Exploring the consequences of a recalibration of causal conditions when assessing sufficiency with fuzzy set QCA

Authors:
Dr. Judith Glaesser (corresponding author) and Prof. Barry Cooper
School of Education
Durham University
Leazes Road
Durham
DH1 1TA
Tel. (JG): 0191-334 8308
Judith.Glaesser@durham.ac.uk
Barry.Cooper@durham.ac.uk

Judith Glaesser is a lecturer in the School of Education at Durham University. Her interests include sociology of education, inequality and meritocracy in education, and research methods, particularly QCA. She gained a PhD at Konstanz University (published as Soziale und individuelle Einflüsse auf den Erwerb von Bildungsabschlüssen). With Barry Cooper, she is applying case-based methods in comparing transitions in English and German secondary schools. A new book, Cooper, Glaesser, Gomm and Hammersley's Challenging the Qualitative-Quantitative Divide: Explorations in Case-focused Causal Analysis was published by Continuum in 2012.

Barry Cooper is Emeritus Professor of Education at Durham University where he was, from 1998 to 2005, Director of Research in Education. He was from 2004-2007 co-editor of the British Educational Research Journal. His interests are in the sociology of education, especially social class, educational achievement and assessment, set-theoretic research methods and the evaluation of educational aid projects. A representative book is, with Máiréad Dunne, Assessing Children's Mathematical Knowledge: Social class, sex and problem-solving. A new book, Cooper, Glaesser, Gomm and Hammersley's Challenging the Qualitative-Quantitative Divide: Explorations in Case-focused Causal Analysis was published by Continuum in 2012.

Acknowledgments:
This work was supported by an Economic and Social Research Council (ESRC) research fellowship [RES-063-27-0240] awarded to JG.

We gave an earlier version of this paper at the conference “Qualitative Comparative Analysis. Perspektiven für Politikwissenschaft, Soziologie und Organisationsforschung“, 1-2 June 2012, Hamburg University. We would like to thank participants for their valuable comments.

Word count: 5,994

Exploring the consequences of a recalibration of causal conditions when assessing sufficiency with fuzzy set QCA

Introduction

Charles Ragin's method Qualitative Comparative Analysis (QCA) (Ragin, 1987, 2000, 2008) focuses on the sufficiency and necessity of conditions or configurations of conditions for obtaining some outcome. Its use is increasing in the social sciences but some of its characteristics, especially those of its fuzzy set variant, are still not well-understood by users. One such feature concerns the paradoxical results that can arise when fuzzy sets are employed, something we have explored elsewhere (Cooper & Glaesser, 2011). In this paper, we focus on another important element of all QCA-based work, the process of allocating degrees of membership in the sets that enter the analysis[1]. We also employ an analytic approach that differs from the procedure embedded in the fsQCA software (Ragin & Davey, 2009), one we have used elsewhere for a different purpose (Cooper & Glaesser, 2010, Cooper, Glaesser, Gomm & Hammersley, 2012). We write on the assumption that users of methods, especially those embodied in easy to use software, always benefit from having an in-depth understanding of the procedures that underlie the results they obtain. Without such understanding, users are less likely, in our view, to produce valid conclusions because there is a danger that they do not apply the methods correctly or appropriately which can lead to flawed results.

[1] Skaaning (2011) also explores how this process of set calibration affects QCA results, but while he varies the fuzzy values of several conditions and the outcome, we will focus on the effect of changing just one value, holding the others constant. Thiem's (2010) work is also concerned with calibration. He explores the use of different membership functions to allocate fuzzy values, and what effect different membership functions have on coverage.

QCA is based on Boolean algebra. It analyses the relationships between sets of cases. It was originally developed for use with crisp sets, that is, with those where a case is simply either in or out of a set (Ragin, 1987). However, this only allows for dichotomous factors[2] and Ragin therefore developed QCA further to allow the analysis of fuzzy sets, that is, those where cases can have partial membership in a set. Fuzzy membership scores run from 0 (completely out of the set) to 1 (fully in the set), with 0.5 being the crossover point at which a case is as much in as out of the set. Allocation of set membership is straightforward for many crisp sets such as "female" or "university graduate" (assuming that there is no ambiguity over which HE institutions are regarded as universities), but it is less clear for some other conditions. The case of adulthood is frequently used to illustrate this problem. A 30 year old is clearly a full member of the set of adults, and a 10 year old clearly is not. But what about an 18 year old? We might allocate a fuzzy membership score of, say, 0.8, based on our knowledge of what adulthood means and what 18 year olds are like. However, if we were to apply a purely legal definition of adulthood, then 18 year olds would be full members of the set of adults in most countries.

This process of allocating membership scores is known as calibration (Ragin, 2008, chapter 4)[3]. Set membership scores form the basis of the analysis of consistency with sufficiency and necessity in QCA. It is clear therefore that we have to pay careful attention to how sets are calibrated. Various procedures are possible. One is, as in the adulthood example, to use some external definition of the set which is being calibrated. This would mean that the age of 18 (or

[2] Though dummy factors can be used to avoid this restriction. This is possible within conventional cs/QCA. Multi-value QCA, which was developed by Lasse Cronqvist, is another technique which allows for variables with at least three categories (e.g., Cronqvist, 2009).

