International Journal of Social Research Methodology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tsrm20

Exploring the robustness of set theoretic findings from a large n fsQCA: an illustration from the sociology of education

Barry Cooper[a] & Judith Glaesser[a]

[a] School of Education, Durham University, Leazes Road, Durham DH1 1TA, UK

Published online: 29 Apr 2015.

To cite this article: Barry Cooper & Judith Glaesser (2015): Exploring the robustness of set theoretic findings from a large n fsQCA: an illustration from the sociology of education, International Journal of Social Research Methodology, DOI: 10.1080/13645579.2015.1033799

To link to this article: http://dx.doi.org/10.1080/13645579.2015.1033799

International Journal of Social Research Methodology, 2015
http://dx.doi.org/10.1080/13645579.2015.1033799

Routledge
Taylor & Francis Group

Exploring the robustness of set theoretic findings from a large n fsQCA: an illustration from the sociology of education

Barry Cooper* and Judith Glaesser

School of Education, Durham University, Leazes Road, Durham DH1 1TA, UK

(*Received 15 July 2014; accepted 19 March 2015*)

Ragin's Qualitative Comparative Analysis (QCA) is often used with small to medium samples where the researcher has good case knowledge. Employing it to analyse large survey datasets, without in-depth case knowledge, raises new challenges. We present ways of addressing these challenges. We first report a single QCA result from a configurational analysis of the British National Child Development Study dataset (highest educational qualification as a set theoretic function of social class, sex and ability). We then address the robustness of our analysis by employing Duşa and Thiem's R QCA package to explore the consequences of (i) changing fuzzy set theoretic calibrations of ability, (ii) simulating errors in measuring ability and (iii) changing thresholds for assessing the quasi-sufficiency of causal configurations for educational achievement. We also consider how the analysis behaves under simulated re-sampling, using bootstrapping. The paper offers suggested methods to others wishing to use QCA with large n data.

Keywords: Qualitative Comparative Analysis; calibration; fuzzy sets; robustness; simulation; bootstrapping

Ragin's (2008) Qualitative Comparative Analysis (QCA) analyses minimally necessary and/or sufficient conditions, or configurations of conditions, for some outcome. Most published work uses QCA in the context of small to medium sized datasets, but there is also work with large datasets. This addresses, for example, education and stratification (Cooper, 2005; Cooper & Glaesser, 2008, 2010, 2012a; Glaesser, 2008), poverty (Ragin, 2006; Ragin & Fiss, 2008), management (Greckhamer, Misangyi, Elms, & Lacey, 2008), and typologies (Cooper & Glaesser, 2011a; Fiss, 2011).

Greckhamer, Misangyi, and Fiss (2013) note that QCA's nature changes when it is applied to large datasets. One challenge is the relative lack of detailed case knowledge. This is important, though as these authors note, one can study selected cases within the framework of a large n QCA. We have used this approach to develop theoretical understanding (Cooper & Glaesser, 2012b). Another challenge is to understand how QCA behaves with random sampling. We believe that the benefits of QCA's configurational approach should be available to all social scientists, not just those studying small samples/populations. However, to make QCA more acceptable to large n researchers, scholars need to explore further the nature of the challenges inherent in using QCA with large n data and develop responses.

*Corresponding author. Email: Barry.Cooper@durham.ac.uk

Here, we draw on experience of using fsQCA with survey data in the sociology of education. In typical work on education and meritocracy, comparisons of systems, across time and space, using longitudinal datasets, are undertaken mostly using correlation-based methods, focusing on the net effects on educational achievement of ability vs. ascriptive factors. We have argued that the question of meritocracy can be usefully re-phrased in terms of necessary and sufficient conditions, creating such hypotheses as, 'for girls from disadvantaged social backgrounds, high ability tends to being necessary but not sufficient for high educational achievement' (Cooper & Glaesser, 2010, 2012a). We have also used QCA simply to explore the pathways to high educational achievement (Cooper, 2005). We will focus our methodological discussion here on that task, exploring results from a set-theoretic re-analysis of the British National Child Development Study (NCDS) dataset (educational qualifications as a function of social class, sex and ability). The NCDS follows individuals born in one week in 1958. The cohort is conventionally treated as if it comprises a random sample from a population. We treat it as such.

Our data comprise 6666 cases.[1] The variables are: social class origin, highest qualification at age 33, sex and 'ability'. Father's social class has been coded to the categories of the scheme employed in Breen and Goldthorpe (1999). Highest qualification has six categories running from 'no qualification' to 'degree or higher'; 'ability' derives from a general ability test taken at age 11 (variable: n920).[2]

We initially present a single solution. Our focus then moves to its robustness. Previous robustness studies (Hug, 2013; Krogslund & Michel, 2014; Lucas & Szatrowski, 2014; Schneider & Wagemann, 2012; Seawright, 2005; Skaaning, 2011) have either focused on the effects of choices scholars must make in calibrating sets, in choosing consistency thresholds for sufficiency/necessity, over frequency thresholds for allowing configurations into the minimised solution, or have simulated the effects of variously dropping one or two cases, dropping conditions or simulating error. The nature of these studies reflects QCA's origins in analysing small to medium sized datasets, often whole populations such as 'welfare states'. In such analyses the ratio of cases to explanatory conditions is often very low and there are usually configurations for which there are no cases (limited diversity). Marx and Duşa (2011) have shown that, with very low cases/conditions ratios, QCA can misleadingly find 'explanatory' models when faced with random data. They show that, with three conditions – the number we employ – around a dozen cases are required to avoid this problem. With our 6666 cases, this problem need not concern us.

