Dealing with Errors in QCA
Author(s): Martino Maggetti and David Levi-Faur
Source: *Political Research Quarterly*, Vol. 66, No. 1 (MARCH 2013), pp. 198-204
Published by: Sage Publications, Inc. on behalf of the University of Utah
Stable URL: http://www.jstor.org/stable/23563603

before using dummy variables created according to this typology to account for contextual effects in a regression performed at lower level of analysis. Dotti Sani and Quaranta (2011) use such an approach to study the work-motherhood relation in Organisation for Economic Co-operation and Development (OECD) countries. Focusing on formally accounting for measurement error, Eliason and Stryker (2009), in another approach, do use goodness-of-fit tests to qualify the fit of fuzzy-set conditions and thereby to adapt fsQCA results to inferential logic based on falsification.

3. Maggetti and Levi-Faur (this issue) discuss strategies for dealing with potential measurement error in QCA, whereas Emmenegger, Kvist, and Skaaning (this issue) review comparative welfare-state research using QCA and find that not all studies carried out robustness checks of their findings.
4. Although the interested reader is encouraged to read these articles for a formal treatment of optimization on lattices and for the relevant mathematical proofs, we shall focus our discussion of this approach on the results necessary for its econometric operationalization.
5. The potential connection of lattice theory with (fs)QCA is also suggested by Zaytsev et al. (2012), who combine QCA with formal concept analysis (FCA) based on lattice theory to address problems of measurement in democracy studies.
6. The regression produces estimates for all four betas as it does not include a constant.
7. If V is continuous, then this result suggests that A is a necessary and sufficient condition for high V. If $\beta_3 \approx \beta_4 < \beta_1, \beta_2$ then this would suggest that A is a necessary and sufficient condition for low V.
8. If V is continuous, the above result would suggest equifinality with respect to a high V outcome. More generally for a continuous V, if the difference between two or more estimated betas is not statistically significant, these configurations are equifinal as they are mutually exclusive yet associated with the same value of the outcome variable.
9. See Mohnen and Röller (2005) for details of the test statistic and of the inequality-constrained minimization problem used to calculate it.

Dealing with Errors in QCA

Martino Maggetti[1,2] and David Levi-Faur[3,4]

DOI: 10.1177/1065912912468269f

Abstract

This paper discusses five strategies to deal with five types of errors in Qualitative Comparative Analysis (QCA): condition errors, systematic errors, random errors, calibration errors, and deviant case errors. These strategies are the comparative inspection of complex, intermediary, and parsimonious solutions; the use of an adjustment factor, the use of probabilistic criteria, the test of the robustness of calibration parameters, and the use of a frequency threshold for observed combinations of conditions. The strategies are systematically reviewed, assessed, and evaluated as regards their applicability, advantages, limitations, and complementarities.

Introduction: Errors and Criticisms of QCA

Strategies to deal with the possibility of error are essential tools in all types of social research. The challenge of error management can be broadly conceived as the challenge of forming a bridge between theory and empirical research in a world where some imprecision, uncertainty, and randomness is unavoidable. Any research study in the social sciences must contend with error, stemming from a variety of sources, including incomplete definitions of the constructs being measured, imperfect operationalization of the ideas contained in the corresponding concepts, and weaknesses of methods of assessment. This holds of course also for QCA, that, some argue, has limited capacity to deal with different types of errors that are commonplace in the social sciences. As QCA methods typically work under deterministic or quasi-deterministic assumptions, standard statistical techniques that are used to correct and minimize measurement error and other types of error do not apply. The researcher cannot straightforwardly

[1]University of Zurich, Switzerland
[2]University of Lausanne, Switzerland
[3]Hebrew University of Jerusalem, Israel
[4]Freie Universität Berlin, Germany

estimate the error term because QCA does not aim to measure the size of the "net effect" of independent variables on a dependent variable. QCA vulnerability to error is said to be even bigger than other methodologies because each of the multiple causal combinations constituting sufficient conditions is analytically relevant to the outcome. In crisp-set analysis (csQCA), case-sensitiveness and variable-sensitiveness are especially high because the inaccurate coding of fundamental set properties such as "presence" and "absence" may radically alter the results. Therefore, a single "wrong" coding of a case can falsify a theory that would otherwise be corroborated.