[3] Since we wrote this paper, a new book has appeared which addresses some of the points discussed here, such as calibration (Schneider & Wagemann, 2012).

whatever the legal definition in the country under study is) acts as a threshold or cut-off point for full membership. Another is to use some distributional criterion. For example, having scored in the top 20% on some test can be calibrated as full membership of the set highly able, the bottom 20% would be fully out of the set, and the median or the mean could be used as the crossover point, i.e. the point where a case is as much in as out of the set. Obviously, the outcome of calibration would then depend on the particular sample being used, and it could result in having different calibrations based on different samples. But there can be good substantive grounds for using a distributional measure. For example, under the old British selective secondary school system, around 20% of children per cohort were allocated places at the selective grammar schools, based on their performance on a set of tests at age eleven which included a cognitive ability test. So it might make sense for some purposes, for example in studies of the degree to which educational selection is meritocratic, to define high ability as the top 20% of the performers on some test. It is also worth noting that we implicitly use distributional criteria when thinking about set membership. For example, any calibration of the set of tall people will be a function of the distribution of actual heights in a particular population. In this sense, the calibration of many sets is inherently distributional, whether based on a specific sample or on criteria based on knowledge of distributions in the world[4]. This is particularly so for sets whose members' properties have "positional" characteristics (Hirsch, 1976).

[4] Ragin (2008, p.80) points out that a sample-inherent measure such as the sample mean is not usually used to calibrate sets. We agree that properties of some sample should not be used in all cases to calibrate a set. Our argument is that such properties can sometimes give an indication of what values of some raw variable may be used as anchors, but this should only be applied on substantive and theoretical grounds.

Using first an abstract illustration and then an empirical example, this paper explores how two possible ways of calibrating a condition affect the degree of reported consistency with sufficiency of this condition with an outcome. We should note that one reason for employing different ways of calibrating the same raw values can be the wish to capture different concepts. For example, for both the sets "likely to have learning difficulties" and "highly intelligent" the raw variable "IQ" can be used as the basis for calibration, but the allocation of membership values to cases would look different. Ragin (2008, p. 33) describes a similar situation where GNP per capita may be used as an index variable to calibrate different sets, e.g., the sets of "wealthy countries" and "middle income countries".

We assume some knowledge of QCA throughout the paper, but by way of a reminder to the reader we will briefly explain the concepts we are using as we proceed.

Calibration and consistency with sufficiency: an abstract illustration

One definition of consistency with sufficiency in the fuzzy set context is that the membership value of the condition set has to be less than or equal to membership of the outcome set, i.e., x <= y, where x is the membership score for some condition and y is the membership score for some outcome (Ragin, 2000). Figure 1 illustrates diagrammatically a relationship of perfect consistency with sufficiency for multiple cases. We have added the y = x line to facilitate our discussion of calibration. All the cases are in the upper left triangle, i.e., on or above the y = x line – the region of consistency with sufficiency.

[Insert Figure 1 about here.]

Typically, we find relationships that are less than perfect, giving us some cases below the y = x line, i.e. in the lower right triangle. Such situations, where some but not most cases fail the criterion for sufficiency, are usually described with the term quasi-sufficiency (Ragin, 2000). We have constructed an invented dataset to demonstrate how a change in calibration can shift more cases into the region of consistency with sufficiency[5]. Figure 2 plots the values on some condition x against some outcome y. The original calibration is shown as diamonds, the new one as triangles. Because the condition values on the new measure (x2) are lower than those on the original calibration (x1), these are shifted to the left on the scatterplot, so that more cases now fall into the region above the y = x line, i.e. the area where the criterion for consistency with sufficiency, x <= y, is being met. For an example, consider the outcome "educational achievement" and the condition "cognitive ability". IQ might be calibrated initially to produce the set X1, "high ability", but, subsequently, higher levels of IQ might be required for any particular fuzzy set membership value for the set X2, "very high ability". It is easy to see that "very high ability" will be more consistent with sufficiency for the outcome than "high ability". However, in practice, either of these calibrations might be labelled by a researcher "high ability" – a potential source of confusion.

[Insert Figure 2 about here.]

The simplest measure of the degree of consistency with sufficiency is the proportion of cases with non-zero membership in x which satisfy the criterion, x <= y. The problem with this measure, however, is that cases that only narrowly miss the criterion count as much against

[5] We will use an example using empirical data later on in the paper. The reason for choosing an invented dataset initially is that this enables us to bring out the issue under discussion more clearly by illustrating a point in the abstract before moving on to more complex "real" data.

consistency as cases which miss it by a large margin. So a case with values of (0.9, 0.8) for X and Y violates the consistency criterion as much as one with values of (0.9, 0.2). Ragin (2006) explains and employs an alternative measure which takes account of near misses. It is calculated using the following formula, where min (X_i, Y_i) is the minimum for each case of the values of X and Y. This is the measure of consistency we use in the remainder of this paper.

$$\frac{\sum(\min X_i, Y_i)}{\sum X_i} \qquad \text{(Equation 1)}$$

QCA uses the measure of coverage to indicate how relevant or important a (set of) condition(s) is with regard to predicting an outcome. The formula to calculate coverage is given by equation 2[6]:

$$\frac{\sum(\min X_i, Y_i)}{\sum Y_i} \qquad \text{(Equation 2)}$$

Calibration and consistency with sufficiency: an empirical example

We now wish to illustrate the effect of alternative calibrations on consistency using real data. As noted, there can be good substantive and/or theoretical reasons for calibrating the same raw variable in different ways. In our example, we use cognitive ability as measured through an intelligence test as our condition. As noted in the introduction, the raw test scores here

[6] There are some problems specific to fuzzy set QCA, one of which is that paradoxical results can arise concerning consistency and coverage. For a discussion of these paradoxes and how they might be alleviated, see again Cooper & Glaesser (2011).

could form the basis of a variety of sets, such as "highly intelligent" as well as "likely to have learning difficulties".