Much of this work on robustness re-analyses studies where this key ratio is very low, breaching or nearly breaching Marx and Duşa's guidelines (for example, Hug, 2013; Krogslund & Michel, 2014; Lucas & Szatrowski, 2014; Schneider & Wagemann, 2012; Skaaning, 2011). Because of this, these studies – some of which report that QCA solutions are very sensitive to a scholar's choices – are strongly affected by the problem of limited diversity. The latter necessitates the analysis of truth tables which are not fully populated. Some rows, termed 'logical remainders', lack cases. Ragin (2008) has shown how, via counterfactual reasoning about all or some of these remainders, drawing on theoretical or case knowledge, scholars can access three types of solutions (parsimonious, intermediate or complex). Much of the reported sensitivity of QCA solutions to scholars' choices arises from the way in which the set of logical remainders changes, in small n contexts, as calibration anchors etc. are varied. In large n contexts, however, unless a large number of conditions are employed, such limited diversity is usually absent, and the three types

of solutions collapse into one. At no point in our analyses does the problem of limited diversity arise.

The aspects of previous work most relevant for us are set calibration, consistency thresholds and measurement error. We incorporate changes in these into our analyses. However, given our large n focus, we will also address a key issue that doesn't arise in small n work analysing whole populations: the consequences of sampling error for the stability of solutions. To summarise, we address robustness by employing Duşa and Thiem's (2013) R QCA package (also: Thiem & Duşa, 2013) to explore the consequences of (i) changing fuzzy set calibrations of ability, (ii) errors in measuring ability and (iii) changing key thresholds for assessing the nature, as well as the consistency and coverage, of set theoretic solutions for educational achievement. We also use bootstrapping techniques to explore the stability of our solutions when faced with sampling error. We focus our attention, for illustrative purposes, as have others, on just one of our 'causal conditions': ability. We treat the measurement and calibration of the others as givens. This makes it easier, in a methodological paper, to present our arguments and suggestions. In addition, the condition 'ability' is particularly interesting. If we take the fuzzy set 'high ability', there is no obvious theoretically grounded way of setting the key calibration points (Glaesser & Cooper, 2014). Measures of ability are already partly constructed to meet conventional distributional criteria. Furthermore, given some calibration of educational achievement, some ability calibrations are likely to be more successful than others (in terms of QCA's measures of 'consistency' and 'coverage' for sufficiency[3]) in modelling pathways to it. Notwithstanding the warnings of various authors concerning data mining (for example, Greckhamer et al., 2013), we see no reason why varying the crossover point[4] for 'high ability' and observing the effect of this on consistency and coverage should be ruled out, given the nature of measured ability itself.[5]

We discuss, in the light of our experiments, whether our initial solution is sound and/or whether some other solution is preferable, taking account of the consistency and coverage of the solutions, but also the PRI measure of consistency, which aims to alleviate the paradoxical results that can arise when fuzzy logic is employed (Cooper & Glaesser, 2011b). We also consider whether it makes sense to talk about a 'best solution'. Lastly, we consider how our initial solution behaves under simulated re-sampling. Then, in our conclusion, we reconsider the differences between assessing robustness in small and large n contexts, and set out the areas we believe need further work.

1. The single solution

Initially, we report the result of a QCA in which our outcome, highest qualification (HQUAL33F in Table 1) is a fuzzy set,[6] as are two of our conditions, class[7] and ability, while sex (MALE) is a crisp set (with males = 1). In the first analysis reported here, ability (0–80) is calibrated by Ragin's (2008) direct method,[8] using the R QCA package, with the point for full exclusion at 25, the crossover at 45, and the point for full inclusion at 65. The raw consistency threshold (for sufficiency) is set at .8[9] (see Table 1 where the rows passing the threshold are shown in bold).

We should comment on our choice of 45 for the initial crossover point for the ability calibration. As it happens, this value is close to the mean of the underlying variable. However, this is not our reason for choosing it. We have already noted that

Table 1. The truth table.

MALE	CLASS	ABILITY	Number	HQUAL33F (the outcome)	Raw consistency	PRI consistency
1	**1**	**1**	**872**	**1**	**0.879**	**0.786**
0	**1**	**1**	**1008**	**1**	**0.834**	**0.695**
1	**0**	**1**	**694**	**1**	**0.830**	**0.655**
1	1	0	564	0	0.761	0.476
0	0	1	835	0	0.753	0.482
0	1	0	454	0	0.725	0.356
1	0	0	1057	0	0.593	0.275
0	0	0	1012	0	0.531	0.170

Note: The rows passing the threshold are shown in bold.

there are no obviously correct calibration points for sets like 'ability'. Varying the crossover point for ability clearly will generate different solutions. For each solution, for any particular crossover value, we must make sense of what the set 'ability' means. For example, if we set the crossover point at 55, then fsQCA reports[10] that ability *on its own* is quasi-sufficient for high achievement (consistency: .821, coverage: .617). In this case the set is best understood as comprising 'very high ability'. If, instead, we set the crossover at 30, then ability is not quasi-sufficient (consistency: .678) but is now quasi-necessary (consistency: .887). Here the set is best understood as something like 'moderate ability'. That 'very high ability' is quasi-sufficient and that 'moderate ability' is quasi-necessary are both interesting findings. However, as sociologists, we are primarily interested in the ways in which factors such as class and sex act together with ability in predicting achievement and, for this purpose, we require a set with a crossover somewhere between these values of 30 and 55. We find that 45 is a choice that provides sociologically interesting findings. Using this calibration we find ABILITY (best read as 'high ability' given the calibration) must be combined with either MALE or CLASS (best read as 'higher class origin') to be part of a configuration quasi-sufficient for the outcome. Table 2 shows this solution. The coverage figures suggest that the second term is empirically more important, but this partly reflects the binary nature of the sex factor. We have (MALE*ABILITY) + (CLASS*ABILITY) ⇒ ACHIEVEMENT[11] where the * indicates set intersection (logical AND) and the + set union (logical OR).