Fuzzy-set analysis (fsQCA) has advantages over csQCA regarding the treatment of errors. The main advantage is that conditions and outcomes are no longer binary but they can be calibrated according to different "degrees of membership" in sets. The negative consequences of coding errors are thus reduced. This allows the researcher a greater precision in the operationalization of conditions, and it increases the analytical leverage of QCA. Therefore, the use of fsQCA is usually recommended to increase measurement accuracy and diminish potential biases. Nonetheless, as Hug (2012) pointed out, the treatment of error is not often explicitly addressed in empirical studies. In particular, two enduring sets of criticisms apply to fsQCA, too. On one hand, fuzzy-set analyses are exposed to distortions due to the occurrence of errors related to the selection, operationalization, and coding of cases and conditions. On the other hand, their alleged inability to distinguish randomly assigned values from real data is frequently mentioned, in connection with the issue of their high degree of sensitivity to model specifications. In addition, although the problem of error is mitigated when the researchers possess a substantive knowledge of each case and when the number of observations is small, this does not apply to larger data sets that are typically used for fsQCA, when researchers lose the intimate knowledge of individual cases and thus the probability of measurement error increases.

In this context, some researchers have developed and applied different strategies to deal with errors in QCA, which remain, however, at present quite dispersed. The goal of this contribution is to systematize, review, and assess these strategies. The next section discusses and evaluates five strategies to deal with five main errors in QCA. They are applied to fsQCA as the most general formulation of QCA, but, with the usual restrictions, they can be considered valid for other types of QCA analyses. The conclusion follows.

Overview and Assessment of Five Strategies to Deal with Error in QCA

We focus on five types of error in QCA: condition errors, systematic errors, random errors, calibration errors, and deviant case errors, which refer, respectively, to the sensitivity of conditions, the inaccuracy of observation devices, unpredictable factors, the (miss)specification of the parameters of calibration, and sensitivity to one or more flawed cases. The strategies that we discuss to deal with them are the comparative inspection of complex, intermediate, and parsimonious solutions; the use of an adjustment factor; the use of probabilistic criteria; the test of the robustness of calibration parameters; and the use of a frequency threshold for observed combinations in the truth table.

a. The comparative inspection of complex, intermediate, and parsimonious solutions to deal with condition errors

Measurement errors may be related to the operationalization of one or more specific conditions. The strategy for handling this type of error in fuzzy-set analysis relies on the inspection of the complex, intermediate, and parsimonious solutions (as defined below). As a preliminary to the analysis, the ratio of selected conditions to cases should be below a certain threshold. Marx (2010) suggests a ratio of conditions to cases ranging from 0.33 for small-medium-*N* to 0.20 for medium-large-*N* and an upper limit of seven or eight conditions to the absolute number of conditions to be included in crisp-set analysis, after which results are less reliable due to the emergence of many unique causal paths. These recommendations are even more valid for fsQCA, for which the problems of "uniqueness" (when configurations consist of a single case) and "limited diversity" (too many nonobserved configurations) persist, whereas their consequences become less visible.

Then, as a first step, the consistency and coverage of solutions should be computed to assess the quality and empirical relevance of the set relations identified with the fuzzy-set analysis. Consistency and coverage allow the researcher to assess and report how closely the set relation is approximated (i.e., the degree to which the cases sharing a given combination of conditions agrees in displaying the outcome) and the empirical relevance of consistent subsets (i.e., the proportion of cases following a specific path). Therefore, the measures of consistency and coverage are useful to make sense of imperfectly consistent set relations, which are prevalent in the social sciences. A widely used consistency level for sufficiency is 0.80, referring to the proportion of consistent membership scores in a causal condition (or combination of causal conditions) on all membership scores in a condition (or combination of conditions). The assessment of necessity typically requires more stringent criteria, for example, a consistency score above 0.90 or 0.95. It is worth remembering that consistency thresholds should not be applied in a mechanical way, but must be adapted to research goals, levels of analysis, data quality, and number of cases. For instance, exploratory analysis requires lower consistency than rigorous hypothesis testing.