Data and measures

Our example is set within the substantive area of sociology of education. We use data from Durham University's Centre for Evaluation and Monitoring (CEM). CEM conducts large scale educational monitoring studies whose main purpose is providing feedback to schools on pupils' performance, including value added analyses. In addition to performance indicators, background data are collected (for an overview of CEM's work see Tymms & Coe, 2003). Here, we use a combined dataset, comprising Yellis (**Ye**ar **11 I**nformation **S**ystem) and Alis (**A**dvanced **L**evel **I**nformation **S**ystem) data from 2005. All our cases therefore are Sixth form students, with n = 3188. During Year 11, pupils take a cognitive ability test (the Yellis test), provide background information and fill out an attitudinal questionnaire. The school later adds GCSE and A level exam results[7]. The Yellis test is designed to measure developed abilities which can predict GCSE performance.

To explore the effects of different calibrations on the results of tests for nearness to sufficiency, we analyse how cognitive ability combines with the factors parental education, gender and type of secondary school to predict A level performance. Using Ragin's direct method (Chapter 5 in Ragin, 2008)[8], we have calibrated the raw Yellis test scores in two

[7] GCSE stands for General Certificate of Secondary Education, the examination taken at the end of compulsory schooling in England, usually at the age of 16. A levels are the English qualification usually required for university entrance, obtained at the age of 18.

[8] Nothing much hinges on our choice of calibration method, given that this is a methodological paper rather than a substantive one. Ragin's direct method is one that is well suited to calibrating continuous data and is commonly applied, which is why we have chosen to use it here.

different ways, thus allocating cases to two fuzzy sets on the basis of their cognitive ability. The first, original, calibration uses the whole cohort, setting the cross-over point (i.e., a fuzzy value of 0.5) at the median for that group ("HIGH_ABILITY_1"), the second, new, calibration uses the more select group of A level pupils and their median as the cross-over point ("HIGH_ABILITY_2"). In the first case, we are effectively setting the reference population for constructing the set "high ability" as all secondary pupils; in the second the reference population comprises A level pupils. Given that test results are higher, on average, among the group of A level pupils compared to the whole cohort, the fuzzy values here are lower than those derived from the group as a whole. This is because we have, in effect, used a stricter criterion for various degrees of membership in the set "high cognitive ability". If we take fuzzy subsethood between two sets X2 and X1 as being defined by cases' membership values in X2 being lower or equal to those in X1 (Ragin, 2000), then HIGH_ABILITY_2 is a fuzzy subset of HIGH_ABILITY_1.

The other measures we use are parental education ("EDU_P) , gender ("MALE", coded MALE = 1 for boys and MALE = 0 for girls) and type of secondary school ("SCHOOL_SELECTIVE", coded SCHOOL_SELECTIVE = 1 for academically selective schools and SCHOOL_SELECTIVE = 0 for non-selective schools). EDU_P is a fuzzy measure based on educational attainment of both parents[9], the other two, MALE and

[9] This measure is based on parental education in five levels. We have converted each parent's level of qualification into fuzzy values (with "little or no formal education" = fuzzy value 0, "left school at the minimum school leaving age" = 0.33, "O levels or similar" = 0.66, "A levels, FE college or similar" = 0.83, "degree" = 1) and then, analogously to set theoretic intersection, we have used the minimum to define combined parental education. The latter is clearly a contestable decision, but it is of no consequence for the core argument in this methodological paper.

SCHOOL_SELECTIVE are crisp. The outcome measure is the average A level result ("ALEVEL_AV")[10], fuzzyfied using expert calibration.

Consistency: one condition

It is straightforward to calculate consistency with sufficiency for the two calibrations of the condition "high ability" with outcome "A level performance". For each case, its contribution to the sum in the numerator in the formula shown in Equation 1 is ruled by the lower of the two values, X and Y, simply because it is the minimum of these two values. Given that values for X are lower with the new calibration, this means that for more cases than before the values for the numerator and denominator will now be the same. In general, the value for any single case given by Equation 1 will be either higher or the same as we move from the initial to a stricter revised calibration of X. If HIGH_ABILITY_1 was already lower than or equal to the outcome, e.g. for a case like 0.8, 0.9, then Equation 1 would return the value 1 for both this calibration and any stricter calibration like HIGH_ABILITY_2 that reduced membership in HIGH_ABILITY. If, for example, the 0.8 became 0.7, Equation 1 would still return 1 for the single case 0.7, 0.9. For cases where X is higher than Y, for example for a case like 0.8, 0.6, then a reduction in the membership in X, produced through a stricter calibration, would increase the value returned by Equation 1. If 0.8, 0.6 became 0.7, 0.6, then the value would increase from 0.6/0.8 (=0.75) to 0.6/0.7 (=0.86).

[10] To obtain this fuzzy measure, we first calculated the mean result of all A levels obtained, with a grade A = 5, a grade B = 4 etc. We then turned this average into six fuzzy values as follows:

A level average	0	0.1 to 1.49	1.5 to 2.4	2.5 to 3.4	3.5 to 4.4	4.5 to 5
fuzzy value	0	0.17	0.33	0.66	0.83	1

The results of Equation 1 will therefore be higher for the new stricter calibration. Using the formula in Equation 1, we obtain a consistency value of 0.857 for the original calibration and of 0.910 for the new calibration. Of course, given that HIGH_ABILITY_2 is a fuzzy subset of HIGH_ABILITY_1, we expect to get at least the same overall consistency with sufficiency. If one condition set is a subset of the outcome set (the criterion for quasi-sufficiency), then another set that is a subset of the first will also be a subset of the outcome set. This is a straightforward consequence of the way subsethood is defined, but its implications for consistency and coverage, as calibrations are varied, need to be fully understood by researchers using QCA.

