Qualitative Comparative Analysis:
Social Science Applications and Methodological Challenges

January 15th-16th, 2015
Tilburg University, the Netherlands

Interesting results - but are they valid?

Alessia Damonte

Dept. of Social and Political Sciences
Università degli Studi di Milano

alessia.damonte@unimi.it

first draft
- please do not cite -

comments extremely welcome

Abstract. QCA's grasp on causation is often questioned from a probabilistic, experimental understanding of validity. QCA results however rely on logical and set-theoretical inferences. Is a difference in languages enough to justify a separate validity yardsticks? And what secures that QCA is delivering valid results?
The review of quantitative and qualitative exemplary yardsticks shows that traditions share validity concerns, yet give them different contents. The article argues that such difference is legitimized by the special assumptions about causation that inform their research processes. It therefore clarifies QCA causal ontology, identifies its special threats, and evaluates the strategies in use to prevent or tackle them - also adding a new one to address over-specified hypotheses. In this, the nomothetic yardstick proves to be a fertile framework, yet hardly a proper guideline for solutions.

Keywords. Causation, Explanation, Qualitative Comparative Analysis, Research paradigms, Validity.

1. Introduction

Validity is the property claimed by any research conclusion that aspires to be credible. The claim is usually grounded in evidence that the research process has avoided, or limited, the effects of possible biases, errors, and ambiguity on findings (Cook & Campbell 1979). Especially in the domain of «why question» research, such evidence is therefore crucial. Belief has it that, whatever the technique used to shape inferences, without evidence of validity no result can be accepted as «proven», «sound», or «true». Those from Qualitative Comparative Analysis (therefore, QCA) make no exception, as recent works have remarked (Tanner 2014). Also, whatever the technique, belief has it that evidence of validity can be achieved - if the research process sticks to special benchmarks.

How general these benchmarks are, however, it has long been the subject of a lively debate across disciplines and research traditions. The point is highly consequential, as unfitting benchmarks can impose inconsistent procedures which disrupt inferences instead of strengthening them. Nevertheless, it is still far from clear whether any yardstick can claim the title of «gold standard» - even within a same tradition.

The article addresses such puzzle by first looking for guidance in the yardsticks which have shaped the debate in the domain of policy studies, where the stakes of the validity game are higher, and the benchmarks clearer. In section 1, the review of exemplary yardsticks (Shadish *et al.* 2002; LeCompte & Goetz 1982; Lincoln & Guba 1986; Yin 2000a) shows that the concept of validity entails common concerns, yet has been given fairly different contents. It is argued that such differences are legitimate as far as they follow differences in key assumptions - about how to conceive of causality, and about how to best seize it. Validity standards simply secure that the technical construction and treatment of data are kept consistent with the special coupling of ontology and epistemology embedded into a research strategy. Thus, a same standard applies when at least one of the two sides of causal knowledge is shared - and separate standards are required when ontology and epistemology differ.

The further operation required to evaluate the applicability of any validity yardstick to QCA, therefore, consists in the identification of the special ontology and epistemology embedded in its inferences. Section 2 identifies the former in the mechanistic understanding of causality as drawn by Bhaskar (1975) and Pawson (1989), and the latter in the set-theoretical technique developed by Ragin (1987, 2000, 2008) after Barton (1955). Threats, and strategies «for ruling them off», are therefore discussed in the light of these

fundamental assumptions - and some new consistent protocols advanced for improving the causal validity of QCA results, before the final remarks in Section 3.

2. Validity standards

As for policy studies and evaluation, the debate about the contents and the applicability of validity standards has unfolded since the late 1960s from the nomothetic camp. Here Campbell and his associates have provided a systematization which, in its evolution, is still considered the main yardstick and, for long, the single best. Such Olympic might has however been reconsidered after critical reactions from within and outside the approach. Within, the ambition to fully valid causal results has been downscaled, together with the capacity ascribed to the research strategy to grasp actual causation. Outside, the notion of a hierarchy of methods - implied by the nomothetic understanding - has been vigorously challenged by qualitative scholars - with the side-effect of spreading the interest in validity. Trying to justify the inapplicability of the nomothetic yardstick to their studies, scholars in the idiographic camp have thus developed their own criteria. Their arguments, and the contents given to the concept, will then be used for clarifying what - if anything - in validity can be considered a cross-cutting standard, and why.

.1. The nomothetic yardstick

The nomothetic understanding assumes that covariation always indicates some causal connection, and its validity revolves around a threefold proof: that the covariation occurs; that it occurs between the hypothesized «treatment» and «outcome» alone; and that such treatment is consequential compared to a «counterfactual» baseline.

The overall operation is therefore quite delicate, and exposed to many threats. In addressing them, Campbell (1957) developed a typology with two relevant dimensions - «internal» and «external» - later specified by two further additions - «statistical» and «construct» (Shadish & al., 2002; Cook & Campbell, 1979, 1983; Campbell & Stanley, 1966). In it,

1) «internal validity» addresses causal-reasoning errors, and revolves around the question «did the experimental stimulus make some significant difference in this specific instance?» (Campbell 1957:297). It hence focuses on the extent to which evidence neatly supports the causal inference, and

weakens whenever the changes in the outcome cannot be *undoubtedly* ascribed to changes in the treatment alone - here, because of, for instance, errors in the identification of the cause; differential unaccounted changes within the cases under analysis; ambiguities; deterioration of gauges and phenomena; or causal models hidden the error term, in fixed effects, or in the control variables.

2) «statistical conclusion validity» addresses the same question than above, but from the perspective of statistical covariation. It requires that the model is properly specified - that all the causal variables are included; that the relationship is not correlational and is imposed neither the wrong functional form nor the incorrect parameter constraints; that the direction of causality is unambiguous. It therefore weakens if false positives or false negatives are unreasonably likely because of low statistical significance or low statistical power; by relevant omitted variables; by violations of the specific assumptions of key tests in the analysis; or by misinterpreted differences among statistical models.
3) «external validity» is defined by the question, «to what populations, settings, and variables can this effect be generalized?» (Campbell 1957:297). It refers to the robustness of the treatment's causal power - the extent to which its relationship with the effect holds beyond the circumstances of the study, so proving its nomological nature. It weakens whenever the predictor interacts with case selection, setting, history, or other treatments - each of which narrows the domain of validity of the causal relationship.
4) «Construct validity» relates to external validity, as it focuses on the link between «constructs» - namely, general traits, prototypical features, or properties used to hypothesize the causal relationship - and «sampling particulars» - that is, the empirical instances and the measures that operationalize such hypothesis. Their relation weakens if slippages occur, so that empirical instances do not match the prototypical features properly - which leaves the researcher with the problem of establishing whether her results depend on how the hypothesis was defined, or on how it was operationalized. In this framework, the problem becomes either deductive - of precision in the selection of empirical instances, as well as of reliability of their gauges - or inductive - of misspecification of the constructs that substantiate empirical conclusions.

According to this literature, all the threats can be reasonably kept at bay. The ones to internal and to statistical validity can be «ruled off» by design. The threats increase in severity as much as the conditions of a study deviate from the optimality of scientific experiments: so, the first best strategy to

secure validity is to adopt an experimental design - especially with Randomized Control Trials, which therefore becomes the «gold standard» -, or, as a weaker alternative, on covariational analyses structured so to approximate the experimental rationale. Whatever the design, internal and statistical validity also require careful sampling and measurement to minimize errors, and a wise utilization and interpretation of tests. External validity follows when the results from internally/statistically valid studies are confirmed under different conditions of time, setting, units of analysis and gauges. Construct validity is made safe when different constructs across replications and reproductions prove clear-cut and non-overlapping - so that alternative, precise gauges of a same construct correlate, whilst measures of different constructs do not.

a) Accepting limitations...