As a second step, consistent solutions with satisfactory coverage can be retained for further examination and interpretation. Following "QCA best practices," the analysis of sufficient conditions should always be performed with and without simplifying assumptions regarding the logical remainders, and all solution formulas should be reported. The parsimonious solution is based on simplifying assumptions on all logical remainders, the intermediate solution is based on theoretically meaningful simplifying assumptions (easy counterfactuals) and the complex solution does not assume any simplifying assumption. All formulas are logically true as they are based on empirical information contained in the truth table but they differ in their degree of precision. Intermediate solutions are expected to be subsets of parsimonious solutions and supersets of complex solutions. In that regard, the comparison of complex, intermediate, and parsimonious solution formulas can provide information on which conditions may be particularly sensitive to error. Conditions that display counterintuitive and theoretically incoherent patterns should be carefully reconsidered, in particular when they are inconsistent along the minimization procedure from the complex to the intermediate solution, which implies the use of theory-backed logical remainders. Conversely, conditions that are part of both parsimonious and intermediate solutions can be considered as "core elements" where the evidence indicates a "strong" relationship with the outcome (Fiss 2011).

As an example of the comparative inspection of solutions, Greckhamer investigated combinations of cultural and environmental attributes associated with differences in compensation level and compensation inequality with fsQCA. He analyzed country-level data encompassing four occupational groups (cleaners, secretaries, mid-level managers, and senior managers) from forty-four countries. To inspect the property space related to the configurations displayed in the truth table, he distinguished between "core" and "complementary" conditions, that is, conditions that are part of both parsimonious and intermediate solutions, and, respectively, conditions that only occur in intermediate solutions. This way, it was possible to focus on those causal recipes that are more stable and accurate, showing that alternative combinations of culture, national development, and welfare-state shape compensation for the four occupations differently, accounting more for variations in low and intermediary occupational levels than for the highest managerial positions.

To summarize this issue, the comparative inspection of complex, intermediate, and parsimonious solutions allows conditions that are more sensitive to error to be detected. This procedure is important for interpreting the findings of any fuzzy-set analysis, especially when theoretical expectations are ambiguous as regards the role of one or more explanatory conditions. Conditions that are uniformly present across solutions should be regarded as core elements. In particular, the parsimonious solution usually provides a simplified formula that is less precise but more stable, because the remaining conditions are present even when all logically possible simplifying assumptions are included in the minimization procedure.

b. The use of an adjustment factor to deal with systematic errors

Systematic errors typically stem from some approximation in the accuracy of measurement due to imperfect tools, circumstances, and methods of observation. When this type of error is suspected from the inspection of the truth table or from the fuzzy plot, it is possible to use an adjustment factor to relax the parameters for necessity or sufficiency (see Figure 1). The original goal of the adjustment factor—to enlarge the applicability of the analysis of subset relations—is now incorporated in consistency and coverage measures. However, the adjustment factor can still be applied before calculating consistency scores in the case of suspected systematic error. This operation will allow more accurate consistency measures concerning the "net quality" of the set relation to be obtained.

With a standard adjustment factor of 0.10, cases are considered as positive instances in the outcome, not only when they configure a perfect subset relation but also when their membership in the causal condition does not exceed an adjustment factor representing 0.10 additional fuzzy-membership points. To qualify conditions that are affected by the adjustment factor, the notions of quasi-necessity and quasi-sufficiency have been introduced. These notions signify that adjusted causal combinations are "almost always" sufficient or necessary for 90 percent of the cases where the causal combination applies. The implications of such an adjustment factor may be presented and reported in a plot showing the distribution of cases for a hypothetical causal condition. All cases under the diagonal—a straight line going from corner to corner—indicate necessity whereas all cases above the diagonal indicate sufficiency (see Figure 1). A fuzzy adjustment of 0.10 raises the diagonal 0.10 points above or below its normal position so that more points are consistent with the adjusted diagonal.