Having established this solution, which, roughly speaking, shows that high ability, to be quasi-sufficient for high achievement, must be combined with one of higher class origins or being male, we now, in Sections (2.1–2.3), explore its robustness by drawing on a database of 15655 solutions created by running R QCA in three nested loops. The outer loop runs through the original ability calibration plus

Table 2. Single solution (from minimising the first three rows of the truth table).

	Raw coverage	Unique coverage	Consistency
MALE*ABILITY	0.369	0.115	0.799
CLASS*ABILITY	0.525	0.270	0.855
Solution coverage	0.639		
Solution consistency	0.813		

100 'new' ability factors, created by adding random measurement error to the original non-calibrated variable. The middle loop allows the crossover point for the calibration of ABILITY to vary (30, 35, 40, 45, 50). The inner loop runs through 31 threshold levels for consistency (.70, .71, … , .99, 1.00). This gives us 101 × 5 × 31 (15655) solutions to examine.[12]

2.1. *Stability of the solution as the ability calibration is varied*

Here we discuss how the solution varies when, holding everything else the same as in (1), we vary the ability calibration. We allow the crossover point for ability to take, sequentially, the values 30, 35, 40, 45, 50. The solutions are in Table 3. The obvious point to note is that the two lower crossover points, which reduce the difficulty of entering the set 'high ability', have the effect of requiring more supportive terms in the conjunction. Importantly, because the term MALE*CLASS*ABILITY excludes all females, we find coverage greatly reduced for this solution. The table is easy to interpret. As the crossover point is lowered, we effectively reduce the 'highness' in the set of 'high ability cases' (ABILITY). At lower levels (30, 35), two ascriptive factors need to be conjoined with ABILITY to achieve consistency above .8. At the three higher levels (40, 45, 50), only one is required.

2.2. *Stability of the solution as error in measuring ability is simulated*

We have created 100 error-affected versions of 'ability' by adding to each score, prior to calibration, a random variable representing measurement error. The added error term is normally distributed with a mean of zero and a standard deviation of 5. Given that our ability variable has a mean of approximately 45 and a standard deviation of 15 in our sample, this is a fairly severe test of robustness. Holding everything else the same as in the analysis under (1) but running through these 100 error-affected solutions (plus the original), we find that all 101 solutions are of the form (CLASS*ABILITY) + (MALE*ABILITY). The overall consistencies range from .804 to .813, the coverage indices from .629 to .639. Solutions are stable under this trial.

2.3. *Stability of the solution as the consistency threshold is varied*

Here we hold everything the same as in the single solution under (1) except the consistency threshold. We have 31 potential solutions. In fact, once the threshold reaches .88, no rows pass from the truth table, leaving us 18 solutions.[13] Varying the threshold does affect the solutions dramatically. Again, coverage drops abruptly as females disappear from the MALE*CLASS*ABILITY (.84–.87) solutions (Table 4).

Table 3. Varying crossover.

Crossover	Overall solution	Consistency	Coverage
30	MALE*CLASS*ABILITY	0.833	0.294
35	MALE*CLASS*ABILITY	0.849	0.284
40	CLASS*ABILITY + MALE*ABILITY	0.789	0.682
45	CLASS*ABILITY + MALE*ABILITY	0.813	0.639
50	CLASS*ABILITY + MALE*ABILITY	0.838	0.589

The solution (CLASS*ABILITY) + (MALE*ABILITY) + (MALE*CLASS) has a good balance of consistency and coverage but is unstable in that a small shift in the threshold up or down changes its nature. Given this instability of (CLASS*ABILITY) + (MALE*ABILITY) + (MALE*CLASS) perhaps the choice here is essentially between the solutions between which it is sandwiched. This choice involves a trade-off between consistency and coverage. Favouring consistency leads to the choice of (CLASS*ABILITY) + (MALE*ABILITY), the solution we found in (1).

3. All 15655 solutions

Of the complete set of 15655 combinations of crossover point, error-affected variables and consistency threshold 7219 (46.11%) produce no solution. Since there are set theoretic relations to be found in these data, this in itself serves as an argument that the analyst may need to explore varying calibrations. Giving up just because the first set of parameters chosen fails to produce a solution would not seem a sensible strategy, although this might be argued for by someone believing that the calibration process should be entirely theoretical and/or substantive. In Section (3.2) we consider the whole run of potential solutions, but only after looking first, in Section (3.1), at just the subset of 155 that employ the original (but then calibrated) ability score.