At the same time, scholars in this tradition agree that a single real study cannot achieve validity along all the dimensions at once. Instead, each analysis necessarily prioritizes, and trades validities - especially internal/statistical for external and construct. To some, trade-offs simply follow constraints to available resources, so that complete validity can be achieved as much as results from reproductions and replications cumulate and are systematized - for instance by meta-analyses (Shadish & al., 2002). Cumulation however requires that replication and reproduction studies meet the criteria of «maximum similarity [...] to the conditions of application which is compatible with internal validity» (Campbell & Stanley, 1966): yet, often optimal comparability is simply not available. Other scholars thus conclude that perfect validity lies well beyond the analytic capacities of any nomothetic design - thus leaving internal/statistical criteria as the main requisites that a study can aspire to meet. As a consequence, even in the fortunate case of a positive result, experimental evidence can only tell that the treatment led to the outcome «in at least some members of some fixed causally homogeneous subpopulations» (Cartwright, 2007:17).

Such conclusion however does not fall short on nomothetic ambitions as much as the heuristic goals of these techniques are made out clear. Variable-oriented studies are mainly concerned with ascertaining the «effect of a cause» - i.e., that «a particular variable or small set of variables makes a marginal difference in some outcome over and above all the other forces» (Shadish & *al.*, 2002:457). Their goal is thus fulfilled when a sound «causal description» is provided - the empirical relation between a predictor and an outcome is «purified» and contrasted with expectations and evidence from a

counterfactual baseline. The analysis therefore narrows on the potential causal power of some property to make an outcome occur, even regardless of its empirical relevance; and only clarifies «whether something *can* happen, rather than whether it typically *does*» or will (Mook 1983: 382). What cannot and should not be expected from such strategy, therefore, is that analyses «completely explain some phenomenon» (Shadish & al., *ibid.*).

Indeed, variable-oriented, probabilistic studies are seldom meant for accounting for why and how an actual causal relationship did (not) obtain across real cases. To some, this is because variable-oriented analyses gauge causation as a constant, probabilistic «arrow» from the treatment to the outcome, and not as variable, deterministic «pretzels» (Cook, 2000).

b) ...and pushing them

A nomothetic technique for complex causation has nevertheless been advanced by Spirtes *et al.* (2000). Developing the agenda set by Pearl (1988), they aim to provide an empirical basis for «intelligent planning» - meant as a judgment about the logical truth or falsity of future conditional sentences («*if X were to be the case, then* Y *would be the case*») on the bases of past counterfactual evidence («*If X had been the case, then* Y *would have been the case*»). Here, an accurate knowledge of the causal structure becomes crucial for identifying what should be manipulated in order to push the outcome in the desired direction: and such structure, they acknowledge, has the shape of conditional causality.

The structural approach hence allows for more than one cause to insists on the outcome - so that either none of them alone is enough for the outcome to occur, or that all are independently sufficient to bring an overdetermined outcome about. Furthermore, here causes can be direct - when they are proximate effectors to the outcome - but also indirect - when their impact on the outcome only unfolds through one «mediator» or more. An indirect cause may also be common to two separate mediators - and, unlike in standard regression-based models, this does not affect their causal relevance as effectors: only, it clarifies the paths of causation to the outcome. The overall structure is then given by causal chains in which indirect causes effect the immediate descendant(s) only, and the last mediators alone effect the outcome.

Complex causation hence can be modeled starting from direct causation. Spirtes defines it first in categorical-like terms. Causes and effects are conceived of as «Boolean variables» - dummy classifications of events «as of a kind», so that each event *A* is paired with its non-occurrence ~*A*. In such

terms, Boolean variable *C* causes Boolean variable *A* if and only if at least the presence or the absence of *C* effects at least the presence or the absence of *A*. As a consequence, in Sprites' terms, being **V** a fixed set of events which includes *C* and *A*, *C* is a direct cause of *A* if *C* is a member of the set **C** included in **V**•~*A* such that *(i)* **C** is cause to *A*; *(ii)* the events in **C**, were they to occur, would cause *A* no matter whether the events in **V**•~(*A*•**C**) were or were not to occur; *(iii)* there is no proper subset of **C** that satisfies conditions *(i)* and *(ii)* (Spirtes *et al.* 2000:21). Interestingly enough, such definition does not delimit causation; rather, it delineates directness, and in a way that makes sense when translated into the language of causal dependence between variables. Here, it reads as: variable *C* is a direct cause of variable *A* relative to **V** provided that *(i) C* is a member of a set **C** of variables in **V**; *(ii)* there exists a set of values **c** for variables in **C** and a value *a* for variable *A* such that, were the variables in **C** to take on values **c**, they would cause *A* to take on value *a* no matter what the values of other variables in **V**; and *(iii)* no proper subset of **C** satisfies *(i)* and *(ii)* (*ibid.*). It so becomes clear that **V** is the overall model, *C* a direct cause in the multi-causal system **C**, *A* is the outcome variable, the events in **V**•~(*A*•**C**) are the error term; and that causality means a statistical covariation of the outcome *A* and the overall causal structure **C** - as defined by effectors which however depend on their «parents» and «grandparents». When **V** is «causally sufficient» to a population - i.e. when every common cause either is perfectly included in the causal structure, or at least takes the same value for all the units in the population - the model can be treated as a deterministic system, in which the values of the «exogenous» uncaused variables determine unique values for all the others.

Such causal sufficiency is achieved when the relationship between the graph of the structure and the related probability distributions meets three intertwined conditions - namely, (1) the Causal Markov Condition: the probability distribution of each vertex in the structure proves independent of the probability distribution of any cause other than its «descendants» and «parents», given parents; (2) the Causal Minimality Condition: each edge in the graph does prevent some conditional independence relation that would otherwise obtain; (3) the Faithfulness Condition: the causal structure displays no further independence relation than the Markov ones, and the overall probability distribution and its graph are «faithful» one another. The first condition is therefore crucial to the remaining two, and special validity weaknesses arise that take the form of challenges to Markov independence - especially after non-homogeneity. So, for instance, when causes in different subpopulations are pooled together in a single variable, the resulting mix of

different probability distributions may prove all possible conditional dependence relations true - with the paradoxical effect that the pooled variable does neither satisfy nor violate the Markov condition. Mixed populations may also lead to probability distributions that statistically satisfy the Markov condition without entailing any substantial causation, but simply because they result from systems with same graph and parameters independently distributed. Also, the Markov condition may not be satisfied for variables so far from each other that no sensible common cause can be envisaged, simply because their distributions have similar trends over time: yet, once again, out of quantum mechanics this would only make sense in some debatable mixture of populations. Furthermore, the same direct cause effecting the opposite outcome in two subpopulations may appear independent to the outcome when the two populations are pooled - thus violating the Faithfulness condition. Hence, a key relevance is given to the consistency of the relationship between variables and homogeneous population, as it can prevent misleading results.

In Spirtes' view, the main threats to the validity of nomothetic complex causation thus stem from the statistical dimension. Again, they can be ruled off by approximating the variables to the conditions of causal sufficiency: once the population homogeneity, the absence of unmeasured common causes, and a clear linear order of variables - especially by time - are given, the values of the probabilistic dependencies in the population are ideally enough to determine the unique graph satisfying the Markov and Minimality conditions stepwise, even without prior theoretical knowledge. Knowledge however still plays a key role in at least the selection of exogenous variables and of last effectors, as well as in the «proper» gauge of all variables - as no reliable inference is possible without.