The adjustment factor is quite common in empirical research (or at least it was before the popularizing of consistency and coverage scores). For instance, Raunio (2005) used an adjustment factor of 0.17 on a seven-value scale in his study of factors explaining cross-national variation in the level of parliamentary scrutiny of governments in European affairs. Considering that the assignment of the membership scores was quite imprecise, particularly in the middle range, this adjustment factor allowed those instances that exceeded the set-membership scores by one step on the fuzzy scale to be counted as positive cases. This way, it was possible to find a single necessary

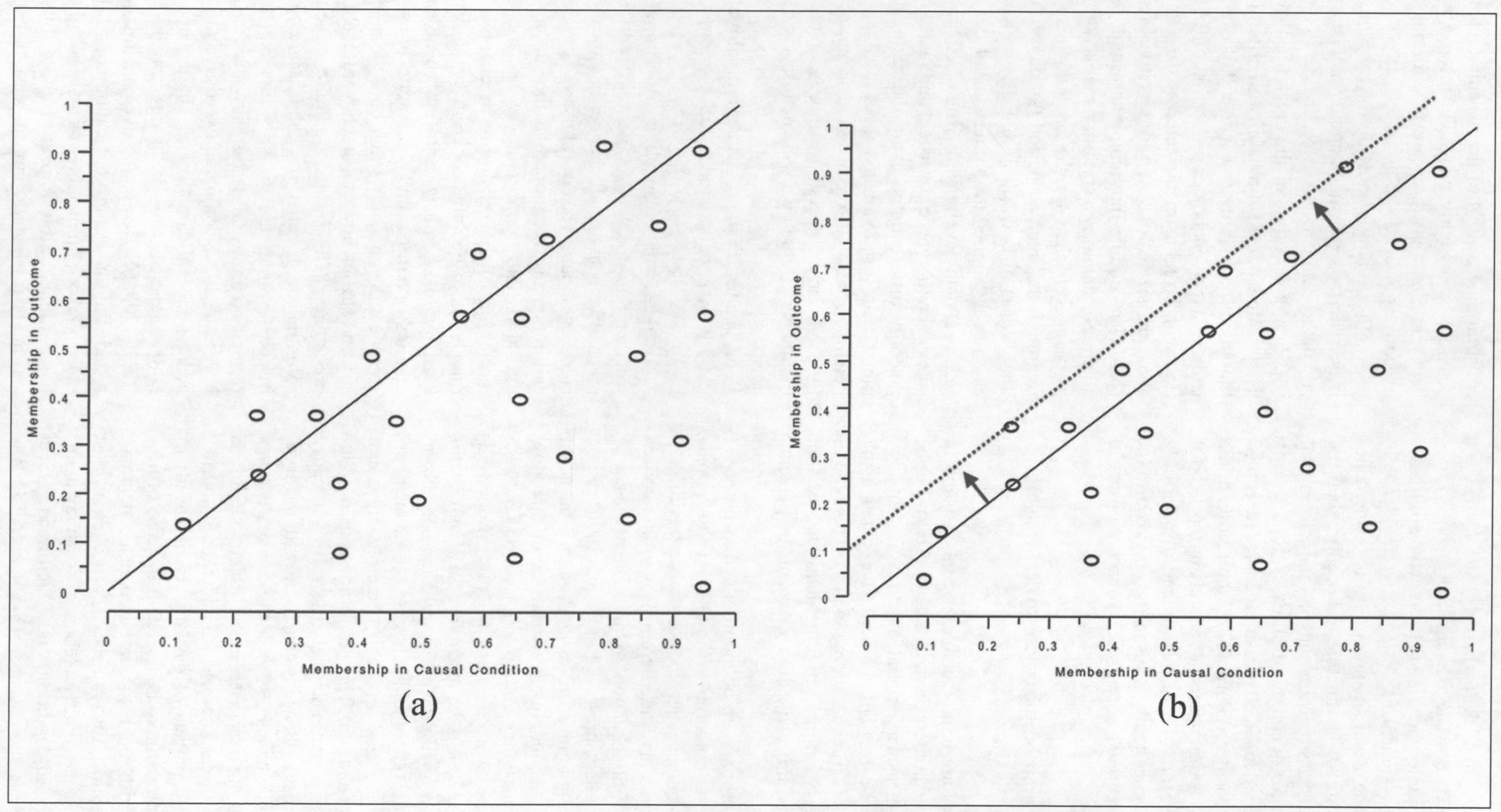

Figure 1. Adjustment factor: (a) test of necessity without adjustment factor, (b) test of necessity with adjustment factor.

condition (the strength of the parliament) and a sufficient combination (a powerful legislature and a more Euroskeptical public opinion) leading to the outcome of tighter scrutiny of the government.