3.1. The 155 analyses using the original ability score (with no error added)

We can learn something from examining the distribution of solutions over the space created by varying threshold and crossover values. To begin with, for simplicity, Table 5 just looks at ability calibrated as a fuzzy set without added error. The maximum number of solutions per cell is therefore 1. Once the threshold reaches .9 there is no solution whatever the crossover point for ability. Below .9 there is an interaction between the crossover point – which partially sets the degree of highness in 'high ability' – and the threshold. Taking the threshold of .84 for illustration, we can see that as 'ability' is calibrated 'more selectively', the solution changes (shown in bold in Table 5). We have no solution when the crossover point is 30. We then, at 35, 40 and 45, obtain the solution MALE*CLASS*ABILITY. Once the crossover reaches 50 then, in place of having to be 'highly able' *and* male *and* from high class origins to reach quasi-sufficiency, it is good enough to be '(more) highly able' and either male *or* from high social class origins.

Table 4. Varying thresholds.

Overall solution	Frequency	Thresholds	Mean consistency	Mean coverage
ABILITY + CLASS	3	0.70–0.72	.698	.867
ABILITY + MALE*CLASS	3	0.73–0.75	.730	.821
CLASS*ABILITY + MALE*ABILITY + MALE*CLASS	1	0.76	.768	.710
CLASS*ABILITY + MALE*ABILITY	7	0.77–0.83	.812	.639
MALE*CLASS*ABILITY	4	0.84–0.87	.879	.254
No solution	13	0.88–1.0		

Table 5. Pattern of solutions and non-solutions by threshold and crossover (155 QCAs using the original ability measure).

	Crossover (across)				
Threshold	30	35	40	45	50
.70	CLASS + MALE*ABILITY	CLASS + MALE*ABILITY	ABILITY + CLASS	ABILITY + CLASS	ABILITY + CLASS
.71	CLASS + MALE*ABILITY	CLASS + MALE*ABILITY	ABILITY + CLASS	ABILITY + CLASS	ABILITY + CLASS
.72	CLASS + MALE*ABILITY	CLASS + MALE*ABILITY	ABILITY + CLASS	ABILITY + CLASS	ABILITY + CLASS
.73	CLASS + MALE*ABILITY	CLASS + MALE*ABILITY	CLASS*ABILITY + MALE*ABILITY + MALE*CLASS	ABILITY + MALE*CLASS	ABILITY + MALE*CLASS
.74	CLASS + MALE*ABILITY	CLASS*ABILITY + MALE*ABILITY + MALE*CLASS	CLASS*ABILITY + MALE*ABILITY + MALE*CLASS	ABILITY + MALE*CLASS	ABILITY + MALE*CLASS
.75	CLASS	CLASS*ABILITY + MALE*ABILITY + MALE*CLASS	CLASS*ABILITY + MALE*ABILITY + MALE*CLASS	ABILITY + MALE*CLASS	ABILITY + MALE*CLASS
.76	CLASS*ABILITY + MALE*CLASS	CLASS*ABILITY + MALE*ABILITY + MALE*CLASS	CLASS*ABILITY + MALE*ABILITY + MALE*CLASS	CLASS*ABILITY + MALE*ABILITY + MALE*CLASS	ABILITY
.77	CLASS*ABILITY + MALE*CLASS	CLASS*ABILITY + MALE*ABILITY + MALE*CLASS	CLASS*ABILITY + MALE*ABILITY	CLASS*ABILITY + MALE*ABILITY	ABILITY
.78	CLASS*ABILITY + MALE*CLASS	CLASS*ABILITY	CLASS*ABILITY + MALE*ABILITY	CLASS*ABILITY + MALE*ABILITY	ABILITY
.79	MALE*CLASS*ABILITY	CLASS*ABILITY	CLASS*ABILITY + MALE*ABILITY	CLASS*ABILITY + MALE*ABILITY	CLASS*ABILITY + MALE*ABILITY
.80	MALE*CLASS*ABILITY	MALE*CLASS*ABILITY	CLASS*ABILITY + MALE*ABILITY	CLASS*ABILITY + MALE*ABILITY	CLASS*ABILITY + MALE*ABILITY

(*Continued*)

Table 5. (*Continued*).

	Crossover (across)				
Threshold	30	35	40	45	50
.81	MALE*CLASS*ABILITY	MALE*CLASS*ABILITY	CLASS*ABILITY	CLASS*ABILITY + MALE*ABILITY	CLASS*ABILITY + MALE*ABILITY
.82	MALE*CLASS*ABILITY	MALE*CLASS*ABILITY	MALE*CLASS*ABILITY	CLASS*ABILITY + MALE*ABILITY	CLASS*ABILITY + MALE*ABILITY
.83	MALE*CLASS*ABILITY	MALE*CLASS*ABILITY	MALE*CLASS*ABILITY	CLASS*ABILITY + MALE*ABILITY	CLASS*ABILITY + MALE*ABILITY
.84	**No solution**	**MALE*CLASS*ABILITY**	**MALE*CLASS*ABILITY**	**MALE*CLASS*ABILITY**	**CLASS*ABILITY + MALE*ABILITY**
.85	No solution	No solution	MALE*CLASS*ABILITY	MALE*CLASS*ABILITY	CLASS*ABILITY + MALE*ABILITY
.86	No solution	No solution	MALE*CLASS*ABILITY	MALE*CLASS*ABILITY	MALE*CLASS*ABILITY
.87	No solution	No solution	No solution	MALE*CLASS*ABILITY	MALE*CLASS*ABILITY
.88	No solution	No solution	No solution	No solution	MALE*CLASS*ABILITY
.89	No solution	No solution	No solution	No solution	MALE*CLASS*ABILITY
0.90–1.0	No solution	No solution	No solution	No solution	No solution

Table 6. Distribution of solutions as threshold and crossover are varied (15655 QCAs using all 101 ability measures).