Apart from a weaker external validity - already considered as a minor dimension in standard experimental studies -, the nomothetic causal structures therefore do not challenge the rationale of the overall strategy: they just widen its causal ontology. Indeed, they still postulates a sequential understanding of causation as a closed linear probabilistic process unfolding, like billiard balls, from remote independent to proximate factors to the outcome. The resulting structure of local «effects of causes-of-causes» is far more informative than the standard model in «effect of a cause» studies; also, it has noble fathers, and finds wide applications (Salmon 1998, Pearl 2000). Nevertheless, it mainly models causality as path-dependency alone: so, despite that the occurrence of the outcome follows the effectors at the end of the causal chain, the attention is shifted to exogenous remote causes as prior determinants - so that desired changes in the outcome are expected after

changes in such remote factors. Interestingly enough, however, these remote causes alone do not seem enough to allow for any reliable prediction about the unfolding of causation. Also, effectors may go statistically undetected, or be statistically indistinguishable - for instance when their effects are «coincidental vanishing partial correlations», which make the whole structure unstable unless cross-kinship ties are removed from the structure.

The nomothetic pretzels are thus modular assemblages of direct covariations, and hold in the restricted domain of precise statistical parameters. Whenever a hypothesis about actual causality violates them in a module, the usual strategy is to drop the hypothesis from the model for the sake of validity. As a consequence, the critiques claim, the rigorous application of nomothetic yardsticks secures representations of causal reality that are self-consistent and clear, yet not necessarily commensurate to actual causation - at the cost of the «usability» of such knowledge.

.2. The idiographic criteria

That case-oriented strategies can be better at seizing actual complex causation is a long-standing claim of qualitative scholars.

Indeed, field research finds its reason in the intensive analysis of the contexts where special actual events and processes bring about specific phenomena of interest. Moreover, case-oriented inquiry seems to resonate better with the mechanistic understanding of causation that has increasingly been attracting attention for its explanation of uneven local causation - up to causation in a single case (Maxwell 2004; Glennan 2002). It assumes that a variable's causal power is a disposition and remains latent unless some mechanism is triggered or defused. The disposition actualizes if the variable interacts with other special conditions which, however, are contextual and irregularly distributed. So, from case to case, a same mechanism may not obtain because of the presence of a hindering factor or the absence of the trigger; or may obtain because of different yet equally effective triggers; or may not obtain, but the phenomenon of interest can still occur because of some alternative mechanism. The approach hence truly focuses on the «causes of an effect», and aims at providing explanations - i.e., at identifying those conditions which, together, account for the (non) activation of a mechanism beneath some local, past (non) occurrence.

As far as mechanisms depend on, or are, concomitances of local events, intensive field research thus can with some reason claim to be better equipped for grasping them. For long, however, the results of qualitative

studies have been only recognized the status of hypotheses, not of proper assessments. Campbell and Stanley (1966) equated case studies to one-group, post-test-only «queasy-experiments» from which ambiguous evidence alone could result, and little be learned with certainty. In a nutshell, such findings could not be taken as valid. A mounting dissatisfaction with the Campbellian yardstick followed. Qualitative scholars complained that experimental requirements had grown into the universal standard for establishing how worth a research was, yet proved impossible to meet - especially outside probabilistic studies. They however agreed that prizing subjectivity and direct observation did not mean that qualitative inferences were free from threats. They therefore stressed the need for fairer validation criteria. The resulting proposals however display remarkable differences in contents, as well as in the distance to the nomothetic yardstick.

a) Radical diversity

From the perspective of ethnographic research, LeCompte and Goetz (1982) refuse the term «validity» for the positivistic meanings it entails. They therefore speak of «credibility», which they see as an issue shared by the two traditions alike. Both are concerned with the «accuracy» of the match between constructs and empirical reality, and by the «replicability» of research results. In their opinion, these shared concerns however do not justify a unique standard, because of the different use of induction. In their words, «experimental researchers hope to find data to match a theory; ethnographers hope to find a theory that explains their data» (*ibid*:34). To them, ethnographic research typically understands phenomena as the «interplay among variables situated in ... an intact cultural scene» (*ibid*.: 33, 54). It deals with such interplay by first providing a thick atheoretical description of the whole complexity, then by unraveling and making sense of it through concepts and hypotheses - starting from those with currency in the context. Such explanations gain the status of causal statements when their «typicality» makes them useful for understanding other contexts, too. Ethnographic generalizability is therefore not of studies, which are unique, but of the resulting constructs alone; for securing it, findings have to «delineate the characteristics of the group studied or of the construct generated so clearly that they can serve as a basis for comparison with other like and unlike groups» (*ibid*:34). Yet, reduction to typicality in ethnographic research is retrospective and lies at the end of the research process; if operated prematurely, it becomes an unmendable source of idiosyncratic bias. So, in contrast to conventional research, here imprecise and redundant

observations become constitutive parts of the good practice, as they later allow the researcher to deal with history fruitfully and to establish for instance «which baseline data remain stable over time and which data change» (*ibid*:45). The dynamic nature of the context also justifies creative adaptations of the observational protocol in the making. The little proceduralization of the research process and design, again, is therefore necessitated by the special epistemology of the tradition, and does not undermine the credibility of results - at least until adjustments are justified and open to scrutiny. Threats rather arise when the researcher becomes so familiar with the context that she gets blind to relevant evidence, or from the distorting effects that her techniques for eliciting responses have on observations - threats which cannot be prevented by collection design, but by «disciplining» subjectivity. Transparency hence becomes the key criterion - that does not secure sound conclusions but indirectly, as it restraints the investigator's behavior by making her constantly aware that her research strategy will later be judged by many different observers.

A similar yet more radical and systematic redefinition of validity comes from Lincoln & Guba (1982, 1986), and from their understanding of the qualitative methods as expressions of a single «naturalistic» counter-paradigm of inquiry. Their key tenet maintains that «all human behavior is time- and context-bound» - which to them entails the hopelessness of law-like knowledge. A better research instead develops from «working hypotheses» about the local unfolding of actions - explainable in terms «of multiple interacting factors, events, and processes that give shape to it and are part of it» (Lincoln & Guba, 1986:17). Indeed their naturalistic inquirer can «establish plausible inferences» about «patterns and webs» of such complex local unfolding: but from data shaped in an open relationship of bargaining, mutual learning, and joint control with the respondents. In their view, the impossibility of research neutrality also compels the inquirer to make the research process into a strategy for improving the social interactions she is observing: so, under the label of «authenticity», they defined new special requirements that naturalistic research has to meet - such as the fair representation of all the values; or the positive effect of the research process and of its results on the actors - in terms of higher awareness of their context and dynamics, improved capabilities, deeper engagement. As for the knowledge so created, Lincoln & Guba redefine Campbellian validity as positivistic «rigor», then pit naturalistic «trustworthiness» against it. In their «trustworthiness»,

1) Internal validity becomes «credibility», indicating the agreement of the inquired that the inquirer's insight is plausible. Credibility is therefore

undermined by sources of «distortion» such as saliences in some situation; biases of the investigators, of the respondents, or of the technique for data collection; and contradictory information;

2) External validity becomes «transferability», that is, the plausibility of applying judgments from a context to another. It weakens when it cannot be established whether the receiving context is dissimilar from the sending one;
3) Reliability turns into «dependability», which indicates that the process and goals of inquiry have appropriately adapted to the evolution of the social processes of which the investigator is part and observer. It weakens when the inquirer is not transparent about such adjustments;
4) Objectivity makes place to «confirmability», meaning that interpretation is transparently grounded in observations. It weakens when reconstructions are not substantiated with proper evidence.

