COMPASSS Working Paper 2012-70

Why organizations (do not) evaluate: a search for necessary and sufficient conditions

Valérie Pattyn
Public Management Institute; K.U.Leuven, Belgium
valerie.pattyn@soc.kuleuven.be

Abstract

The wide acceptance of evaluation in this evidence-based society might hide significant variation in the extent of evaluation activeness between public sector organizations. In explaining these differences, evidence is only fragmentally available. Admittedly, multiple explanatory factors can be identified in the evaluation community, mainly in the evaluation capacity building literature. Yet, common to the practical character of the field, insights are mainly of anecdotic nature and have seldom been systematically tested. Thus far, the only certainty is that 'contingency' matters. The inherently contingent nature of evaluation practices may not discourage us, however, from collecting more systematic insight in explaining differences in the extent of evaluation activeness. It is not clear, indeed, to which degree the contingency reigns. The question is whether more parsimonious patterns can nonetheless be discerned, when attacking the complexity. The present paper takes up this challenge. Via a systematic comparison of 27 public sector organizations of the Flemish administration (Belgium) through the application of several configurational comparative techniques (MSDO/MDSO & csQCA), the analysis identifies a range of necessary and sufficient (combinations of) conditions for the (non)conduct of evaluations.

1. Rethinking contingency in understanding organizational differences in evaluation activeness

"Evidence-based policy" as a term is impossible to think away in current-day policy discourse. The marriage of 'evidence' and 'policy' has become so evident, that it has ironically been asked what other types of policy can exist (Gray & Jenkins, 2011). Policy evaluation as one particular form of evidence follows the same trend. As Eliadis, Furubo, & Jacob (2011) stated, it is difficult to imagine a society without a general use of evaluation. The wide acceptance of evaluation might hide, however, significant differences in the extent of evaluation activeness between public sector organizations. Not all organizations have equally adopted policy evaluation practices in this evidence-based era (see e.g. Mackay, 1999; Segerholm, 2003).

In *explaining these differences in the extent of evaluation activeness*, evidence is only fragmentally available. Admittedly, multiple explanatory factors can be identified in the evaluation community, mainly in the evaluation capacity building literature. Yet, common to the practical character of the field, insights are mainly of anecdotic nature and have seldom been systematically tested. Thus far, the only certainty is that 'contingency' matters. The inherently contingent nature of evaluation practices may not discourage us, however, from collecting more *systematic* insight in explaining differences in the extent of evaluation activeness. To date, it is not clear to which extent the contingency reigns. The question is whether more parsimonious patterns can nonetheless be discerned, when attacking the complexity. The present paper will take up this challenge.

From the multiple approaches possible to the 'extent of evaluation activeness', we will narrow the research to *explaining the mere presence or, instead, the absence of evaluation activities among public sector organizations.* Having a better understanding of this rather basic dimension is after all a a prerequisite to proceed to studies about more complex nuances of evaluation activeness (such as: the extent of quality measures taken by public sector organizations; the number of evaluations conducted; etc.). Policy evaluation practices, being ultimately shaped by public sector organizations, the research will take the latter as units of analyses.

To address this research question, we make use of configurational comparative methods (CCM) and different related techniques (i.e. MSDO/MDSO and csQCA).[1] CCM are uniquely attractive to tackle the contextual complexity of evaluation activeness, as they conceptualize cases as combinations of attributes, and start from the assumption that it is these very combinations that give cases their unique nature (Fiss, 2011). The output of this research should bring more systematic insights on the role, importance and interrelationships of the explanatory factors which circulate nowadays. More in particular the analyses should allow us to identify:

1. Whether there are any *necessary* conditions[2] that make organizations evaluate (1a); or that hinder organizations from conducting evaluations (1b).
2. Which (combinations of) conditions are *sufficient* to explain why organizations evaluate (2a); or why not (2b).

Having more fine-grained evidence on the (combination of) factors clarifying differences in evaluation activeness is not merely relevant for theoretical purposes, but could also provide insight for evaluation capacity building, if entities would seek to develop the latter.

The data for this study have been collected in 27 public sector organizations (29 analytical cases) [3] from 8 different policy domains[4] of the Flemish (Belgian) administration. The choice of the Flemish administration is particularly relevant, in light of the government-wide NPM-inspired reforms[5] that were implemented in 2006. The Framework Decree of this reform explicitly assigned the evaluation of policy implementation (instruments used, outputs, and outcomes, etc.) to the departments. Evaluation would hence become a tool for policy

[1] The current study builds on the techniques introduced by De Meur (1996), Ragin (1987, 2000) and Rihoux & Ragin (2009).

[2] Configurational methods bring along their particular vocabulary, such as the term 'condition'. While a 'condition' corresponds with an independent variable in statistical analysis, it is not an independent variable in the statistical sense. There is no assumption of independence between conditions. CCM instead focus primarily on combinations of conditions to be relevant (Rihoux & Ragin, 2009).

[3] To take full account of the potential impact of the issue specific characteristics of an organization (see below), it was decided to split up two organizations into 2*2 analytical cases, if the organizations were active on areas of very dissimilar nature, with potential different evaluation realities. Whether a differentiation was necessary, was checked with the respondents in the survey. As such, while there are 27 organizations involved in the analysis, these correspond with 29 analytical cases.

[4] We chose to focus our research on the evaluation of substantive policies. The policy sectors included in this study concern (1) education and training; (2) work and social economy; (3) mobility and public works; (4) well-being and public health; (5) housing and spatial planning; (6) economy and innovation; (7) agriculture; (8) environment, nature and energy. The entities active on a more horizontal administrative level are not included in the analysis.

