STRATEGIES FOR SOCIAL INQUIRY

Set-Theoretic Methods for the Social Sciences

A Guide to Qualitative Comparative Analysis

CARSTEN Q. SCHNEIDER
CLAUDIUS WAGEMANN

CAMBRIDGE

1.2 The calibration of set membership

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

1.2.1 Principles of calibration

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

1.2.2 The use of quantitative scales for calibration

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

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

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

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

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

Table 1.2 Calibration of condition "many institutional veto points"

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

Source: Emmenegger (2011)

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

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

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

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

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

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

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

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

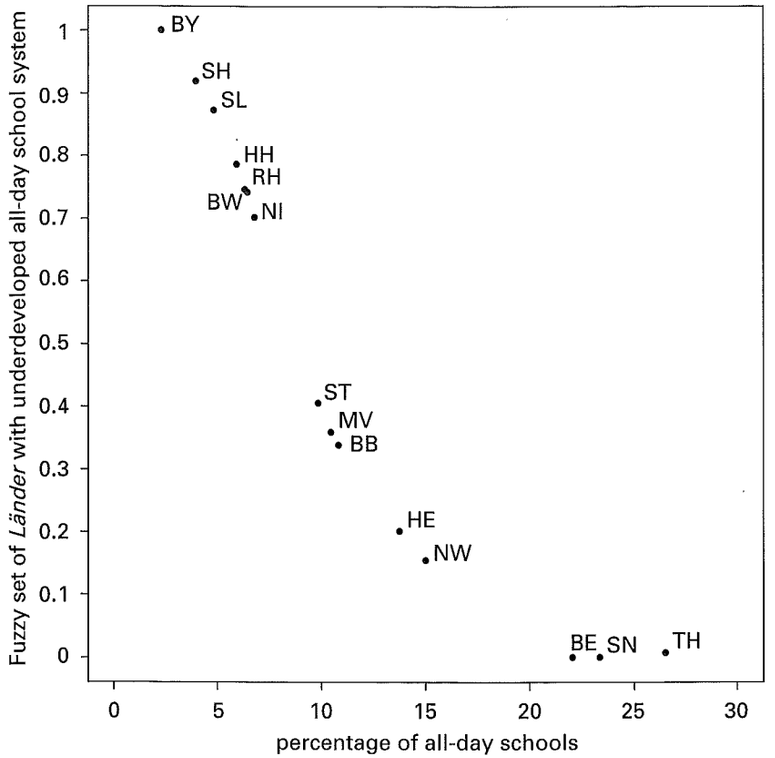

Figure 1.1 Membership in fuzzy set of *Länder* with underdeveloped all-day schools plotted against percentage of pupils enrolled in all-day schools

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

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

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

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

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

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

1.2.4 Does the choice of calibration strategy matter much?

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

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

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

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

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

Adapted from Emmenegger (2011)

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

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

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

1.2.5 Assessing calibration

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

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

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

At-a-glance: the calibration of set membership

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

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

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