A HANDOUT FOR GETTING CALIBRATION RIGHT

ROEL RUTTEN, TILBURG UNIVERSITY

JULY 2025

Being a set-analytical method, QCA requires calibration to determine cases' (degree of) membership in sets. **Calibration distinguishes between qualitatively different cases**; between cases in and out of the set (crisp sets) or between cases above and below the crossover point (fuzzy sets). Put differently, calibration places a threshold between cases that have enough of the characteristics of a concept to be recognized (qualified) as an instance of that concept and cases that do not. QCA is a threshold method. It investigates whether being above or below the threshold for X makes a difference for being above or below the threshold for Y. It makes calibration the most decisive analytical decision in a QCA study. This handout provides calibration guidelines for both new and experienced users of QCA. It explains what calibration is, what it is not and how to do it correctly.

WHAT IS CALIBRATION

Calibration attaches meaning to measurements. Most fundamentally, calibration answers **two questions**: (1) **What does it mean to be a case of X** and (2) **How do we know a case of X when we see one**. The first question is about definition, about meaning. The second question is about identifying a plausible measure (empirical value) that captures this meaning. It is important to keep the two questions apart so as not to conflate meaning and measurement. For example, to calibrate the set of poor households, researchers first define what it means to be a poor household. A poor household may be defined as one that cannot make ends meet at the end of the month. Second, researchers set an empirical value that identifies this thresholds. A researcher can now follow **two approaches**. Either they can investigate every one of their cases and determine whether or not each case (household) makes ends meet. In this approach researchers **calibrate cases**. Or researchers **can calibrate a source variable**, e.g., income, and set a generic crossover point (a particular income) for all cases. Calibrating cases is the most accurate but also the most labour intensive approach to calibration and may not be feasible in large-N studies. Calibrating source variables distinguishes between relevant and irrelevant variation; however, researchers may not know whether all cases below the crossover point (below the income threshold) fail to make ends meet (are actually instances of the concept). In both approaches, researchers give a definition first and look at measurements second.

WHAT CALIBRATION IS NOT

Calibration is **not data conversion**. Calibration is not in the first place about data, instead it is about meaning. Calibration may look like data conversion, however, obtaining set-membership values is always about setting meaningful thresholds between qualitatively different cases. Cases above the threshold for X are cases that we recognize as instances (examples) of X because they have enough of the characteristics of X to qualify as instances of X. Cases below the threshold for X do not have enough of the characteristics of X. Consequently, **means, averages and percentiles are not a basis for calibration**. Cases just above and below the average are all typical examples of average cases. The

average does not distinguish between qualitatively different cases. Neither do means and percentiles. They are arbitrary thresholds because there is no meaningful difference between cases on either side of a mean or percentile. Using arbitrary thresholds for calibration produces spurious results that cannot be meaningfully interpreted. A case is not an instance of X because it is above the 50th percentile of a distribution. It is an instance of X because it has enough of the characteristics of X to qualify as such. Only when researchers define their sets as, e.g., "above average crime rate area" or "below average student" can population distributions be used for calibration.

SETS (CONDITIONS) VERSUS VARIABLES

Variables and conditions (sets) are both analytical constructs. They are, however, very different things. **Variables describe the distribution of a characteristic** (e.g., income) across a population of cases (households). Instead, **conditions qualify (identify) cases** as having a particular characteristic (e.g., being a poor household). Having the condition (in degree) identifies a case (a household) as a member (in degree) of the corresponding set (the set of poor households). Sets are not necessarily representative of a given population and they are unconcerned with how a characteristic is distributed in a population. Researchers must always define sets (conditions) using adjectives. Income **(noun) is a variable**. Low-income household **(adjective) is a set** (condition). Better yet, poor household, to emphasize the meaning of the set (being poor) rather than the source variable (income).

CRISP AND FUZZY SETS

Set-membership values are truth values, not empirical values. Having a membership of 1 in the crisp set of poor households means that the statement: this is a poor household, is true (because it cannot make ends meet). Having a membership of 0 in the set of poor households means that the statement is false. The set-membership values 1 and 0 are unconcerned with how far above or below the threshold for being poor a household is. **1 and 0 only signal that it is true or false** that the household is poor (cannot make ends meet). **Fuzzy sets also allow statements to be true in degree**. A household may be, e.g., fully poor, mostly poor, moderately poor or marginally poor. Having said that, also in fuzzy set QCA (fsQCA), the crossover point, i.e., the threshold between more true than false (>0.5) and more false than true (<0.5), is what matters mostly – because also fsQCA is a threshold method. That is, crisp sets are not analytically inferior to fuzzy sets.

