions about Calibration

At the start 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 and this chapter has shown that 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, admittedly, there are also

some QCA articles where it is almost impossible to decipher the calibration process, much less to replicate the results.[11] In sum, the best way to address concerns about calibration is with conceptual clarity, data transparency, and clear calibration criteria (see also the "good practices" in the final section of this chapter).

Second, from the user perspective, 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 mean 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 an approach misses 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 to allow for gradations in set membership. This may be true in some cases. But often it appears that users apply crisp sets simply because it is more convenient to work with binary values. However, as Schneider and Wagemann (2012, 277) point out, crisp sets tend to yield higher consistency and coverage measures. Hence, a weak set-theoretic relationship can look stronger because crisp sets were used. This is one reason why fuzzy sets should be the preferred choice whenever the data allows this, besides the advantage of more fine-grained measures.

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 2), the differences in the results are rarely substantial. What matters are the positions of the qualitative anchors, but it is of secondary concern whether the direct of method of calibration was used or whether the values were assigned manually,

[11] For my QCA courses at the University of Erfurt, one of the tasks is for students to try and replicate published studies. It was sobering to see that very few studies could be replicated with matching results. In many cases this was because of missing information on calibration thresholds or analytical decisions or a lack of access to the raw data used in a study.

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 a list of good practices of calibration, as summarized in Box 1. This goes in line with efforts to formulate coherent methodological standards for QCA.[12]

Box 1: Good Practices of Calibration

- Studies should clearly indicate the total number of conditions that were used throughout the analysis. (This is particularly relevant for studies that use various "models" of conditions).
- Data sources should be made transparent. The raw and calibrated data should be included in the publication itself, in an online appendix on the journal's website, or in an openly accessible data repository.
- 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 there should also be histograms and plots for raw and calibrated data (in supplementary documents).
- Calibration thresholds should be verbally justified, and their impact should be discussed (e.g. why certain cases were considered below or above the threshold, what their inclusion or exclusion means for the results). For critical cases, there should be an alternative analysis based on different calibration thresholds (reported in a supplementary document).
- Set labels should be as concise and unambiguous as possible (long acronyms do not work particularly well).

[12] 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).

- Set names should indicate the directionality of a given set (e.g. *generous* welfare state, *supportive* public, *high* unemployment, and so forth).
- Robustness tests with alternative data sources and/or other plausible calibration thresholds should be conducted to validate analytical results.

Summary

This chapter started on the distinction between traditional measurement and set-theoretic calibration, before presenting the main calibration approaches, the use of raw data, and the technical routine for the direct method calibration. The chapter closed with several calibration examples from published studies, a summary of common misconceptions and good practices of calibration.

References

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. 2018. *QCA with R. A Comprehensive Resource*. Cham: Springer.

Economist. 2019. "Economic & Financial Indicators." *The Economist* 18 May.

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*. New York, NY: Worth Publishers.

Goertz, Gary. 2006. *Social Science Concepts: A User's Guide*. Princeton, NJ: 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.

Mello, Patrick A. 2014. *Democratic Participation in Armed Conflict: Military Involvement in Kosovo, Afghanistan, and Iraq*. Basingstoke: Palgrave Macmillan.

Mello, Patrick A. 2019. "Paths towards Coalition Defection: Democracies and Withdrawal from the Iraq War." *European Journal of International Security* (14 June): https://doi.org/10.1017/eis.2019.10.

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

Ragin, Charles C. 2008a. "Measurement versus Calibration: A Set-Theoretic Approach." In *The Oxford Handbook of Political Methodology*, edited by Box-Steffensmeier, Janet M., Henry E. Brady and David Collier. Oxford: Oxford University Press, 174-98.

Ragin, Charles C. 2008b. *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago, IL: University of Chicago Press.

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.

Schneider, Carsten Q., and Claudius Wagemann. 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.