To conclude, the use of an appropriate adjustment factor (e.g., 0.05 or 0.10) is valuable, especially when the researcher expects a small amount of imprecision that may concern the fuzzy coding of all cases, representing the tolerance interval in which measurements are accepted before they are considered flawed. The application of the adjustment factor is quite transparent and appropriate for small-*N* analyses, too, but we recommend using it only when this type of uncertainty is suspected and to do so for an interval representing a very limited subset of observed cases (e.g., 10%).

c. The use of probabilistic criteria to deal with random errors

Random errors are unpredictable and inconsistent in their magnitude or in their direction. They are said to be unavoidable in measuring social phenomena, but they are expected to be particularly pervasive when using some research tools, such as survey inquiry. This type of error is indeed very common in large data sets based on questions that require estimation or relate to personal attitudes.

Probabilistic criteria can be incorporated in fuzzy-set analysis to address "randomness" in QCA. They are particularly recommended when the researcher has limited knowledge of cases or limited control over data collection and data operationalization. In this context, the use of probabilistic criteria as presented in Ragin (2000) can be useful to assess not only the proportion of consistent instances against a predefined benchmark (a task now performed by consistency and coverage scores) but also the significance levels related to this proportion. Following Ragin (2000), the researcher must first evaluate the proportion of cases with nonzero membership in the set defining the outcome that have outcome membership scores that are less or equal to their membership scores in a given condition. This proportion is then evaluated in relation to a predefined benchmark and significance level to see whether evidence is consistent with the argument of sufficiency or necessity. For instance, Pennings (2003) examined the necessary and sufficient conditions for high constitutional control in democratic countries, that is, the formal powers of parliaments and heads of state to constrain executive behavior. In the case of necessary conditions, the analysis was based on a benchmark proportion of "almost always necessary" (when at least 80 percent of instances of the outcome also display membership in the causal condition) and a significance level of 0.05. As a result, the absence of presidentialism was found as a necessary condition for a high level of constitutional control of the executive. This finding was considered as highly plausible because a strong presidency implies that there are few formal constraints on executive power.

Another example showing the usefulness of probabilistic criteria is Herrmann and Cronqvist's (2009) analysis of the Vanhanen data set to identify necessary and sufficient conditions for the breakdown of democracies. They implemented probabilistic criteria by using a binominal probability test with a 0.05 significance level and a relaxed benchmark proportion of 0.65, in combination with an adjustment factor of 0.3, which represents about the size of one step in their five-step membership scale. This double fine-tuning appeared to be indispensable for discovering a condition that is both almost necessary and sufficient for the breakdown of a democratic regime in the interwar period (an uneven distribution of knowledge). It is worth noting that probabilistic criteria can also be applied to the measure of the consistency of fuzzy-set data developed by Ragin (2006b) mentioned in subsection (a) where consistent and inconsistent cases are not only differentiated using the main diagonal of the plot, but are also evaluated as regards their relative membership score, with credit for near misses and penalties for condition membership scores that exceed the outcome membership score by a wide margin.

d. The test of the robustness of calibration to deal with model misspecification errors

To deal with errors, it is also important to check whether the model is robust, that is, if results are not much affected by small changes in the calibration parameters. Two main types of robustness checks for fuzzy-set calibration are mentioned in the literature: calibration thresholds and membership scores. First, an extensive discussion of the sensitivity of changes in calibration thresholds, especially concerning the crossover points used for fuzzy values, is offered by Skaaning. With a textbook example, and 242 replications with different specifications, he shows that both the analysis of necessity and sufficiency are somehow affected by calibration procedures. The results are quite mixed, because the solutions are affected to a "nontrivial degree," but most of the formulas are virtually identical to the baseline result, and almost all terms show the expected direction. Maggetti (2012) compared two different calibration procedures for real data to assess the independence of regulatory agencies, using the "indirect method" described by Ragin (2006b). Two qualitative benchmark codings were tested; the one being very cautious and the other using more relaxed parameters. Results showed that this choice has a nonnegligible impact on the results of the fuzzy-set analysis; hence, careful, case-based, qualitatively informed calibration turned out to be very important.