The discussion above was based on the relationship of one just condition set (calibrated in two different ways) with some outcome. However, one of the strengths of QCA is that it can examine conditions in conjunction with other factors and not in isolation, taking account of the conjunctural context in which conditions have their effects (Ragin, 1987, 2000, 2008). We therefore now turn to exploring how the analysis of the quasi-sufficiency of the *configurations* of factors is affected by this change in calibration.

Consistency: Truth table analysis

Initially, we explore how consistency with sufficiency changes using these two different measures of cognitive ability alongside the other conditions by examining and solving truth tables containing the conditions under study. In fuzzy set QCA, truth tables are constructed by allocating cases to rows according to whether they have values above or below 0.5 in the relevant configurations. We present two truth tables, each with the same conditions and the same outcome, but the first one using the original calibration of the cognitive ability test and the second using the new calibration. In these tables, the column headed "number" gives the

frequency of cases with a membership over 0.5 in each configuration. We will term such cases "good cases" for the configuration, since they are more in than out of the underlying set. ALEVEL_AV is the outcome "A level performance".

[Insert Table 1 and Table 2 about here.]

Following re-calibration, there are fewer members of the set HIGH_ABILITY_2 with values over 0.5 than of the set HIGH_ABILITY_1 and therefore we have different numbers of good cases in the different rows. For example, in Table 1, there are 616 good cases with the configuration HIGH_ABILITY_1*male* SCHOOL_SELECTIVE * EDU_P, but in Table 2 just 502 good cases showing the combination HIGH_ABILITY_2*male*SCHOOL_SELECTIVE* EDU_P. The order of the truth table rows has changed, too, reflecting the ways that the change in calibration affects the value returned by the minimum operator that is used to create membership values of the cases in each row[11], and hence consistency values.

QCA solves truth tables for quasi-sufficiency. The first step is to order the rows from high to low consistency with sufficiency, as we have already here. Then a threshold, usually around 0.8, is chosen for quasi-sufficiency. Rows, i.e. configurations, with a consistency score above the chosen threshold go forward into a minimised solution. The minimisation procedure is simple to understand, though can be hard to implement manually when there are more than a few conditions (Ragin, 2008). Basically the process repeatedly collapses pairs of rows that pass the threshold but only differ in having a single condition either present or absent. In

[11] The degree of membership in a configuration is calculated by taking the minimum of the memberships in the individual conditions (Ragin, 2000).

Table 1, for example, the first row, 0010, and the third row, 0011, differ only in having EDU_P either absent or present. Since both configurations have a high consistency score, and therefore can be considered to be quasi-sufficient for the outcome, we can say that 001- is quasi-sufficient for the outcome, where the dash indicates that the presence or absence of the fourth condition is logically irrelevant.

We can now examine initially the two overall solutions we obtain using a cut-off threshold that is suitable for both tables. For this purpose, 0.88 is suitable, since, in both tables, it marks a relatively large jump in consistency between the ordered rows. The 1s and 0s in the outcome column reflect our choice of threshold. The two solutions for quasi-sufficiency are:

For Table 1, using HIGH_ABILITY_1:

	raw coverage	unique coverage	consistency
	---	---	---
SCHOOL_SELECTIVE	0.360	0.084	0.873
HIGH_ABILITY_1*male	0.558	0.282	0.898

solution coverage: 0.642
solution consistency: 0.863

For Table 2, using HIGH_ABILITY_2:

	raw coverage	unique coverage	consistency
	---	---	---
SCHOOL_SELECTIVE	0.360	0.095	0.873
HIGH_ABILITY_2	0.676	0.411	0.910

solution coverage: 0.771
solution consistency: 0.875

It is easy to see that in the first solution high ability has to be conjoined with being female to be quasi-sufficient, but in the second it is quasi-sufficient on its own. However, it is not straightforward to interpret the differences between these two solutions arising from the recalibration of HIGH_ABILITY. Were we to employ just one calibration of HIGH_ABILITY but to vary the threshold for consistency, we would expect there to be a trade-off between consistency and coverage figures, with one rising as the other falls (Ragin, 2003). However, moving between these two calibrations with the threshold held constant, we obtain a rise in both overall consistency and coverage. The small rise in consistency results from the more selective definition of HIGH_ABILITY. The rise in coverage results from the fact that there are different configurations in these two solutions[12]. Ten rows enter into the first solution, while twelve rows enter the second solution. This produces the increase in explanatory coverage.

There are other ways to compare the effect of the recalibration. One is to compare the change in consistency figures for particular configurations. However, the fact that configurations may not contain exactly the same number of non-zero cases when using the original and the new calibration means that a direct comparison of consistency figures for truth table rows is likely to be misleading[13]. Furthermore, were we to use just good cases, rather than all non-

[12] We should note that moving from a less strictly calibrated to a more strictly calibrated measure of ability, while it will, *for any single case*, raise consistency with sufficiency scores, will also lower coverage scores. Consider an invented illustrative case, moving from 0.8, 0.7 to 0.5, 0.7. Consistency rises from 0.875 to 1. Coverage, however, falls from 1 to 0.714. Substantively, in this simple example, where coverage for sufficiency equates with consistency with necessity, we could say that while high ability is necessary for the outcome, very high ability is not. The situation may be different, as we have seen, for overall solutions.