Solution	15655 set	Percent (%)	155 set	Percent (%)
ABILITY	201	1.28	3	1.94
ABILITY + CLASS	774	4.94	9	5.81
ABILITY + MALE*CLASS	574	3.67	6	3.87
CLASS	59	0.38	1	0.65
CLASS + MALE*ABILITY	1089	6.96	9	5.81
CLASS*ABILITY	574	3.67	3	1.94
CLASS*ABILITY + MALE*ABILITY	1648	10.53	18	11.61
CLASS*ABILITY + MALE*ABILITY + MALE*CLASS	826	5.28	9	5.81
CLASS*ABILITY + MALE*CLASS	355	2.27	3	1.94
MALE*ABILITY	9	0.06	0	0.00
MALE*CLASS	13	0.08	0	0.00
MALE*CLASS*ABILITY	2314	14.78	23	14.84
No solution	7219	46.11	71	45.81

Now, what about a 'best solution'? For the moment we continue to consider just the original ability measure, so that everything except the ability calibration and the threshold is taken as fixed, and choose from the 155 QCAs. Of these 155 (5×31) potential solutions 71 are non-solutions, leaving 84 to consider (see Table 5). Amongst these CLASS*ABILITY + MALE*ABILITY appears 18 times (shaded cells), with a mean consistency of .817 and a mean coverage of .629. The mean crossover for these is 45.83 and the mean threshold .804. The mean PRI consistency is .695. No other solution offers as much. MALE*CLASS*ABILITY does appear more times (23) but, overall, its parameters are less good. Though it has a slightly better mean consistency of .862, its mean coverage is .270. Our candidate 'best solution' is, therefore, at this stage, the familiar (CLASS*ABILITY) + (MALE*ABILITY).

3.2. The complete set of 15655 analyses

We now consider whether this conclusion is supported by the full set of solutions, including the additional analyses of 100 error-affected ability variables. The overall solutions and their distribution for the whole set (15655) and the smaller set (155) are shown in Table 6.[14] The distributions are very similar, but not identical. In around 1 in a 1000 cases we obtain solutions amongst the 15655 that did not appear in the set of 155 (MALE*ABILITY; MALE*CLASS). The mean consistency for our potential 'best' solution (CLASS*ABILITY) + (MALE*ABILITY) is .813, its mean coverage .619, and its mean PRI .687. Again, MALE*CLASS*ABILITY appears more frequently, but with a much lower overall coverage. Moving from the 155 analyses to the full set of 15655 leaves the most plausible 'best solution' as CLASS*ABILITY + MALE*ABILITY.[15]

4. Bootstrapping/resampling

Here we employ a method developed by Efron and Tibshirani (1993) to explore the stability of QCA solutions. Efron's non-parametric bootstrap is a well-known procedure (for an introduction, see Shalizi, 2010). One assumes that the available sample

of data contains all the available information about the underlying population from which it is drawn.[16] A series of samples *with replacement* is taken from the available single sample. For each of these the parameter of interest is calculated and then, usually, the standard deviation of the distribution of the parameter of interest is used to establish this parameter's confidence intervals. Where one is interested in establishing confidence intervals for a parameter such as the mean the procedure is straightforward and many statistical packages incorporate the optional calculation of bootstrapped estimates. However, a complication arises in the context of QCA. Here, the nature of the solutions themselves may vary with resampling.[17] A 'causal pathway' might appear in some solutions but not others. It is, of course, exactly this possibility – that QCA solutions are unstable under resampling – that led us to employ bootstrapping here. We therefore focus on using the resampling approach to explore the stability of the *form* of solutions under resampling.

Because QCA minimises those rows of truth tables passing a consistency threshold, we will consider whether the set of rows passing the threshold changes under repeated resampling *with replacement* from our 6666 cases, initially taking samples of size 6666 to match our case numbers, as is usually recommended in the literature. We use the original ability measure for this work, with a calibration crossover of 45 and a consistency threshold of .8 (thus returning to the parameters employed in our initial solution). We carry out this resampling 1000 times. Table 7 shows the descriptive statistics for consistency for the eight rows for 1000 runs.[18] Looking at the minima and maxima, it can be seen that the three rows 111, 101 and 011 (shown in bold) pass the .8 threshold in each of the 1000 runs. No other row passes the threshold in any run. The solution (CLASS*ABILITY) + (MALE*ABILITY) is therefore stable under this exercise.

Initially we took repeated samples of the same size (6666) as our empirical sample. It is, however, instructive to explore the behaviour of the solutions when smaller samples are taken. We first take 1000 samples of size 1000. Here, in a minority of samples, some rows whose consistency in Table 1 did not exceed .8 become consistent at this level, and some whose consistency did exceed .8 now fall below this level. As a result, not all 1000 runs exhibit the same solution (see Table 8). However, the solution (CLASS*ABILITY) + (MALE*ABILITY) dominates the results, appearing in 91% of the analyses.[19] Looking at this from a different perspective, we can see that there will be some sample size, *for this dataset*, where the proportion of solutions that are (CLASS*ABILITY) + (MALE*ABILITY) rises above a 95% conventional confidence level. This seems to provide us with a possible way of applying a statistical approach to the *form* of the solution. If we take samples

Table 7. Descriptive statistics from 1000 samples *with replacement* of size 6666.