DEGREES OF MEMBERSHIP

Degrees of membership must be a **function of the fine-grainedness of** (the definition of) **a concept**, not of the fine-grainedness of the source variable (i.e., the data). It is not meaningful to define many degrees of set-membership for concepts that are coarse. Some concepts are meaningfully defined dichotomously and require only two set-membership values (1 and 0), other concepts may be defined using multiple degrees. However, **only** when using a **ratio-scale source variable** is it meaningful to **calibrate continuous fuzzy sets**. And only then is it meaningful to use the direct method of calibration

(see below). When calibrating from anything other than ratio-scale source variables, and when calibrating cases, researchers identify degrees of membership with linguistic hedges.

Each linguistic hedge (degree of membership) is a threshold in its own right and requires a definition (meaning) and a corresponding measurement (empirical value). That is, researchers answer the above two questions for each linguistic hedge. For example: What does it mean to be a mostly poor household? And how do we know a mostly poor household when we see one? The number of linguistic hedges that can meaningfully be identified is a function of (the definition of) the concept, not of the fine-grainedness of the empirical data. **Set-membership values are assigned to linguistic hedges fairly arbitrarily** (unless calibrated from ratio-scale source variables). Calibrations A and B in Table 1 are equally plausible ways of assigning set-membership values to linguistic hedges. This makes set-membership values an ordinal scale – unless calibrated from ratio-scale source variables using the direct method of calibration. The crossover point (0.5) is set between intensifying hedges (fully, mostly, largely, considerably) and diluting hedges (moderately, somewhat, marginally, not).

Table 1: Degrees of membership

linguistic hedge	calibration A	calibration B
fully a case of X	1	1
mostly a case of X	0.85	0.90
largely a case of X	0.70	0.75
considerably a case of X	0.55	0.60
moderately a case of X	0.45	0.40
somewhat a case of X	0.30	0.25
marginally a case of X	0.15	0.10
not a case of X	0	0

THE MEANING OF THE CROSSOVER POINT (0.5)

The crossover point is the threshold that distinguishes between cases that qualify (are recognized) as instances of X (>0.5) and cases that do not (<0.5). Cases with a set-membership >0 but <0.5 are members of the fuzzy set of X but are not instances (examples, cases) of X. To qualify as an instance of X (>0.5), cases must have enough of the characteristics of X to be recognized as such. Cases below the crossover point may still have some of the characteristics of X but not enough to qualify as examples (instances) of X. This makes it logically impossible for cases to have a membership of 0.5. Cases cannot have too many of the characteristics of X to be not-a case of X but too few characteristics to be a case of X. Moreover, the truth table minimization ignores 0.5-cases (because they are not difference makers), which makes calibrating 0.5-cases pretty much the set-analytical equivalent of throwing away data. Calibrating from source variables, the crossover point should be set at a value that does not exist in the data. For example, if the source variable 'income' is in Pounds but not pennies, the crossover point for poor households may be set at, e.g., £1,500.50 or £1,499.50 but not at £1,500.

CALIBRATING CASES

Calibrating cases is different from calibrating source variables. In both approaches, researchers start with the definition of the set (concept): What does it mean to be a case of X? Calibrating cases, researchers then **establish whether each of their cases meets the criteria for being a case of X** (or the degree in which they do for fuzzy sets). For example, researchers establish whether each of the households in their case population makes ends meet. This calibrates households as poor (1) or not poor (0). Or poor in degree, depending on what specifically a household can(not) afford (e.g., housing, food, clothing, etc.). Calibrating cases makes that there is **not necessarily a blanket crossover point for all cases**. A household earning £1,700 may not make ends meet in expensive London, however, a household in poverty-stricken Blackpool may make ends meet earning only £1,400. A family earning £2,000 but fails to make ends, because they spend "too much" on leisure and luxury, may not qualify as a poor households, whereas a family that manages to make ends meet earning just £1,200 may still be calibrated a poor household. Calibration is (always) a dialogue between ideas (definition of the concept) and evidence (households making or not-making ends meet). Calibrating cases, this dialogue is conducted on case-level to **allow for relevant context**. Calibrating cases is the most accurate way of calibration but also the most labour intensive and may be unfeasible in large-N studies.

CALIBRATING SOURCE VARIABLES

Calibrating source variables, researchers **set a floor** (the threshold of full non-membership), **crossover point** (see above) **and ceiling** (the threshold of full membership) **in the distribution of their source variable**. For example, calibrating the set of poor households from the source variable 'income', a researcher may decide to set the crossover point at the minimum wage (in this example, £1,500) – on the plausible assumption that the minimum wage is set to allow households make ends meet. Although we do not know whether all households above the minimum wage make ends meet (some will, some will not), the minimum wage is a plausible value to distinguish between relevant and irrelevant variation in the source variable. Calibrating source variables thus **abstracts from cases**. The **dialogue between ideas and evidence is conducted on the level of the source variable to distinguish between relevant and irrelevant variation**. Consequently, some cases may end up on the wrong side of the crossover point. It underlines the difference between meaning (being poor) and measurement (minimum wage). The floor and ceiling of a source variable should be set in a way that is similar to setting the crossover point (see, setting thresholds).