Here again, threats can be controlled. Credibility is increased by prolonged engagement and observation, triangulation, peer debriefing, negative case analysis, and checks and feedbacks from the inquired. Transferability requires that the inquirer's insight goes with a thick description of the context, so that the similarity of the «receiving» one can be evaluated - «although it is by no means clear how "thick" a thick description needs to be» (Lincoln & Guba, 1986:19). Both dependability and confirmability call for transparency of inferences and of the underlying processes, and can be secured by a competent external, disinterested audit of process and products, respectively.

b) Normalized diversity

In contrast with relativistic approaches to validity, Yin (2013; 2000) has institutionalized a successful account of the case-study method in which «rigor» can fruitfully apply to qualitative findings, too. In his view, case studies do produce sound scientific knowledge: whereas they cannot justify inferences about populations, they are nevertheless capable of «analytic generalizations» to theory. Especially when driven by a «why» question, they can corroborate a concept against some rival hypotheses, or provide empirical reasons for developing new ones. Otherwise said, they too can prove causality by rejecting alternative explanations: only, while probabilistic strategies conflate such alternatives in error terms of which they prove the irrelevance, case-studies adjudicate on few explicit rivals which are assumed relevant to the events under analysis (Yin 2000). This explanatory capacity however unfolds when, as in conventional quantitative studies and unlike much

qualitative inquiry, the research design is detailed, and consistency is maintained between empirical evidence and the starting question. Yin finds the case-study rationale so close to the nomothetic standard that he takes the Campbellian typology onboard with almost the same labels - and some adjustment mainly for disconnecting validity from probabilistic data and treatments. So, in his proposal,

1) Construct validity becomes the first criteria, requiring the identification of the «correct» measures to operationalize theoretical concepts and hypothesize relations *before* fieldwork. The operation is meant for preventing the researcher from stacking the deck in favor of some pet idea while collecting data;
2) Internal, or logical, validity requires that inferences are unambiguous - i.e., that «compelling» relationships are established between antecedents and consequences. Data collection has hence to be drawn so to anticipate possible ambiguities during the analysis, and to make conclusions «airtight» to counterarguments;
3) External validity relates to the generalizability of results, and entails the definition of the theoretical domain in which the findings hold;
4) Reliability refers to the need of minimizing errors and biases along the research process, and requires evidence that the key research operations would lead to the same results if reproduced by a different researcher.

Differences nevertheless resurface in the strategies for case-oriented research to meet the four criteria. External validity means theoretical generalization alone, and increases when the study revolves around wide-ranging «if-then» hypotheses. Reliability passes through the transparency of the research process, and is secured by explicit protocols and the creation of case study databases. Construct validity improves through triangulation, member checks and feedbacks, and consistent theoretical-empirical connections. Internal validity is strengthened by establishing explicit expectations about key empirical patterns before the analysis, by building consistent empirical explanations, and by providing logical proofs of the explanatory (ir)relevance of rival hypotheses. On this latter point, Yin agrees with nomothetic scholars on the importance of research design, and of counterfactual reasoning in assessing «rivals». Ambiguous conclusions can hence be prevented if multiple cases are selected purposefully - because they are critical, extreme, unusual, common, or revelatory of theories, so that they can provide clear evidence about special conditional statements inspired by counterfactual thinking. In this way, the local relevance can be tested of key alternative explanations - such as chance, history, maturation, instability, mortality, implementation, or a wider «suspect» than hypothesized.

.3. A rose by any other name?

Beneath their differences, all positions can agree with Maxwell (1992:283) that, by and large, validity focuses on the «relationship between an account and something outside - whether this something is construed as objective reality, the constructions of actors, or a variety of other possible interpretations». As such, validity is a cross-cutting concern about the consistency of the relationship between the starting question, the selection of information, its transformation into treatable aggregates, analysis, and conclusions about causation. Otherwise said, both camps have developed strategies to secure that their gauges and processes of induction lead to results as free from biases and ambiguities as possible. Also, regardless of labels and classifications, valid causal inference always entails some use of counterfactual thinking - usually operationalized in the research design as «negative cases» or baseline «control groups». Yet, the few but influential examples above provide evidence that these common points become striking differences when such strategies are detailed.

In the nomothetic camp, requirements for a sound inference mean that (1) *data* are defined to operationalize a clear covariational hypothesis; come from gauges of the prototypical cause and effect as precise as possible; are collected from a wide and homogeneous sample; and are aggregated into variables whose behavior fits the assumptions for a sound statistical treatment. Also, (2) the *analysis* runs convincing tests of the insulation of the prototypical relationship from any other source of influence, ambiguity and bias - so that the treatment is left as the only cause, and its effect can be evaluated against some baseline. Moreover, (3) *information* is provided that makes the study replicable and reproducible, so that the prototypical relationship can be cumulatively proven a nomological nature or, more reasonably, be given a clear empirical domain.

Quite differently, a sound idiographic inference instead follows when (1) *observations* are selected after some theory, although at the beginning they may be only loosely related to constructs; are rich enough, and as balanced (or fair) as possible, to provide arguments for the application of a leading theory and of some explicit rivals; are constructed transparently (though not always dispassionately) from one or more theoretically relevant cases; may or may not be coded, but are always aggregated into statements of facts which are non-contradictory and commensurable with rivals - or part of them. Also, (2) the analysis displays extensive evidence that statements of facts are

grounded, and logically connects them to theoretical concepts. Moreover, (3) information is provided for other researchers to evaluate the similarity of the case to other contexts of interest - which define the applicability of a theory.

Why similar concerns turn into such different benchmark cannot simply be explained by the type of data used or by the method of treatment. Such preferences do not clearly discriminate among research strategies - at least because of the empirical eclecticism of idiographic studies. Differences rather run at a deeper level - of the core tenets about the nature of causation, of inference, hence of good results. From this perspective, in the nomothetic camp we find that

a) *causation* indicates that an «if-then» law-like relationship exists; takes the form of a probabilistic sequence of classes of events; and is properly seized in statistical terms as a covariation;
b) to *infer* causation means to prove that a change in the probability of the prototypical effect follows a change in the probability of the prototypical cause - either simple or compound - neat of other influences;
c) *good results* allow for prediction of the future values of the effect, at the aggregate level, given the cause and a stable domain of validity.

Quite the opposite, the idiographic camp maintains that

a) *causation* indicates the complex local co-generation of some phenomenon of interest; takes the form of situated interactions among multiple social and individual factors; and is properly seized in theoretical terms, as the discovery/unveiling/construction of connected constructs;
b) to *infer* causation means to disentangle the factors beneath the occurrences in the context of analysis, then to see how they logically fit a theory (or its rivals);
c) *good results* make broad sense of past local interactions as patterns or relationships with theoretical and evaluative relevance.

Thus, it is the difference in ontological and epistemological assumptions about causality which makes the two strategies into alternative research paradigms, and legitimizes the difference in the prescriptions for securing consistency to the research processes - hence, in the validity yardsticks.

What does it mean to QCA?