MALE	CLASS	ABILITY	Minimum	Maximum	Mean	Std. deviation
0	0	0	.506	.561	.531	.0087
0	0	1	.730	.773	.753	.0068
0	1	0	.693	.754	.725	.0092
0	**1**	**1**	.815	.850	.834	.0055
1	0	0	.564	.617	.593	.0089
1	**0**	**1**	.809	.850	.830	.0065
1	1	0	.731	.786	.761	.0088
1	**1**	**1**	.860	.896	.879	.0051

Table 8. Distribution of solutions under resampling (sampling $n = 1000$).

Overall Solution	Frequency	Percent
ABILITY	3	.3
CLASS*ABILITY	33	3.3
CLASS*ABILITY + MALE*ABILITY	912	91.2
CLASS*ABILITY + MALE*ABILITY + MALE*CLASS	40	4.0
MALE*ABILITY	10	1.0
MALE*ABILITY + male*CLASS	1	.1
MALE*CLASS*ABILITY	1	.1
Total	1000	100.0

Note: The "male" in lower case refers to the negation of the set "MALE", i.e. to females.

with replacement of size 1500, we obtain the solution (CLASS*ABILITY) + (MALE*ABILITY) in 97% of the 1000 analyses. The required sample size here to achieve a result over 95% is clearly between 1000 and 1500.[20]

5. Discussion

We aimed to illustrate some methods for addressing the robustness of large n set-theoretic analyses. Our initial QCA solution, generated against the background of the calibrations of sex, class and achievement used in Cooper (2005), entered a plausible set of calibration points for 'ability' into Ragin's 'direct method' of calibration and used the often employed threshold for quasi-sufficiency of .8. We explored the sensitivity of this solution to changes of calibration, to changes in the threshold that allows configurations to pass from the truth table into the minimised solution, and to measurement error. We also used a bootstrapping procedure to explore the stability of our initial solution under random sampling. The overall results suggest that, at least for the model and dataset under examination, the solution that would have been produced by a fairly standard application of fsQCA to our dataset, notwithstanding the problems of calibrating sets in a large n context where we have little case knowledge, would have been a good one to accept.[21] However, there are some issues that differentiate our work from earlier robustness studies that require discussion.

First, as we noted in discussing the literature on robustness, much earlier work addressed studies in which the cases/variables ratio was very low. Schneider and Wagemann (2012), for example, re-analysed a study with 19 cases and 6 conditions. With 6 conditions a truth table will have 64 (2^6) rows, and clearly, with only 19 cases, many will be either empty or have just one or two cases. In such cases, drop-one-case sensitivity tests, via their effect on which rows of the truth table remain non-empty, will often produce changes in minimised solutions. In our case, with a ratio of 6666 to 3, and no rows with no or few cases, such an approach can be expected to have no dramatic effects on a solution. For this reason, we chose to undertake the resampling exercise which, for us, seems, for those analysing samples rather than small entire populations, to provide a large n alternative to the drop-one-case test employed in earlier studies. We believe we have shown this to be a potentially useful approach deserving of more development in the large n QCA context.[22]

Second, in comparison with others, we have allowed both (i) our key anchor point for calibrating the ability set and (ii) our consistency thresholds to vary over a wider range. Skaaning (2011) argues that, during robustness analyses, changes in

calibration should be small enough so as not to undermine the theoretical arguments underpinning the original choice of anchor points. This is a sound argument but, in the case of ability, as we argued earlier, it does not rule out the use of a wide range of anchor points. For us, there is a complex relation between varying anchor points to test the robustness of solutions and varying them to answer a range of theoretical questions (Glaesser & Cooper, 2014). For the analysis of some outcomes, one .5 anchor point (creating the set 'moderately high ability') might be fruitful but, for some other outcome, or some other societal setting, a second anchor point might lead to more meaningful solutions. A related argument can be applied to our use of consistency thresholds from .70 to .99. Given our substantive interests, we do not only want to know how solutions change as small changes are made in these thresholds (the robustness focus), but also how the solutions vary as we make larger changes, making quasi-sufficiency easier or harder to achieve. A table like our Table 5 can be used for both these purposes.

The main issues not addressed in this paper are (i) the potential systematic bias resulting from differential attrition (by type of case) from longitudinal studies, (ii) the concerns arising from the claim that QCA is more liable than correlational approaches to incorporate random variables into a solution (Krogslund, Choi, & Poertner, 2013), and (iii) the consequences of changing the membership function used in the calibration of fuzzy sets. Concerning (i), we have discussed elsewhere how the weights of types of cases can affect consistency and coverage (Cooper & Glaesser, 2015). Any differential drop-out from a cohort will, via its effects on case weights, change these summary measures. Linking the attrition/weights issue with the bootstrapping approach employed in this paper might be a valuable and informative exercise. Concerning (ii), it would be interesting to compare QCA's response to random variables with those of other more conventional approaches. Concerning (iii), Thiem (2014a) has recently reported some results concerning membership functions, focussing on their effects on coverage, but this is another area where more work would be very welcome.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work has been supported by the UK's ESRC.

Notes

1. As with most longitudinal surveys, there has been attrition. The issue of attrition in the context of set-theoretic analysis is one we intend to address in future work. We bracket it out here where our concern is mainly with the stability of a given solution. We simply use all cases where we have data on these variables.
2. Breen and Goldthorpe (1999) note that this test cannot be regarded as simply measuring 'innate' ability.
3. Consistency assesses the extent to which the evidence supports the claims that a condition, or a configuration of conditions, is sufficient for the outcome (Ragin, 2008). Coverage assesses the empirical importance of this route to the outcome.
4. Membership in fuzzy sets ranges from none (0) to full (1). At the crossover point of .5 cases are as much in as out of the set.