Next, Eliason and Stryker developed a goodness-of-fit test for assessing the fit between evidence and causal hypotheses while accounting for measurement error in membership scores. Their strategy is based on the idea of comparing the observed distance of cases in a fuzzy-set graph from the diagonal representing perfect necessity and sufficiency with the distance that is expected should the underlying causal hypothesis be true. Cases just below or above the main diagonal should not be treated as very strong evidence against a set relationship, as this small distance may be due to imprecise measurement and coding, the fit of the data to expectations should comprise a certain degree of measurement error. An additive error in the membership score is thus assumed. The maximum value is set at the midpoint (0.5), which diminishes smoothly to the endpoints 0 and 1, because it seems plausible that the coding of membership scores is less certain at the maximum point of ambiguity and more certain when coding extreme membership scores.

Therefore, it is possible to check to what extent results are affected by changes in calibration thresholds and membership scores. In a further step, when the model is not robust enough, the researcher should identify the crucial assumptions underlying the results. It is advisable to try (at least) a conservative calibration and one with more relaxed parameters as regards anchors and the crossover point, to see how much they affect the solution.

e. The use of frequency thresholds to deal with deviant cases errors

When the researcher is not closely familiar with the cases, the occurrence of infrequent combinations of conditions might stem from measurement or coding error and cannot provide strong evidence of the set relations.

A way to deal with this type of measurement error, which is straightforward and suitable for a large data set, is to set a frequency threshold for combinations of conditions. In fact, with truth table analysis, it is possible to arrange all logically possible configurations that originate from a given set of causal conditions in the rows of a table to see which configurations have been empirically observed (corresponding to cases that have a membership score greater than 0.5 in each combination) and whether they are leading to positive or negative outcomes. Logically possible configurations lacking empirical instances—because these cases do not exist or because related information is inadequate—are treated as remainders and deleted from the table. Relevant configurations are included in the analysis instead. The usual rule for discriminating between relevant configurations and remainders is to differentiate between combinations of conditions presenting zero and at least one empirical instance. Thus, configurations lacking a single case with an adequate membership score are treated as reminders, while the combinations of conditions with at least one case are retained for the fuzzy-set analysis. However, especially when the number of cases is large (more than fifty), it is advisable to establish a frequency threshold greater than one for configurations. This way, rare configurations can be treated the same as those with no empirical substance.

For instance, following this approach, Glaesser (2008) studied factors influencing pupils' selection processes within the highly stratified German secondary school system. She applied csQCA to a large database of 1,014 individual cases. She established a frequency threshold for the relevance or viability of causal combinations "in order to ensure that the analysis is not based on such rows with very small-N which may be unduly influenced by measurement error." Therefore, only configurations that represent more than three cases were included in the Boolean analysis. The results underlined, among other factors, the importance of high marks in secondary school for predicting the attainment of higher qualifications.

To sum up, the issue of sensitivity to case-based errors can be addressed by excluding very rare configurations, that is, cases that fall below a certain frequency threshold in the truth table. This manipulation allows for dealing with a particular type of measurement error, that is, the uncertainty introduced by a few doubtful (or nonexistent) cases. In addition, it is worth noting that the exclusion of combinations of conditions below a reasonable threshold will possibly narrow down but not dramatically alter the analysis, because QCA results tend to remain stable when moving from the set of investigated cases to a particular subset (but they are likely to be more unstable when moving from the set to a superset, whereas the opposite is true for conventional statistical analysis). Therefore, when a frequency threshold is set, the precision of the fsQCA solution could be reduced but its validity is expected to be reinforced.