[13] In our particular example, these differences in the number of non-zero membership cases are very small, but in other cases they may be large.

zero cases in the calculation of consistency, as we suggest in Cooper and Glaesser (2011), then we would be comparing rows with quite different numbers of good cases. For example, we can compare the configurations HIGH_ABILITY_1*male* SCHOOL_SELECTIVE * EDU_P and HIGH_ABILITY_2*male* SCHOOL_SELECTIVE * EDU_P. However, as noted above, the former comprises 616 good cases and the latter, 502. It is also the case that, if we were to set a minimum number of good cases for a row to be entered into the minimisation process, in order to base any conclusions on an adequate number of cases, then, as a result of the recalibration, different rows would be omitted from Tables 1 and 2. Using a cut-off of ten, for example, would omit row 0110 from Table 1 but row 1110 from Table 2[14]. We need therefore a mode of comparison that allows us to hold the context for comparing configurations across the two calibrations constant. To achieve this, we will therefore undertake a different form of comparison, one that takes account of the constellation of factors while keeping the focus on the effects of the change in calibration of the ability measure, but within particular configurationally defined sets of cases. We will draw on an approach that we have developed for another purpose in a previous methodological paper (Cooper & Glaesser, 2010; expanded in Cooper et al., 2012).

Consistency: Within configurations

This approach moves us away from truth table analysis. Instead, we calculate the consistencies for the two measures of high ability within configurations defined by the other factors of interest. Gender and type of school (non-selective vs. selective) are already dichotomous, and for this analysis we have changed the fuzzy measure for parental education to one with four categories, which results in 2 x 2 x 4 = 16 configurations. As expected, within each configuration consistency is higher for the newly calibrated measure,

[14] The use of counterfactual analysis (Ragin, 2008) would further complicate things here.

"HIGH_ABILITY_2". We can see in Table 3 that, if we were to take 0.8 as the cut-off for quasi-sufficiency, two of the sets of cases fall well below it when HIGH_ABILITY_1 is used (with one other type of case just below this level), but all types of case, bar one (which nearly reaches it), surpass this level when HIGH_ABILITY_2 is employed. On the newly calibrated measure, even the lowest consistency value – that for boys in non-selective schools who do not have highly educated parents – reaches nearly 0.8. If we were to employ a stricter cut-off for quasi-sufficiency, say 0.9, we would produce a different account of the role of "high ability" dependent on which of these two calibrations we chose to use.

[Insert Table 3 about here.]

These results are summarised in Figure 3 which shows the consistencies for the 16 configurations both for the original and newly calibrated measures, grouped by gender.

[Insert Figure 3 about here.]

For those groups who showed fairly high consistency on the original measure, the difference between this and consistency of the new measure is rather small. This points to ceiling effects operating here: for students in selective schools in particular, consistency is so high on the original measure already that re-calibration hardly affects this[15]. We can move to a substantive interpretation more easily if we agree to a relabeling of HIGH_ABILITY_1 and

[15] Consider X1. If a case has high consistency on this measure this will be because X1 is lower than or equal to the outcome or very near to being so. Take the case of X1 being lower than or equal to the outcome. Equation 1 will return a value of 1 for such cases. If we move to using X2, then each case X2 will be lower or equal to X1 on the condition membership score, but there is no scope for this to improve the value returned by Equation 1, since it is already 1.

HIGH_ABILITY_2 as, respectively, highly able and very highly able. We can then say that for students in such schools, being in the set of the "highly able" is a quasi-sufficient condition for high achievement. Being "very highly able" does not have any additional benefits for the outcome measure we are using[16], given the favourable circumstances they find themselves in[17]. For most other types of students, only being a member of the set "very highly able" is quasi-sufficient for high achievement. This illustrates once more the importance of configurational context: the effect of conditions such as cognitive ability has to be analysed against the background of specific other factors, since their effect at least partly depends on this context.

Conclusion

As well as providing an insight into the effects of changing calibration on the consistency of a condition, assessing consistency within configurations can be useful as a complement to "standard" truth table based QCA, especially for researchers who are interested in the relationship of two particular factors and who want to explore the way the relationship changes as the configurational context is changed. In a previous paper, we used this approach to explore the degree to which measured ability, within configurations defined by parental class, grandparental class and gender, was quasi-necessary and/or quasi-sufficient for certain levels of educational achievement. We were able to show, for a cohort of British cases born in 1958, that high ability tended to be necessary but not sufficient for cases from lower class origins, and sufficient but not necessary for those from higher class origins (Cooper &

[16] Clearly, this outcome could also be recalibrated for some purposes.

[17] All such conclusions are only as good as the model specified (Steel, 2011). It is possible that these findings would be different were we to be able to include measures of ambition, motivation, etc.

Glaesser, 2010; Cooper et al., 2012). We also showed that conventional correlational procedures failed to model this causal complexity.

Similarly, in this paper, our approach has been useful in exploring, set theoretically, the consequences of different levels of ability for different types of cases, with these here being configurationally defined in terms of gender, parental education and type of schooling. While a lower level of ability is quasi-sufficient for high achievement for the most socially advantaged types of case, only a higher level of ability can be considered to be quasi-sufficient for the less advantaged. This provides an example of the benefits of applying distributional criteria to calibration, as discussed in the introduction: given higher cognitive ability, on average, in the group of A level students, we were able to use this group as our reference sample to calibrate the set "very highly able", having used the whole group as the reference sample to calibrate the set "highly able". This threw additional light on the way ability interacted with other factors in our dataset.