5. The techniques we use to explore the calibration of 'ability', and error in its measurement, could, in principle, be applied to our other conditions and our outcome. However, in the case of sex, there is an obvious way to allocate membership. In the case of social class, it could also be argued, certainly from within the European sociological tradition, that there are more constraints of an ontological kind to bear in mind than in the case of measured ability. As far as our outcome measure, highest qualifications achieved, is concerned, there might be an argument for taking its population distribution into account when calibrating it, given its partly 'positional' nature (Hirsch, 1977). We bracket out these considerations.
6. Highest qualification at age 33 orders qualifications into six categories, from 'no qualification' to 'degree or higher'. We have fuzzified it as follows (Cooper, 2005): No qualification: 0; CSE 2-5/NVQ1: .17; O Level/NVQ2: .42; A Level/NVQ3: .67; Higher qualification/NVQ4: .83; Degree/higher/NVQ5,6: 1.0 .
7. This variable is derived from the variable n2385 (father or father figure's occupation, coded to Socio-Economic Groups) taken at respondent's age 16. Following Breen and Goldthorpe (1999), and Heath and McDonald (1987), we derive an approximation to the Goldthorpe 7-class schema. The latter is fuzzified as follows (Cooper, 2005): Upper service class: 1, lower service class: .83, routine non-manual: .583, petty bourgeoisie: .583, supervisors etc.: .417, skilled manual: .17, semi- and unskilled manual: 0.
8. See Thiem (2014a) for a recent discussion of the direct method.
9. This threshold determines which configurations pass from the truth table into the minimised solution.
10. Using a threshold of .8 for quasi-sufficiency.
11. Equivalently, ABILITY*(MALE + CLASS) ⇒ACHIEVEMENT.
12. Of the 101 versions of ability, one (the original) will have real measurement error and the other 100 this plus our added simulated error.
13. Given a fully populated truth table of 8 rows, 6 of which are above the lowest threshold of .7 used in these solutions, there are no more than 6 possible solutions, and so the 18 discussed here will be distributed across no more than these 6 'possibles'.
14. We are assuming here that the 155 that employed the original ability measure themselves include real measurement error, which is why we continue to include these 155 here.
15. Running these analyses again, but ignoring the output generated using thresholds lower than .8, produces the same pattern of results.
16. Our sample of 6666 cases has been taken from a longitudinal study in which there has been attrition. We have taken the view of most scholars who employ these data that the nature of this attrition is not such that large biases are introduced into analyses. Also, our purpose here is methodological – to illustrate how bootstrapping can be used with QCA in the large n context. In future work we intend to address this issue of drop-out and its effects on large n QCAs.
17. A similar issue arises with regression estimates, of course. Under resampling, as significance levels change, an interaction term may be entered into some solutions but not others.
18. Given space constraints, we will not report parallel results for PRI here.
19. There are two main alternative solutions. CLASS*ABILITY + MALE*ABILITY + MALE*CLASS appears when the consistency for row 7 in Table 7 creeps above .8, and CLASS*ABILITY when the consistency for row 6 falls below it.
20. We should also note, re possible limited diversity, that the lowest number of cases in any truth table row resulting from our samples of size 1000 is 44. For samples of 1500 it is 74. For samples of 6666 it is 387.
21. In this paper we have concentrated our attention on sufficiency analyses. However, the same issues characterise analyses of necessity. Given an already calibrated fuzzy set capturing levels of educational achievement, 'ability' calibrated with a high crossover will be much less likely to be considered as necessary for achievement than 'ability' calibrated with a lower crossover value. Assessments of necessity will also vary with the threshold chosen and as measurement error changes. We see no reason therefore why the techniques used here shouldn't be applied to assessments of necessity.
22. Thiem's (2014b) concerns about Hug's (2013) use of simulations to assess QCA are also relevant here.

Notes on contributors

Judith Glaesser is a senior lecturer in the School of Education at Durham University, UK. Her interests include sociology of education, comparative education, policy, research design and research methods, and Qualitative Comparative Analysis (QCA). She has held two Economic and Social Research Council (ESRC) grants which have enabled her to explore QCA, its foundations and applications. Her most recent book is *Young people's educational careers in England and Germany: Integrating survey and interview analysis via Qualitative Comparative Analysis*, Palgrave Macmillan.

Barry Cooper is a professor in the School of Education at Durham University, UK, where he was (1998–2005) the director of Research in Education. From 2004–2007, he co-edited the *British Educational Research Journal*. His interests are in the sociology of education, set-theoretic research methods and the evaluation of educational aid projects. His key publications include, with Máiréad Dunne, *Assessing Children's Mathematical Knowledge: Social class, sex and problem-solving*, and, recently, with Glaesser, Gomm and Hammersley, *Challenging the Qualitative-Quantitative Divide: Explorations in Case-focused Causal Analysis* (Continuum). He currently works as co-investigator with Glaesser on an ESRC-funded project on Qualitative Comparative Analysis.

References

Breen, R., & Goldthorpe, J. H. (1999). Class inequality and meritocracy: A critique of Saunders and an alternative analysis. *The British Journal of Sociology, 50*, 1–27.