3. A validity of its own

QCA is often understood as a hybrid of qualitative and quantitative analysis - a mixed method. Indeed, on the one side, QCA displays many

elements of the idiographic strategy, especially from Yin's explanatory case studies: it is theory-driven; it relies on cross-case comparisons; it does not speak the language of probability, instead looking for the «dead causes» of an effect. On the other side, its source of counterfactual evidence lies in all the cases which meet a clear scope condition; its analysis is systematic; and its results may evoke those of multi-causal and cluster analyses. At a deeper level, however, QCA (Ragin 1987, 2000, 2008; Rihoux and Ragin 2009; Schneider and Wagemann 2010, 2012) really stands out as a third research paradigm, as its understanding of causation is substantially different from any other. Indeed,

a) QCA agrees with the idiographics that *causation* indicates the complex generation of some phenomenon of interest; with structural nomothetics, that such complexity is conditional, set-theoretical, and Boolean. Yet, QCA basically focuses on «chemical» reactions activated by the presence or the absence of special properties, or conditions, consistently with Bhaskar's philosophy of science. Causality is the potential power of a property which only unfolds after the activation of some deep mechanism unobservable to the researcher. The researcher can nevertheless *explain* the uneven occurrence of an outcome: first, by guessing the functioning of the activating mechanism; then, by deducing which special enabling and triggering system conditions would activate the mechanism, were the guess true; and eventually by verifying that the outcome actually occurred in those systems alone displaying these conditions as expected (Bhaskar 1975; Pawson 1989; Ragin 1987; Befani & Sager 2006). Therefore, causation is seized in set-theoretical terms, as an asymmetrical relationship of sufficiency between the joint occurrence - i.e. conjunctions - of generative properties and an outcome. Such set-theoretical relationship can take many shapes. «Standard QCA» appraises one-shot conjunctions of reagents (Ragin 1987). If the order in which the reagents are added to the system is supposed to influence the activation of the mechanism, conjuctions can be conceived of as temporal ordered sequences by «time-QCA» (Ragin & Strand 2008). If different reagents are hypothesized to trigger after different contextual catalysts, such nested causation can be modeled in a «two-step QCA» (Schneider & Wagemann 2006). All these shapes of causation are not discovered at the end of the process, but hypothesized at the beginning whence they shape up the research protocol.
b) *Inference*, again like in the idiographic strategy, takes high complexity as its starting point and consists in reduction. However, as in the nomothetic understanding, this complexity is restricted to few conditions - selected

after the theoretical expectation that they would have jointly caused the outcome, had they been observed. The starting hypothesis is therefore a conjunction which includes all the triggering and enabling conditions that can be theoretically supposed to activate the mechanism of interest when present in a system. Cases hence become instances of the starting hypothesis when they display all the theoretical conditions and associate it to the occurrence of the outcome: yet, there may be cases lacking one or more of the theoretical conditions and still leading to the outcome, and cases with some generative conditions where the outcome did not occur. Inference depends on the evidence of the generative power of such alternative «primitive configurations», which counterfactually demonstrates the irrelevance of single properties to the occurrence of the outcome given other reactants: when a property is the only varying part in two otherwise alike configurations with same outcome, its contribution proves irrelevant to the explanation, and it can therefore be dropped. Each reactant can therefore be found irrelevant to some explanation yet not to others, depending on how configurations are matched. Hence, there are as many explanations as «minimization paths». These solutions, or «prime implicants», detail which mix of reacting properties account for the occurrence of the outcome in special clusters of cases, or single cases. The starting hypothesis can however be falsified when even one prime implicant is contradictory - covering cases with positive and negative outcomes.

c) Good *results* therefore are unambiguous explanations - that is, minimal sufficient configurations accounting for the occurrence, and separately the non-occurrence - of the outcome across cases in a population at a given time point. As such, results have evaluative and theoretical relevance - as they can clarify which configurations succeeded and which failed.

In the light of these special tenets, it makes sense that, in QCA, a sound inference follows when (1) *row data* are identified after a theoretical hypothesis about the capacity of conditions for activating a mechanism; are collected from a population of cases, selected neither with the aim of homogeneity nor for their exemplarity, but because of a scope condition consistent with the theoretical hypothesis and the operationalization of the explanatory conditions; may be either qualitative or quantitative, yet have to provide information about at least the two basic statuses of presence and absence of any causal property in each case - so that the case can be matched with a theoretically possible configuration; are «calibrated» into crisp, fuzzy, or multi-value conditions - i.e., dummy, continuous or categorical degrees of membership to the underlying property-set - after the transparent and

reasoned identification of thresholds corresponding to critical points at which the membership status changes. Also, (2) the *analysis* addresses contradictions in observed configurations before minimizations, as well as in the counterfactual use of unobserved configurations during minimizations. Moreover, (3) *information* is provided that make the operations open to scrutiny - but results are «time- and context-bound» explanations of the uneven activation of a mechanism across specific cases. Thus, the starting theoretical conjunction can be redundant enough for explaining different populations at different time-points: but the results of an analysis only apply to the population under analysis.

Ontological and epistemological assumptions therefore can provide researchers with a criterion for sorting proper from improper threats to inference - and relevant solutions from misleading ones. In so doing, however, previous classifications may prove confusing. Instead, we can boil the many reasons for flawed results down to a cross-cutting threefold problem: the researcher was fooled - by gauges, by technicalities, or by design.

.1. Fooled by gauges

One of the earliest concerns about QCA validity focuses on how accurate and clear the correspondence can be of raw variables and property-sets *via* conditions. As the first version of the technique came in crisp values, at issue there was the excess of information lost: binary coding entails the explicit choice of maximizing the differences between members and non-members of a special property-set, at the expenses of the differences within each group, which made inference especially exposed to errors. The improvement came with fsQCA, where the transformation is more fine-grained. With the aim of turning natural language into degrees of membership to the property-set, Ragin (2000) built a conversion table of a scale of «natural evaluations» (from «fully in» to «fully out» the property-set, through the maximum ambiguity of a «crossover» threshold) to membership scores (from 0 to 1, with the crossover conventionally at 0,5) mainly useful for calibration *via* expert evaluation. For continuous raw measures, an algorithm was provided that pegs the raw values to a sigmoid function by feature-scaling transformations with three anchors - the inclusion threshold, above which differences in the raw value are irrelevant as the cases are already fully in; the exclusion threshold, below which again the null membership is insensitive to differences in raw values; and the crossover, shared with csQCA, which sorts «almost in» from «almost

out» cases. The decision of where the thresholds should be set is left to the researcher, required to be transparent in her choice; however, suggestions have been made about external and theoretically informed criteria as the first best, followed by decisions based on the meaning of the distribution similar to clustering after «natural gaps» - the worst being the unjustified use of basics statistics borrowed from the routines of the nomothetic camp.

Despite the numeric nature of the operation, the original intention of calibration does seem closer to the naturalistic understanding of the research process as a continuous recursive adjustment of theory and evidence unless a coherent picture is provided. So, in textbooks, threshold shifts are also suggested as a way to solve possible contradictory lines in the truth table without refining the starting hypothesis - and alternative to either dropping the case as not informative enough, or treating its value on the ambiguous condition as a «don't care». The choice however requires sound justifications, given its consequentiality. As errors can be detected in the relationship of the conditions and the outcome, every change in gauges affects it - and it is not always clear whether, by recalibrating, we are adjusting evidence to theory, fixing an error, forcing dirty data into the working assumptions of the method, or manipulating reality for the sake of technical fit. Indeed, recalibration moves at least one case which would otherwise falsify the starting hypothesis as such from an observed primitive configuration in the truth table to another. In itself, this treatment in csQCA may create a false positive in an allegedly explanatory model. In fsQCA the problem can be better detailed with some yet no definite improvement. Here, the recognition of a «measurement error» can follow evidence that actual ambiguous cases are too closed to the crossover - of which it is reasonable to suppose the wrong classification. Recalibration can therefore stand as a strategy to treat high case ambiguity so to maintain the starting hypothesis untouched. Apparently, the consequences are less relevant. By crossing the threshold, the ambiguous case transfers its ambiguity from an observed configuration to another, which so is associated to an outcome -let's say the negative. The effect on the consistency values of other negative configurations can prove almost irrelevant, and the new observed ambiguous configuration will not enter fuzzy minimizations for the non-outcome, so that its contribution to the causal paths is minimal, too. Nevertheless, the move improves the consistency of a now positive configuration which so will likely enter minimizations and solutions. The risk of having so created a false positive, although in the complementary field, is still present. To some scholars, this is enough for rejecting QCA - especially in its crisp version - as an unreliable method; others consider it as something «inescapable in the practice of data analysis»,

which however calls for at least sensitivity tests (Varsey 2014). Indeed, recalibration has been increasingly used with this aim. Having weak reasons for a certain setting of thresholds, alternative calibrations can be applied to a same condition, and if this lead to remarkably different causal paths, results are considered unreliable. The severity of sensitivity can also be judged on the basis of the special changes that recalibration sorts on the different kinds of solution: gauges are deemed robust if the change does not affect the results of the parsimonious solution (Fiss 2011). Thiem (2010) better details such source of unreliability as the joint effect of the crossover with the functional shape of the membership score. He proves the logistic shape, embedded in the calibration command of the most popular software, to be very sensitive to such changes - which makes it suboptimal for stabilizing results when compared to, for instance, linear transformations. However, the weakness of the original S-shape may turn into a conservative assumption for avoiding false positives. In any case, recalibration to fix a measurement error can raise fewer problems if substantive reasons can justify the move. As such, it is deemed to affect QCA especially in large-N (Maggetti & Levi-Faur 2013), or in «inductive» applications.