In the present paper, we have shown how markedly calibration can affect analytic results. This emphasises the need for us to take a very careful and considered approach when calibrating raw variables, bearing in mind the danger of arranging the calibration of the data to suit some desired result[18]. A theory might make a claim about "high ability", but we have to be able to judge whether some empirical test is using the term in quite the same way as the

[18] We have noted that stricter calibrations of a condition produce lower fuzzy values and that this leads to higher consistency. However, lower fuzzy values mean that there are fewer "good" cases, or instances of a condition, all else being equal. A lack of enough good cases can make it difficult to justify claims about relations of sufficiency and necessity (Ragin, 2006, p. 295). This must also be borne in mind.

theory[19]. Once a factor has been calibrated in a particular way, and given a verbal label, it is easy to forget the dependence of the solution on calibration decisions. A similar argument applies, of course, to the operationalisation of concepts within conventional approaches such as regression analysis. As increasing numbers of social scientists begin to use QCA, it is important that they focus as much attention on the role and effects of calibration as they would on such operationalisation decisions.

We have focused here on fuzzy set QCA. Given that crisp set QCA is, in some senses, a special case of fuzzy set QCA, we would expect a similar phenomenon to occur[20]. If a criterion for set membership of a crisp condition set was made stricter, this set would be smaller and therefore, given no changes to the outcome set, the condition set would be likely to come closer to being a perfect subset of the outcome set and consistency with sufficiency would be higher.

[19] Alongside this possible confusion of meaning, there is, as we have just mentioned, also a potential problem akin to curve-fitting, where calibrations might be modified purely to raise observed consistency levels. It is also the case that a shift to stricter calibration of a fuzzy condition will reduce the number of good cases, i.e. those with membership over 0.5, in some configurations in any truth table (while increasing it in others). Furthermore, if Ragin's truth table algorithm is used, which includes all cases with non-zero membership in a configuration in tests for sufficiency, this recalibration would tend to push more cases into the paradoxical region for the measurement of consistency (see Cooper & Glaesser, 2011, for more detail).

[20] As far as we are aware, multi-value QCA (Cronqvist, 2009) currently does not allow for less than perfect subset relations. Therefore, no consistency measure exists and our argument does not apply. However, in principle a parallel argument could be made.

References

Cooper, B., & Glaesser, J. (2010). Contrasting variable-analytic and case-based approaches to the analysis of survey datasets: exploring how achievement varies by ability across configurations of social class and sex. *Methodological Innovations Online, 5*(1), 4-23.

Cooper, B., & Glaesser, J. (2011). Paradoxes and pitfalls in using fuzzy set QCA: Illustrations from a critical review of a study of educational inequality. *Sociological Research Online, 16*(3), http://www.socresonline.org.uk/16/13/18.html.

Cooper, B., Glaesser, J., Gomm, R., & Hammersley, M. (2012). *Challenging the qualitative-quantitative divide: explorations in case-focused causal analysis*. London & New York: Continuum.Cronqvist, L. (2009). Multi-value QCA (mvQCA). In B. Rihoux & D. Berg-Schlosser (Eds.), *Configurational comparative methods. Qualitative Comparative Analysis (QCA) and related techniques* (pp. 69-86). Thousand Oaks, CA: Sage.

Hirsch, F. (1976). *Social limits to growth*. Cambridge, Massachusetts: Harvard University Press.

Ragin, C. C. (1987). *The Comparative Method. Moving beyond Qualitative and Quantitative Strategies*. Berkeley, Los Angeles, London: University of California Press.

Ragin, C. C. (2000). *Fuzzy-Set Social Science*. Chicago and London: University of Chicago Press.

Ragin, C. C. (2003). *Recent advances in fuzzy-set methods and their application to policy questions* (Working paper): http://www.compasss.org/wpseries/Ragin2003a.pdf.

Ragin, C. C. (2006). Set Relations in Social Research: Evaluating Their Consistency and Coverage. *Political Analysis, 14*(3), 291-310.

Ragin, C. C. (2008). *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago: University of Chicago Press.

Ragin, C. C. & Davey, S. (2009). *fs/QCA*, Version 2.5. Tucson: University of Arizona. Website: http://www.u.arizona.edu/~cragin/fsQCA/software.shtml .

Schneider, Carsten Q. & Wagemann, Claudius (2012): Set-theoretic methods for the social sciences. A guide to Qualitative Comparative Analysis. Cambridge: Cambridge University Press.

Skaaning, S.-E. (2011). Assessing the robustness of crisp-set and fuzzy-set QCA results. *Sociological Methods & Research*, *40*(2), 391-408.

Steel, D. (2011). Causality, causal models, and social mechanisms. In I. C. Jarvie & J. Zamora-Bonilla (Eds.), *The SAGE Handbook of the Philosophy of Social Sciences* (pp. 288-304). London: Sage.

Thiem, Alrik (2010): Set-relational fit and the formulation of transformational rules in fsQCA. Compasss working paper 2010-61. http://www.compasss.org/wpseries/Thiem2010.pdf

Tymms, P., & Coe, R. (2003). Celebration of the success of distributed research with schools: the CEM centre, Durham. *British Educational Research Journal*, *29*(5), 639-653.