The use of frequency thresholds is quite common in empirical research but it rarely follows an explicit rationale. What is a reasonable threshold? This question is sometimes mentioned in the literature. In this regard, Ragin (2008, 133) says that

> important considerations include the total number of cases, the number of causal conditions, the degree of familiarity of the researcher with each case, the degree of precision that is possible in the calibration of fuzzy sets, the extent of measurement and assignment error, whether the researcher is interested in coarse versus fine-grained patterns in the results, and so on.

However, there are no systematic discussions of the procedure for setting an adequate frequency threshold. Here, we propose a very simple procedure. We need a conventional starting point to find a suitable threshold for excluding dubious cases that at the same time would not

Table 1. Summary of recommendations.

Problem	Definition	Solution	Relevance	Application
Condition errors	Errors related to one or more conditions	The comparative inspection of the solutions (the parsimonious, the intermediate, and the complex) to detect conditions that are more or less sensitive to measurement error	Especially when the theory is ambiguous as regards the role of one or more condition that affects the outcome	Solutions must be plausible and coherent, otherwise the operationalization of conditions has to be reconsidered; core conditions are less sensitive to errors
Systematic errors	Errors due to inaccuracy in coding (because of imperfect tools, circumstances, and, methods of observation)	The use of an adjustment factor	When the researcher expects a small amount of imprecision that may concern the fuzzy coding of all cases	The parameters for necessity or sufficiency can be relaxed to account for systematic approximation in measurement, also for small-*N*
Random errors	Errors that are unpredictable and inconsistent in their magnitude or direction (e.g., because of estimation and personal factors in surveys)	The use of probabilistic criteria	Especially when the researcher has limited knowledge of cases and/or limited control on data collection and data operationalization	It is possible to evaluate the number of consistent cases in front of a benchmark proportion while performing a test of significance
Model misspecification errors	Errors in the specification of the parameters of calibration (thresholds and membership scores)	The test of the robustness of calibration	Especially when there is no a priori guidance on the best model specifications	The researcher can check to what extent results are affected by changes in calibration parameters
Deviant case errors	Errors related to sensitivity to one or more flawed cases	The use of a frequency threshold	When there are very rare, unexpected, counterintuitive or not-theory-backed configurations in a large data set	Configurations that fall below a certain frequency threshold in the truth table are treated as logical remainders

affect substantive results. The minimal number of cases for a configuration is at least one. For big *N* fsQCA with more than 50 cases, we recommend at least two instances. In addition, using variable-oriented research as a benchmark for large-*N* studies, we propose ruling out configurations whose frequency is lower than the standard level of significance used to justify a claim of a statistically significant effect, that is, 0.05. A threshold of 0.01 can also be considered for fine-grained analyses.

After having identified the configurations whose frequency falls below this threshold, but it is greater than or equal to one, we recommend going back to the cases and deciding if this pattern is theoretically and empirically plausible or if it is necessary to recode the related conditions. Substantial knowledge is thus required, if not about the individual cases, certainly about the categories or classes in which specific cases are included. This dialogue between (theoretical) ideas and (empirical) evidence is the foundation of QCA. When it is needed, but impossible for technical or substantial reasons to recode, we recommend excluding these cases from the truth table analysis. The procedure should be repeated until all configurations are included, recoded, or deleted at the lowest-level cut-off point (one), below which configurations must anyhow be deleted. This principle of precaution reduces the sensitivity of the fsQCA analysis to a few flawed cases by making the most of the property of subset stability. Further research is required to determine whether this procedure alters the substantive results too much. The ideal test would be to perform repeated simulations on several random data sets to test the sensitivity of the analysis at different frequency thresholds.

Concluding Remarks

This contribution discussed the advantages and limitations of strategies to deal with different types of error in QCA as well as their complementarities. Table 1 presents a summary of the errors, solutions, their relevance, and possible applications.

The treatment of measurement errors, the improvement of validity, and the reporting of uncertainties are integral parts of any research enterprise, starting with research design and culminating with the publication and critical evaluation of results. We hope that by discussing some strategies for the handling of measurement errors, we may contribute to better methodological standards. Methodological best practices are however only the tools of research and it is advisable to avoid a mechanical application of these procedures and to pursue the constant dialogue between ideas and evidence that lies at the heart of Qualitative Comparative Analysis. This applies also to the strategies for dealing with errors and reporting procedures that we reviewed here.