Recalibration however can be misused if it treats as a measurement error what could instead be due to the bad determination of the truth table, or by a misspecified scope condition and other design problems.

.2. Fooled by technicalities

Another quite disputed point is how reliable the technicality is for induction, because of two main reasons: ambiguity, and limited diversity.

a) Ambiguous sufficiency

All the theoretically possible alternative configurations in the truth table are potential statements of sufficiency, which cases actualize and associate to an outcome. Especially in fsQCA, however, such statements do not come all with the same strength. The subset relationship of a configuration to any outcome may be far from perfect - indicating an ambiguous triggering capacity. Conclusion based on such evidence may therefore be flawed.

The technique then addresses the problem by excluding such ambiguous configurations from the minimization to the outcome. A gauge of the strength of the relationship is provided by a special parameter of fit, the «consistency of sufficiency» («S-consistency» for short), ranging from 0 (total inconsistency, hence perfect subset relation to the complement of the

outcome) and 1 (total consistency, hence perfect subset relation to the outcome). Conventionally, ambiguity is deemed to affect those configurations with S-consistency below 0,75, but caution raises the threshold to at least 0,80. Only configurations associated to the outcome with a stronger relationship than the threshold will be minimized for solution. In such way, the results will be free from false positives - although sometimes at the cost of generating «coverage outliers» (an unknown problem in csQCA).

b) Limited diversity

Indeed, non-contradictory truth tables and consistent minimizations are not enough to automatically secure valid results - mainly because of limited diversity. The problem goes beyond the too-many-variables-too-few-cases curse. No matter the number of cases, a truth table from actual data is hardly saturated, because some of the theoretically possible configurations simply prove empirically or (onto)logically impossible to happen in the population of interest. Nevertheless, the experimental rationale requires that all these counterfactuals are taken into account, for solutions to be valid. To those who take the rationale seriously, a simple minimization of the observed configurations alone, like in complex solution, does not seem enough to infer causal results - simply because the matching is fatally entrenched in records and «missingness» (Varsey 2014).

The use of unobserved configurations - «logical remainders» - is therefore the unavoidable requisite of causal results. In QCA, this entails that unobserved configurations are given a value in the truth table, so that they can match observed configurations with same «truth» values. But where do these truth values can come from?

A first possibility is: from the heuristics of parsimony. Unobserved configurations can simply be used under the assumption that they always contribute to the same outcome than observed configurations, if this gets to more general results. This, in the analysis of the positive outcome, would imply that the mechanism can trigger regardless of how conditions combine - except for the observed negatives, which would be anyway excluded from minimizations. The assumption then really challenges the hypothesis because it basically assumes its empirical irrelevance. This occurs by virtue of the special rationale of QCA minimization, especially the one embedded in the Quine–McCluskey algorithm. What the algorithm looks for in counterfactuals is not a reason to keep a condition from a theory, but a reason to drop it as irrelevant. The paradoxical result is that if all the «logical remainders» are true, then the starting theory is false. Vice-versa, if all the

logical remainders are false, then they cannot be minimized with the observed configurations to the positive outcome and the starting theory is true - but then, all we can end with are our observed primitive configurations alone, that is, little more than tautological descriptions. And these actually are the rationales beneath the first two kinds of solutions displayed in Standard QCA - namely, the «complex» (all remainders false) and the «parsimonious» (all remainders true).

The disturbing element of parsimonious heuristics is that the algorithm can make a contradictory use of unobserved configurations while getting to the solution. A same logical remainder can therefore enter negative and positive minimizations alike - i.e., it can be used as it was capable of generating a different outcome depending on its match. This is hard to accept if we maintain that the truth table shares with any other analytical causal space some key features, necessary to perform induction: hence, that it is a close, single, non-contradictory space (Bakhsar 1975, Lazarsfeld & Burton 1956). So, especially when dealing with logical remainders, some criteria are required to avoid their contradictory use. Moreover, QCA is meant for delivering as realistic solutions as possible. Thus, we may want to sort untenable from plausible truth values of logical remainders, so that the latter alone enter minimizations. Criteria for guiding the researcher's decisions on whether using or barring special remainders to an outcome are usually found in (1) theory, (2) ontology, and (3) empirics (Rihoux & Ragin 2009, Schneider & Wagemann 2012).

Theory implies expectations that a special status of a condition, or a special conjunction of key conditions, leads to some outcome. Consequently, the logical remainders which display the condition or the conjunction may be given a positive value, consistently with expectations. As an alternative, in order to preserve these conjunctions from minimizations so that they will certainly appear in the solution, counterfactuals can be barred if, once used, they would make the key conjunction irrelevant. A more sophisticated view allows minimizations of observed configurations and «easy counterfactuals» alone: so, if evidence tells us that a configuration obtains in *absence* of condition C, and C is theoretically *expected* to contribute to the outcome, then the matching counterfactual in which C is *present* can be associated to the outcome, too. This rationale can be also extended to «uneasy counterfactuals» - i.e., logical remainders which violate the theoretical assumptions about the generative power of a condition - that so can also be barred. A possible critique to such theory-driven use of counterfactuals is that it imposes a confirmatory bias on results - the only added value lying in some detail, and

perhaps in some new path. The critique however applies to those manipulations alone aiming at the preservation of some pattern by shielding it from minimizations to irrelevance. In any other case, theoretical expectations will rather work *against* the preservation of the hypothesis as such. Indeed, easy counterfactuals are the conventional correction that the software applies to parsimonious minimizations in QCA Standard Analysis, resulting into the third kind of solution - the «intermediate» - which is usually presented as the refined QCA findings. Nevertheless, easy counterfactuals may not be enough to get solutions rid of the inconsistencies of simplifying assumptions. Further criteria may hence be required to adjudicate on the outcome of a contradictory logical remainder.

Ontology. Another reason that can justify the direct assignment of truth values to logical remainders is the researcher's consideration of the plausibility of the (onto)logical assumptions. Nonsensical configurations postulating impossible combinations of conditions can be considered false and barred - either from the minimizations to the outcome to which they do not make sense, or from minimizations *tout court*. This second choice however would constrain the possibility of reductions in the complementary analysis.