SETTING THRESHOLDS

To set thresholds, most notably the crossover point but also the thresholds of full membership and full non-membership (when calibrating source variables) as well as setting empirical thresholds for linguistic hedges, researchers must rely on the following:

Substantive criteria are informed by theoretical and substantive knowledge of a concept. For example, substantive knowledge of poverty suggests that poor households cannot make ends meet, i.e., cannot afford the basic costs of living: housing, food, fuels, water and clothing. Substantive criteria may come from the scientific and professional literature but also from talking to experts. Substantive criteria

inform what researchers must look for when calibrating cases, or where to set plausible thresholds in source variables. Substantive criteria are the most robust and valid basis for calibration.

External criteria are empirical values that are external to researchers' data; they do not follow from the distribution (means, averages, percentiles) of the source variable. For example, a researcher may have collected income data from households and then use an external criterion to set the crossover point between poor and not-poor households. The minimum wage is such an external criterion because it is unrelated to income distribution and set by someone else (the government, in this case) with intimate knowledge what it means to be poor. Like substantive criteria, external criteria may come from the scientific and professional literature and from expert opinion. External criteria (empirical values) are often based on substantive criteria, such as the minimum wage for poor households.

Ranking works well with larger-N studies, particularly when the case population (almost) overlaps with a given population – because gaps in the data may suggest genuine thresholds rather than missing data. Suppose one investigates poverty in a population of 50 metropolitan regions and suppose these are all the metropolitan regions in a country. Rank the region with the highest poverty rate as 50 and rank the region with the lowest poverty rate as 1. Plotting the source data (regional poverty rates) on the X-axis and the rankings on the Y-axis (usually) produces an S-curvy line. Inspect the gaps and the bends in the line to set the floor, crossover point and ceiling in the source variable. Where the bottom of the S-curve stops being flat and starts to increase identifies a plausible floor. Where the top of the S-curve starts to flatten out identifies a plausible ceiling. Where the S-curve changes from increasing increase to decreasing increase identifies a plausible crossover point. Inspect cases on either side of a gap or bend to establish whether it is a plausible threshold or whether a nearby gap or bend is a more plausible threshold. That is, conduct a dialogue between ideas and evidence.

Inspecting gaps in the data can be used for smaller-N studies. Gaps may identify clusters of similar cases, particularly when the source data values on both sides of a gap are far apart. Each cluster of cases may correspond to a linguistic hedge. Inspect cases on both sides of a gap to establish whether the gap is a plausible threshold (i.e., dialogue ideas and evidence). Keep in mind that, with smaller Ns, gaps may merely suggest missing data (omitted cases).

Note that ranking and gaps produce inferior calibrations compared to substantive and external criteria. Note further that **calibration is always a dialogue between ideas and evidence** that is rarely, if ever, a one-shot process. Researchers **'update' their calibrations over the course of a research project**.

CALIBRATION METHODS

Depending on their cases and data, researchers can use a variety of calibration methods. It is perfectly acceptable to use different calibration methods for different sets (concepts) in the same study.

Manual calibration. Researchers 'manually' establish the (degree of) membership of each of their cases. Based on familiarity with their cases they decide whether a case (or in which degree) meets the

criteria specified in the definition of the concept (set). Manual calibration is particularly suitable for calibrating cases.

Indirect calibration may be performed on source variables that are not ratio-scaled. Researchers first decide on the number of degrees of set-membership (linguistic hedges). They then assign either the same set-membership value to all cases (source data) in a hedge (a cluster of cases) (Table 2, Column A). Or they assign set-membership values to cases (source data) within each hedge in a linear way (Table 2, Column B). Depending on whether they consider the variation within a hedge relevant or not. For example, largely poor households may be those earning between £1,400 and £1,350. A researcher can assign all largely poor households set-membership value 0.70 – because they all live the same largely poor lifestyle. Or assign all incomes between £1,400 and £1,350 a set-membership value between at 0.65 (for £1,400) and 0.75 (for £1,350) in a linear way – because a £1,375 household is still less poor than a £1,350 household.