Figure 1: Relationship of perfect consistency

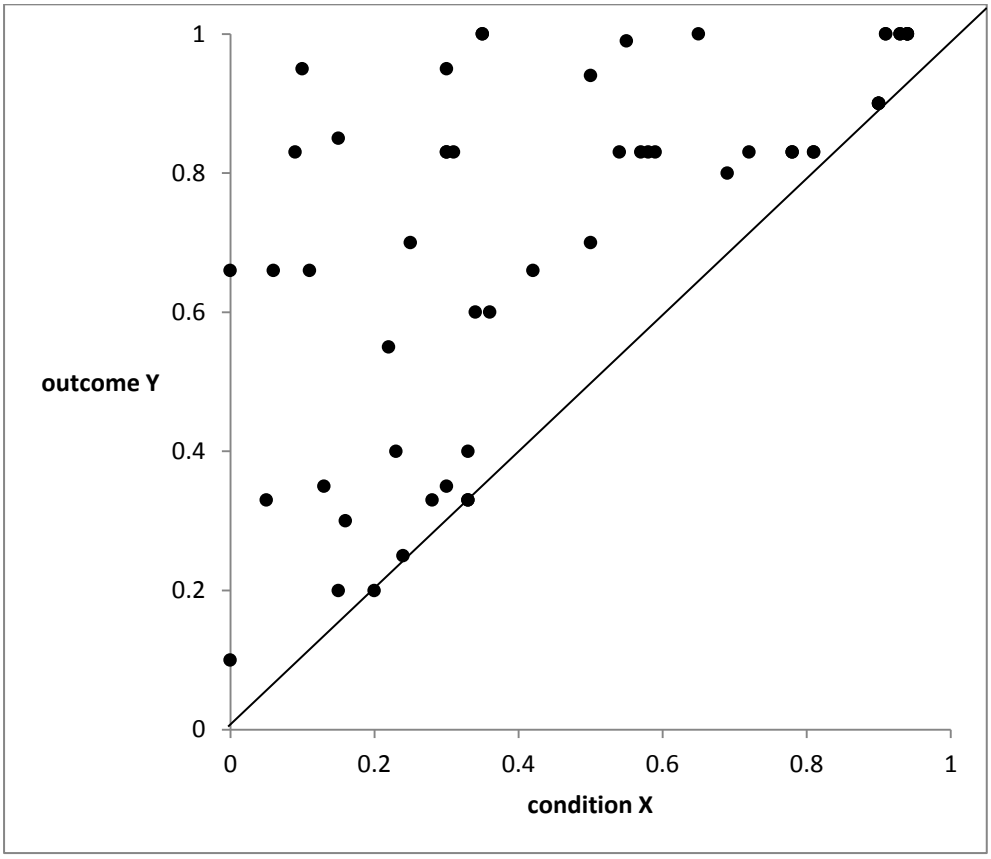

Figure 2: Two ways of calibrating some condition, consistency with an outcome

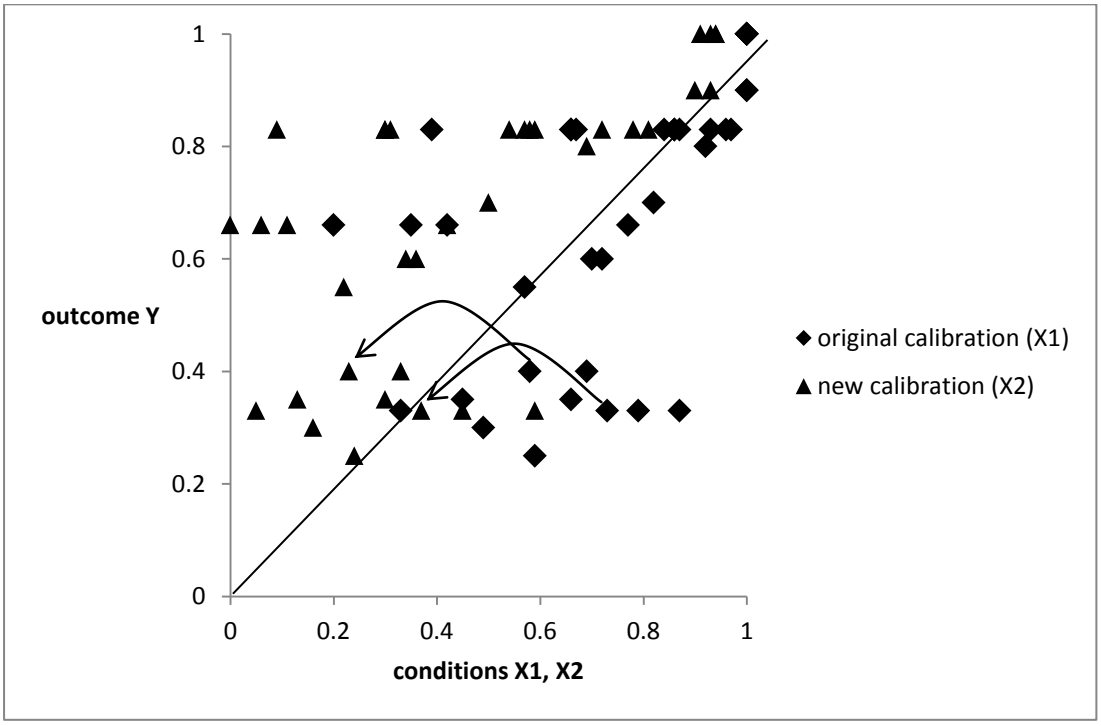

Figure 3: Consistency within groups, original and new calibration

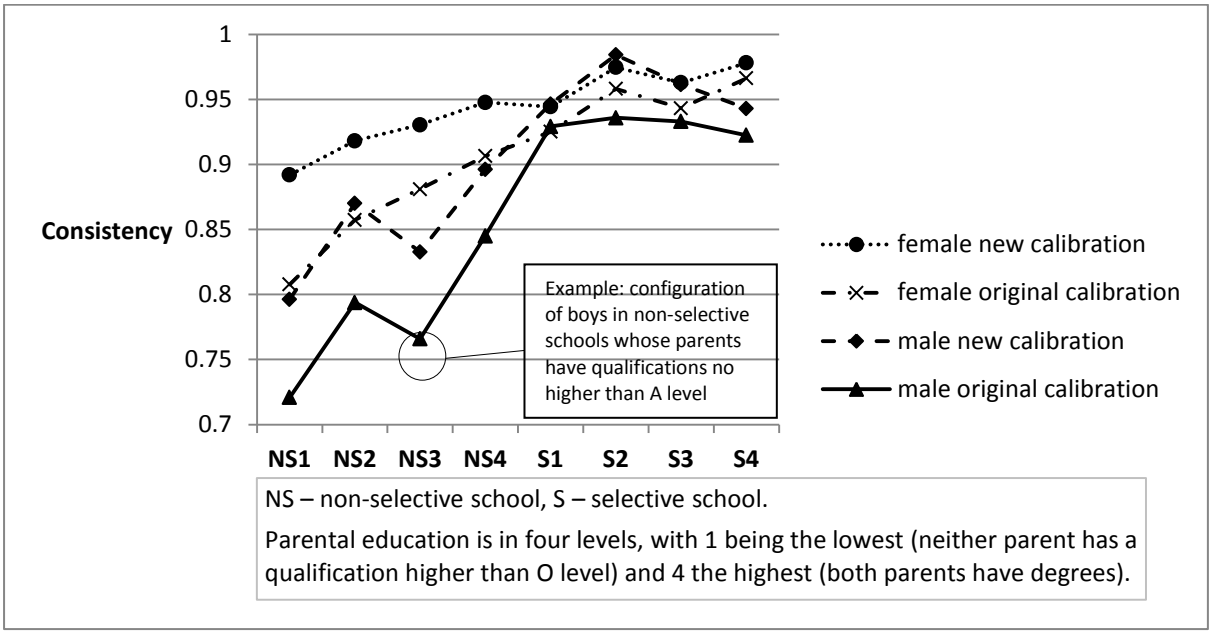

Table 1: Truth table, original calibration of cognitive ability

HIGH_ABILITY_1	MALE	SCHOOL_ SELECTIVE	EDU_P	number	ALEVEL_AV	consistency
0	0	1	0	15	1	0.993
1	1	1	0	12	1	0.993
0	0	1	1	48	1	0.985
1	0	1	0	104	1	0.983
1	0	1	1	616	1	0.974
0	1	1	0	6	1	0.955
0	1	1	1	16	1	0.945
1	1	1	1	83	1	0.941
1	0	0	1	517	1	0.908
1	0	0	0	259	1	0.904
1	1	0	1	510	0	0.844
0	0	0	1	246	0	0.840
1	1	0	0	243	0	0.839
0	1	0	0	122	0	0.830
0	0	0	0	212	0	0.823
0	1	0	1	179	0	0.806

Table 2: Truth table, new calibration of cognitive ability

HIGH_ABILITY_2	MALE	SCHOOL_ SELECTIVE	EDU_P	number	ALEVEL_AV	consistency
1	1	1	0	7	1	0.993
1	0	1	0	77	1	0.988
0	0	1	0	42	1	0.986
1	0	1	1	502	1	0.983
0	0	1	1	162	1	0.976
1	1	1	1	62	1	0.961
0	1	1	0	11	1	0.960
1	0	0	1	317	1	0.948
1	0	0	0	133	1	0.945
0	1	1	1	37	1	0.908
1	1	0	1	355	1	0.891
1	1	0	0	141	1	0.883
0	0	0	1	446	0	0.824
0	0	0	0	338	0	0.806
0	1	0	0	224	0	0.801
0	1	0	1	334	0	0.777

Table 3: Consistency within configurations

Gender	School type	Parental education	n	Consistency original calibration	Consistency new calibration
male	non-selective	O level is highest parental qualification	413	0.721	0.796
female	non-selective	O level is highest parental qualification	523	0.808	0.892
male	non-selective	A level is highest parental qualification	226	0.794	0.870
female	non-selective	A level is highest parental qualification	274	0.857	0.918
male	non-selective	one but not the other parent has a degree	249	0.766	0.833
female	non-selective	one but not the other parent has a degree	254	0.881	0.930
male	non-selective	both parents have degrees	166	0.845	0.896
female	non-selective	both parents have degrees	183	0.906	0.948
male	selective	O level is highest parental qualification	18	0.929	0.946
female	selective	O level is highest parental qualification	127	0.925	0.944
male	selective	A level is highest parental qualification	17	0.936	0.984
female	selective	A level is highest parental qualification	116	0.958	0.975
male	selective	one but not the other parent has a degree	36	0.933	0.961
female	selective	one but not the other parent has a degree	224	0.943	0.963
male	selective	both parents have degrees	46	0.922	0.943
female	selective	both parents have degrees	316	0.966	0.978