Evidence. Intervention on logical remainders may also be justified if previous evidence indicates that a condition has an empirical explanatory power such that it should appear in solutions, but the parsimonious minimizations make it disappear. This is especially the case of necessary conditions, detected at the beginning of the analysis when the parameters of fit are calculated for all the explanatory conditions. These parameters reveal whether each condition is individually necessary (i.e. a superset) to the occurrence of the positive or the negative outcome; how perfectly (consistency); and with which empirical relevance (coverage). Convention has it that a condition is necessary to the outcome if its «consistency of necessity» value (or «N-consistency») is higher than 0,95. In such case, the condition will appear in almost all the paths of the solution to that outcome - so that it can be factorized. From its set-theoretical definition, we also know that when a condition is necessary, there cannot be cases carrying the outcome which do not display it. Therefore, to preserve a necessary condition into solutions, we can bar from minimizations all those contradictory logical remainders which would violate this rule - and use them for the minimization of the complement. On a similar vein, a further strategy to deal with contradictory minimizations can be deduced from empirical sufficiency. The analysis of necessity also calculates the value of the «consistency of sufficiency» («S-consistency») when it provides the empirical relevance of

necessity relationships. Again, the threshold for a non-ambiguous sufficiency is usually set between 0,80-0,85; the higher the value, the more perfect the subset relationship of the condition to the outcome. From the definition, we know that a sufficient condition is such that all the cases which display it also display the outcome, so that no case with the condition can be expected to lead to the complement. High values of S-consistency can thus provide a guideline for decisions about contradictory simplifying assumptions in the complement.

Empirical adjudication on contradictory simplifying assumptions may be preferred to theoretical preserving ones as it avoids confirmation biases and keeps minimizations consistent to clear distributional evidence. Whatever the strategy, however, the methodological literature suggests a further standard that results should meet to prove the inference valid: complex, intermediate, and parsimonious solutions should prove perfect nested supersets of increasing generality (Schneider & Wagemann 2012).

Yet, a good inference cannot certify validity if it does not apply to a sound design.

.3. Fooled by design

In a sense, QCA techniques assume sufficient causality, and so specific that its occurrence cannot be a simple random effect. Were it random, the N- and S-consistency parameters of the conditions in the hypothesis would be low to the point of leading to solutions not consistent enough for minimization. Also, Quine-McCluskey's minimizations do challenge the validity of the starting statement of sufficiency extensively: «The method of elimination is superior to both Mill's method of agreement and his indirect method of difference because the focus is on eliminating causal conditions, not confirming them» (Ragin 1987). To many, however, this does not mean that what is left is causal - and reasonably so. In a nutshell, if correlation is not causation, neither is configuration (Pawson 2008, Schneider & Rohlfing 2013, Schneider & Wagemann 2010).

The correspondence of results to some ontological connection can be quite shaky when QCA is used inductively, for discovering a theory from somehow related conditions. Indeed, there is no technique which can secure causality unless it treats factors or properties which are deemed causal for reasons external to the technique itself - namely, theoretical. Possible new causal relations may rather be suggested when a theoretical hypothesis proves logically true yet unable to account for all the cases - which indicates

that the starting hypothesis is under-specified. Coverage outliers thus provide a puzzle that may lead to refined or enlarged hypotheses, operationalization, and testing. The underlying requisite however remains that the starting hypothesis relates to the triggering and defusing of some mechanism of interest - which provides the proper rationale for making sense of configurational solutions, and of their limits. Only so, coherence can be of the whole analytical process - not only of the technical part of induction.

The selection of theoretically consistent conditions cannot be enough to minimize the risk that, despite the technical machinery above, results are still flawed. A further unaddressed problem lies in the over-specification of the starting hypothesis. The problem is somehow neglected by standard QCA, which assumes that minimizations will tackle it properly. However, this may not be true. Once a condition is considered as constitutive of the first statement of sufficiency, all the analysis will treat it as it were causal - dropping it only if logical remainders can prove its irrelevance. Yet, its presence can still inflate solutions and contribute to exacerbate problems of measurement, contradictions, ambiguity, and limited diversity (Marx 2006).

Consolidated solutions to over-determination rest on the aggregation of conditions in the starting hypothesis into higher-order constructs (Schneider & Wagemann 2010, Elman 2005), which however may still not be enough. A different approach suggests that a two-step QCA can be applied (Schneider & Wagemann 2006): yet, this can be only appropriate when the starting hypothesis is really made of explanatory factors belonging to two separate ontological levels - that of remote catalysts, and of proximate reactants.

A different rationale considers the opportunity of dropping conditions on the basis of their explanatory power. Especially necessary conditions can prove trivial: because of a N-consistency so high - a distribution so close to a constant - to appear almost tautological to the outcome; or because of a N-coverage so low to look like an idiosyncratic explanation (Braumoeller & Goertz 2000, Schneider & Wagemann 2012). Yet, dropping a condition on the basis of its distribution may prove unwise: once in solutions, necessary conditions usually improve the fitting of the model; moreover, the absence of an almost-constant condition may be the only difference in the configuration of a case otherwise contradictory.

On a similar vein, it has been suggested that the streamlining of the starting hypothesis can result from the automatic identification of «small causal chains» among the conditions in the hypothesis (Baumgartner 2013). The rationale here again borrows concerns from the nomothetic camp and implicitly equates QCA chemical configurations to linear billiard-balls causal

structures. This can of course make sense only if the original hypothesis is formulated in time-QCA terms, and the order of reactants deemed causal. Otherwise, the concern may lead to an unreasonably underdetermined explanatory hypothesis. Actually, that conditions are not totally «independent» makes almost no problem in QCA. Rather, conditions are required to prove a theoretical capacity of activating a mechanism, hence generating the outcome. As such, it is true enough that better explanatory hypotheses focus on the triggering, hindering and enabling conditions that are proximate to the outcome- there where causality unfolds. Yet, as much as these conditions are not totally overlapping - i.e., do not cover exactly the same cases, thus suggesting that they operationalize almost the same property - no actual suspect of mutual dependency can justify the exclusion of one of them. Neither can they be substituted by some alleged «common underlying factor» of which they may be «effecting mediators» - unless, once again, poor fit and QCA outliers would indicate that the model is underdetermined instead, so that the common factor would play as a necessary condition for the generative capacity of the two effectors[1]. In a nutshell, the identification of a properly determined starting hypothesis from an array of theoretical generative candidates cannot follow distributional or linear criteria. Rather, selection criteria should be identified that are consistent with the rationale of the method - which require some Boolean proof of the difference-making capacity of the explanatory candidates.

A more consistent solution can hence consist in the step-wise determination of the explanatory model from a purposefully wide array of theoretically generative candidate conditions. The construction can rely on the information about the N- and S-consistency values provided by the analysis of necessity in the early steps of the research protocol. The starting point for its construction would hence be constituted by the two conditions with higher consistency values of necessity and sufficiency. These first two of course generate contradictions in the population - which should be addressed by looking for the further condition capable of unraveling the highest number

[1] However improper, the concern highlights a feature of QCA of which researchers should be aware: a certain degree of under-determinacy should be considered endemic to QCA solutions. At best, QCA usually finds sufficient, and seldom necessary and sufficient, explanations to the outcome at the case level (although N- and S-consistency exactly account for the distance of a case from the line of necessity-and-sufficiency, which is therefore embedded as benchmark of causality). The reason for «sufficient only solutions» is quite commonsense: the statement of sufficiency unfolds and is tested within a special necessary condition - namely, the scope condition - which cannot be operationalized, as it would really be a useless constant, hence with which the starting hypothesis does not interact. The basic assumption about the scope conditions rather is that it implies many other models, and causes, which in the studies are nevertheless deemed irrelevant for the conditions in the hypothesis to obtain - unless some widening of the analytic scope creates new contradictions which may require the operationalization of some property of the old scope conditions. Therefore the relevance of a clear and reasoned scope condition for case selection.

of contradictions, and by adding it to the explanatory model, recursively unless the resulting truth-table is totally consistent. Such protocol will thus result in a properly determined starting hypothesis - and hopefully reduce its complexity to a conjunction with the minimal number of conditions required to account for the outcome properly. Positive effects will also reverberate on limited diversity. Also, the strategy promises to allow the analysis of causality in designs with higher or lower numbers of cases - provided that they can be thought as proper sub- and super-populations with respect of the model: wider scope with too many cases missing may make any evaluation unreliable, as little but theoretical expectations can drive the decision about contradictory simplifying assumptions. Nevertheless, the stepwise protocol comes at some cost. While it may not hinder the capacity of detecting measurement errors in threshold settings from the interplay of conditions and outcome, it nevertheless assumes symmetry in conditions, and may work properly with standard fsQCA alone. Under other kinds of calibrations - especially if the complement is left open and unexplored (Thiem 2014) - the procedure may develop on a less clear empirical ground.