Cooper, B. (2005). Applying Ragin's crisp and fuzzy set QCA to large datasets: Social class and educational achievement in the National Child Development Study. *Sociological Research Online, 10*. Retrieved from http://www.socresonline.org.uk/10/2/cooper1.html

Cooper, B., & Glaesser, J. (2008). How has educational expansion changed the necessary and sufficient conditions for achieving professional, managerial and technical class positions in Britain? A configurational analysis. *Sociological Research Online, 13*. Retrieved from http://www.socresonline.org.uk/13/3/2.html

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*, 4–23.

Cooper, B., & Glaesser, J. (2011a). Using case-based approaches to analyse large datasets: A comparison of Ragin's fsQCA and fuzzy cluster analysis. *International Journal of Social Research Methodology, 14*, 31–48.

Cooper, B., & Glaesser, J. (2011b). Paradoxes and pitfalls in using fuzzy set QCA: Illustrations from a critical review of a study of educational inequality. *Sociological Research Online, 16*. Retrieved from http://www.socresonline.org.uk/16/3/8.html

Cooper, B., & Glaesser, J. (2012a). Set theoretic versus correlational methods: The case of ability and educational achievement. In B. Cooper, J. Glaesser, R. Gomm, & M. Hammersley, *Challenging the Qualitative-Quantitative Divide: Explorations in case-focused causal analysis* (pp. 170–207). London: Continuum.

Cooper, B., & Glaesser, J. (2012b). Qualitative work and the testing and development of theory: Lessons from a study combining cross-case and within-case analysis via Ragin's QCA. *Forum: Qualitative Social Research, 13*. Art. 4.

Cooper, B., & Glaesser, J. (2015). Analysing necessity and sufficiency with Qualitative Comparative Analysis: How do results vary as case weights change? *Quality & Quantity.* Retrieved from http://link.springer.com/article/10.1007/s11135-014-0151-3

Duşa, A., & Thiem, A. (2013). *QCA: Qualitative Comparative Analysis*. R package version 1.0-5.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Fiss, P. (2011). Building better causal theories: A fuzzy set approach to typologies in organization research. *Academy of Management Journal, 54*, 393–420.

Glaesser, J. (2008). Just how flexible is the German selective secondary school system? A configurational analysis. *International Journal of Research & Method in Education, 31*, 193–209.

Glaesser, J., & Cooper, B. (2014). Exploring the consequences of a recalibration of causal conditions when assessing sufficiency with fuzzy set QCA. *International Journal of Social Research Methodology, 17*, 387–401.

Greckhamer, T., Misangyi, V. F., Elms, H., & Lacey, R. (2008). Using Qualitative Comparative Analysis in strategic management research: An examination of combinations of industry, corporate, and business-unit effects. *Organizational Research Methods, 11*, 695–726.

Greckhamer, T., Misangyi, V. F., & Fiss, P. (2013). The two QCAs: From a small-n to a large-n set theoretic approach. *Research in the Sociology of Organizations, 38*, 49–75.

Heath, A. F., & McDonald, S.-K. (1987). Social change and the future of the left. *The Political Quarterly, 53*, 364–377.

Hirsch, F. (1977). *The social limits to growth*. London: Routledge & Kegan Paul.

Hug, S. (2013). Qualitative Comparative Analysis: How inductive use and measurement error lead to problematic inference. *Political Analysis, 21*, 252–265.

Krogslund, C., Choi, D. D., & Poertner, M. (2013, April). Fuzzy sets, shaky ground: Testing calibration, measurement, and specification sensitivity in fsQCA. *APSA Annual Meeting Paper*. Chicago, IL, USA.

Krogslund, C., & Michel, K. (2014). A larger-N, fewer variables problem? The counterintuitive sensitivity of QCA. *Qualitative & Multi-Method Research, 14*, 25–33.

Lucas, S.R., & Szatrowski, A. (2014). Qualitative Comparative Analysis in critical perspective. *Sociological Methodology, 44*, 1–79.

Marx, A., & Duşa, A. (2011). Crisp-set Qualitative Comparative Analysis (csQCA), contradictions and consistency benchmarks for model specification. *Methodological Innovations Online, 6*, 103–148.

Ragin, C. C. (2006). The limitations of net effects thinking. In B. Rihoux & H. Grimm (Eds.), *Innovative comparative methods for policy analysis* (pp. 13–41). New York, NY: Springer.

Ragin, C. C. (2008). *Redesigning social inquiry* (pp. 190–212). Chicago, IL: Chicago University Press.

Ragin, C. C., & Fiss, P. (2008). Net effects versus configurations: An empirical demonstration. In C. C. Ragin, *Redesigning social inquiry* (pp. 190–212). Chicago, IL: Chicago University Press.

Schneider, C. Q., & Wagemann, C. (2012). *Set-theoretic methods for the social sciences*. Cambridge: Cambridge University Press.

Seawright, J. (2005). Qualitative Comparative Analysis vis-à-vis regression. *Studies in Comparative International Development, 40*, 3–26.

Shalizi, C. (2010). The bootstrap. *American Scientist, 98*, 186–190.

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

Thiem, A. (2014a). Membership function sensitivity of descriptive statistics in fuzzy-set relations. *International Journal of Social Research Methodology, 17*, 625–642.

Thiem, A. (2014b). Mill's methods, induction, and case sensitivity in Qualitative Comparative Analysis: A comment on Hug. *Qualitative & Multi-Method Research, 12*, 19–24.

Thiem, A., & Duşa, A. (2013). *Qualitative Comparative Analysis with R*. New York, NY: Springer.