Table 2: Indirect calibration

source variable income (£/month)	linguistic hedge	column A	column B	column C
1,775	not poor household	0	0	0.05
1,750			0	0.05
1,725	marginally poor household	0.15	0.05	0.06
1,700			0.10	0.08
1,675			0.15	0.11
1,650			0.20	0.14
1,625	somewhat poor household	0.30	0.25	0.18
1,600			0.30	0.23
1,575			0.35	0.29
1,550	moderately poor household	0.45	0.40	0.35
1,525			0.45	0.42
1,500			0.49	0.49
1,475	considerably poor household	0.55	0.51	0.57
1,450			0.55	0.63
1,425			0.60	0.69
1,400	largely poor household	0.70	0.65	0.75
1,375			0.70	0.80
1,350			0.75	0.84
1,325	mostly poor household	0.85	0.80	0.87
1,300			0.85	0.90
1,275			0.90	0.92
1,250			0.95	0.94
1,225	fully poor household	1	1	0.95
1,200			1	0.95

Direct calibration is for ratio scale source variables only. Applying a logarithmic formula on the source variable assigns a unique set-membership value to each data value. This effectively makes the direct method of calibration a data conversion. However, even with ratio scale source variables, set-membership values of more than two digits may still be a form of spurious specificity. The direct method is automated in QCA software but this does not suggest it is the preferred calibration method. Using £1,749; £1,499 and £1,224 as critical values for respectively the threshold of full membership, the crossover point and the threshold of full non-membership, the direct method of calibration

produces the set-membership values in Column C. Note that the logarithmic conversion makes differences close to the crossover point more pronounced (because they are more meaningful) and condenses those close to the threshold. Note also that the logarithmic conversion results in 0.95 and 0.05 as maximum and minimum set-membership values.

CALIBRATION FROM DIFFERENT KINDS OF DATA

QCA is agnostic to the kind of data that are used for calibration. However, different kinds of data require different approaches to calibration.

Calibration from qualitative data, such as interviews, are best used to develop in-depth case knowledge. Qualitative data are the basis for manual calibration of cases.
Calibration from quantitative data, either from existing data bases or surveys, is best done using the indirect or direct method of calibration. Depending on whether the source variable is ordinal, interval or ratio scale. Quantitative data are used for calibrating source variables.
Likert scales already have meaning, however, researchers must avoid using the neutral value as crossover point. This unhelpfully creates many 0.5-cases (see above). Instead, researchers decide whether the neutral value (e.g., 3) counts as 'in' or 'out' of the set. Accordingly, they then set the crossover point at 2.5 or 3.5, since those values do not exist on the Likert scale. Likert scales can be used for calibrating cases as well as calibrating source variables.
Calibration from existing scales is meaningful when those scales are commonly used, such as the World Sustainability Index, the Global Social Progress Index and various psychometric scales. Those scales are often composites of multiple indicators and researchers can treat these scales as source variables. Existing scales already have meaning and researchers may use this 'intrinsic meaning' to inform their calibration. More generally, **indicators of a concept** are best calibrated separately and then aggregated using logical ANDs and ORs, depending on the meaning of the concept.
Normalized scores, standardized scores and Z-scores have no meaning. They merely put the distribution of multiple source variables on the same scale. This is very helpful in correlational studies but useless in set-analysis, because it abstracts from the meaning of the scale. Therefore, standardized distributions cannot be a basis for calibration. If a source variable happens to be standardized, use the ranking method for calibration to distinguish between relevant and irrelevant variation.

VALIDITY OF A CALIBRATION

A good calibration produces a solution that can be interpreted into a plausible explanation of the outcome. Calibrations should be 'updated' if an interpretable solution does not follow. Calibrations must be transparent and reproducible and they must be plausible in the light of substantive and theoretical knowledge. There will often be multiple plausible calibrations from the same data meaning that **calibration is a judgement, not an exact science**. The empirically most robust calibration is not necessarily the best calibration. Less robust calibrations may produce solutions that are better

interpretable. The best calibration does not follow from an empirical exercise but from plausibly and transparently connecting meaning to measurements.

SUMMARY OF CALIBRATION METHOD AND PRACTICE

The table on page 8 gives a summary of calibration method and practice as discussed in this handout.

SUMMARY OF CALIBRATION METHOD AND PRACTICE

CALIBRATION ***sets meaningful thresholds between qualitatively different cases*** **WHAT ↓ HOW →**	**MANUAL CALIBRATION** ***identify linguistic hedges***	**INDIRECT METHOD OF CALIBRATION** ***identify linguistic hedges***	**DIRECT METHOD OF CALIBRATION** ***logarithmic conversion***
CALIBRATING CASES ***establishing (a degree of membership) for each case individually***	Qualitative data ①② Quantitative data ①②③④ Likert scales ①②	Qualitative data ①② Quantitative data ①②③④ Likert scales ①②	not applicable
CALIBRATING SOURCE VARIABLES ***putting a floor, crossover point and ceiling in the source variable***	Quantitative data ①②③④ Likert scales ①②	Quantitative data ①②③④ Likert scales ①②	Quantitative data ①②③④ (ratio scale source variables only)

CRITERIA TO SET THRESHOLDS: ① **substantive criteria** ② **external criteria** ③ **ranking** ④ **gaps in the data**

means, averages and percentiles are arbitrary criteria and produce spurious results