4. Some final considerations

By itself, all that QCA can answer is the twofold question alone of «why did it fail here, but not there?». However limited these questions may seem, it is worth noting that no previous method has been able to answer satisfactorily before - nomothetic strategies, for their inner limitation to average considerations; idiographic methods, for their mainly conceptual ambitions and little systematic proofs.

The usefulness of QCA knowledge is therefore hard to deny: it provides an interesting operationalization to conditional causation; it can address issues about fairly stable variables; it explains the past of single cases as types. Its strategy greatly improves cross-case comparisons as it is not simply based on the more-or-less justified selection of few cases dissimilar in all but the outcome to look for commonalities, or similar in all but the outcome to look for differences: rather, it builds a structure of systematic counterfactual arguments for which some properly selected population of cases provide evidence, and solutions depend on the identification of irrelevant conditions by the Boolean treatment of pairwise matches.

If the starting hypothesis operationalizes a mechanism, is non-contradictory and properly determined, and if the counterfactuals are treated properly, then the method *can* lead to valid conclusions - about which

configuration explained the outcome in one or more cases. The lessons for the cases in negative configurations to be learned may be that, be they willing to improve their outcome, they should shift in one of the nearest positive configurations at their choice. In so doing, it of course makes assumptions - which however are explicit and open to scrutiny.

QCA thus comes with assumptions, and shares many concerns with twin analyses in the two other camps. Despite they can offer interesting suggestions, improvement can only be developed within the ontological and epistemological assumptions of the method.

References

Barton, A.H. (1955). The concept of property-space in social research. In Lazarsfelf, P.F. & Rosenberg, M. (eds.), *The Language of Social Research*. New York: The Free Press, 40-53.

Baumgartner, M. (2013). Parsimony and Causality. *Quality & Quantity*, 1-18.

Befani, B., & Sager, F. (2006). QCA as a tool for realistic evaluations. In Grimm, H. & Rihoux, B. (eds.), *Innovative Comparative Methods for Policy Analysis*. New York: Springer, 263-284.

Bhaskar, R.A., (1975). *A Realist Theory of Science*, London: Verso.

Braumoeller B.F. & Goertz, G. (2000). The methodology of necessary conditions, *American Journal of Political Science*, 44(4): 844-858.

Brewer, M.B. (2000). Research Design and Issues of Validity. In Reis, H.T., & Judd, C.M. (eds.). *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press, 3-16.

Campbell, D., & Stanley, J. (1966). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Cartwright, N. (2007). Are RCTs the Gold Standard?. *BioSocieties*, 2, 11-20.

Cook, T. (2000), «Toward a Practical Theory of External Validity». In Bickman, L. (ed.), *Validity and Social Experimentation*. London: Sage, 3-44.

Cook, T., & Campbell, D. (1979). Quasi-Experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally.

Cook, T., & Campbell, D. (1983). The design and conduct of quasi-experiments and true experiments in field settings. In M. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology* , Chicago: Rand McNally, 223–326.

Elman, C. (2005). Explanatory typologies in qualitative studies of international politics. *International Organization*, 59(2): 293-326.

Fiss, P.C. (2011). Building better causal theories: a fuzzy set approach to typologies in organization research. *Academy of Management Journal*, 54(2): 393–420.

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(4): 387-401.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, 69(3): 342-53.

Guba, E.G., & Lincoln, Y.S. (1982). Epistemological and methodological bases of naturalistic inquiry. *Educational Communication and Technology Journal* 30 (4): 233-252.

LeCompte, M.D., & Goetz, J.P. (1982). Problems of reliability and validity in ethnographic research. *Review of Educational Research*, 52(1), 31-60.

Lincoln, Y.S., & Guba E.G. (1986), But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Directions for Program Evaluation*, 30(3): 73-84.

Maggetti, M. & Levi-Faur, D. (2013), Dealing with errors in QCA. *Political Research Quarterly*, 66(1): 198-204.

Marx, A. (2006), Towards more robust model specification in QCA. *COMPASSS WP*, http://www.compasss.org/wpseries/Marx2006.pdf

Maxwell, J.A. (2004). Using qualitative methods for causal explanation. *Field Methods*, 16(3):243-64

Mook, D.G. (1983). In defense of external invalidity. *American Psychologist*, 38(4), 379-387.

Pawson, R. (1989). A Measure for Measures, A Manifesto for Empirical Sociology. London: Routledge.

Pawson, R. (2008). Causality for beginners. *NCRM Research Methods Festival 2008*, eprints.ncrm.ac.uk/245/1/Causality_for_Beginners_Dec_07.doc

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge UP.

Ragin, C.C. (1987). *The Comparative Method*. Berkeley: University of California Press

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

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

Ragin, C.C., & Strand, S. (2008). Using Qualitative Comparative Analysis to study causal order. *Sociological Methods & Research*, 36(4): 431-441.

Rihoux, B., & Lobe, B. (2009). The case for qualitative comparative analysis (QCA): Adding leverage for thick cross-case comparison. In Byrne, D., & Ragin, C.C., *The Sage Handbook of Case-Based Methods*, Thousand Oaks: Sage, 222-242.

Rihoux, B., & Ragin, C.C. (eds.), (2009). Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques. London: Sage.

Salmon, W.C. (1998). *Causality and Explanation*. Oxford: Oxford UP.

Schneider, C.Q., & Rohlfing, I. (2014). Case studies nested in fuzzy-set QCA on sufficiency. *Sociological Methods & Research*, 43(2): 1-43.

Schneider, C.Q., & Wagemann, C. (2006). Reducing complexity in Qualitative Comparative Analysis (QCA). *European Journal of Political Research*, 45(5), 751-786.

Schneider, C.Q., & Wagemann, C. (2010). Standards of good practice in qualitative comparative analysis (QCA) and fuzzy-sets. *Comparative Sociology*, 9(3): 397-418.

Schneider, C.Q., & Wagemann, C. (2012). *Set-Theoretic Methods for the Social Sciences*. Cambridge: Cambridge UP.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Spirtes, P, Glymour, C., & Scheines, R. (2000), *Causation, Prediction, and Search*, Cambridge: MIT Press.

Tanner, S. (2014). QCA and Causal Inference: A Poor Match for Public Policy Research. *Qualitative and Multi-Method Research*, 12(1): 15-25.

Thiem, A. (2010). Set-relational fit and the formulation of transformational rules in fsQCA. COMPASSS WP 61, http://ecpr.eu/filestore/paperproposal/0f5605a5-bb78-4bc4-b2be-541bd9ac089b.pdf.

Thiem, A. (2014). Unifying Configurational Comparative Methods Generalized-Set Qualitative Comparative Analysis. *Sociological Methods & Research*, 43(2), 313-337.

Vaisey, S (2014), Comment: QCA Works - When Used With Care. *Sociological Methodology*, 44(2): 108-112.

Yin, R.K. (2013). *Case Study Research: Design and Methods* [5th ed]. Thousand Oaks: Sage.

Yin, R.K. (2000). Rival explanations as an alternative to "reforms as experiments." In Bickman, L. (ed.), *Validity & Social Experimentation: Donald Campbell's Legacy*. Thousand Oaks: Sage, 239–266.