[5] The label of the reforms is 'Better Administrative Policy' (in Dutch: 'Beter Bestuurlijk Beleid').

adjustment or improved steering of the autonomous agencies which are responsible for policy delivery. To date, there is hardly any evidence available on the implications of this rhetoric in reality.

2. Evaluating or not? Differences in kind or differences in degree?

As mentioned, while evaluation activeness can in principle be disentangled in many dimensions, we chose to approach the concept in an elementary way, by distinguishing between organizations that do not evaluate on the one hand (referred to as: **Non=1**), and these that indeed conduct evaluations on the other hand (referred to as: **Do=1)**.

Organizations can be rather straightforwardly assigned to either the category of Non=1 or the category of Do=1, on the basis of our definition of policy evaluation. In this paper, we conceive the latter as a "*scientific analysis of a certain policy (or part of a policy), aimed at determining the merit or worth of the evaluand on the basis of certain criteria (such as: effectiveness, efficiency, sustainability, etc.)".* With the notion 'scientific', we refer to the application of objective science-based methods. The evaluations of our interest concern the large array of possible evaluanda, including evaluations on the content of the policy, the process, or the results, effects or impact of the policy (Bressers and Hoogerwerf, 1991). We took into account evaluations on certain policy measures, projects, programs, the general policy of a single policy field or forms of intergovernmental/horizontal policy.[6] This rather generic approach has been mainly motivated by the still limited overall practice of evaluation in our area of analysis (Flanders-Belgium).

Applying this definition, 18 of our cases can be classified as 'conducting evaluations' (Do=1), while 9 cases as not active in any evaluations, and having no intentions to do this in the future (Non=1). Two other cases have a hybrid nature, as they are not conducting any evaluations yet, but have concrete plans to do this in the near future (We label these organization as **Plan=1**). We can classify the latter as not belonging to Do=1, and neither belonging to Non=1.

It might surprise the reader why we use two different categories of evaluation activeness, Non and Do, instead of just one. Key question behind is whether the two 'stages' of evaluation

[6] In the remaining of the paper, we will refer to the generic term of 'policy' evaluations to refer to all these subtypes of policy.

activeness reflect differences in 'degree', or rather in 'kind'. If the former would be true, this would mean that Non and Do are articulations (or subsets) of the same 'set', and subject to the same set of explanatory factors. In case the latter would be rather valid, both types of evaluation activeness are subject to potentially different explanatory factors. In the absence of profound insights on the dynamics at play behind each of the two categories of evaluation activeness, the latter option seems most justified as a starting position. In the paper, we will consequently depart from two different 'outcomes' to be explained, Non and Do. The research should reveal whether there are indeed different conditions for each of these two expressions of evaluation activeness, and whether the 'multiple sets approach' thus proves valid.

3. Explaining evaluation activeness from a perspective of actor centered neo-institutionalism (ACI)

3.1. The choice for ACI as guiding framework

As mentioned, current explanations on organizational differences in evaluation activeness are mainly restricted to normative insights, collected in a variety of empirical settings. In absence of sound evidence to make a prior selection of relevant conditions, we believe that the best way to contribute to the field is an *open, comprehensive approach,* in which we scrutinize a large variety of possibly interesting factors.

Our analysis will nevertheless be ordered by a "theoretical framework". The *framework* deemed most relevant for our purposes is 'actor centered institutionalism' (ACI). Following the founding father of the framework, Fritz Scharpf, we emphasize the notion 'framework'. Given the complexity and contingency of social phenomena, it is hardly possible, neither appropriate to formulate a fully-fledged theory of broad scope with empirical specificity. Instead of providing detailed 'general law like regularities', ACI provides a 'descriptive language', which allows to compare unique qualified cases, by showing how a particular constellation of factors could bring a specific effect (Scharpf, 1997). This understanding of the social reality is hence fully in line with the contingent nature of the evaluation reality.[7]

[7] We will fall back on the principles of ACI, as they were originally developed by Mayntz & Scharpf. Only in a later stage, Scharpf has elaborated his approach by giving game theoretic modelling a central position in it as a tool for analysis (Witte, 2006).

ACI deliberately revolves around concepts formulated at high levels of abstraction, which should be concretely operationalized to serve as useful explanatory factors for our analysis, in confrontation with the empirical reality under study. To this extent, we applied a mixed strategy of literature screening[8] on the one hand (mainly evaluation capacity building literature) and semi-structured interviews on the other hand, with representatives from our particular administrative area. In line with our 'open approach', we continued our search for variables until we reached a point of 'saturation', in which no new variables were encountered. The choice for a single area of analysis, Flanders (Belgium), enabled us to control for a large number of variables.

This mixed approach yielded a large list of potentially interesting conditions, which we subsequently ordered according to the main mediating categories of ACI (see below).

3.2.The mediating categories of ACI

As its name suggests, actor-centered institutionalism in essence proceeds from the conviction that 'social phenomena' are to be explained as the outcome of interactions among intentional actors, but that these interactions are structured, and the outcomes shaped, by the characteristics of the institutional settings in which they occur" (Scharpf, 1997). The underlying assumption is that people do not act on the basis of objective reality and objective needs, but on the basis of their subjectively defined interests, preferences and capabilities which are, but not entirely, shaped by the institutional environment (Scharpf, 1997). In explaining social phenomena, one should thus make a distinction between these two categories of causes: (1) *institutions* and (2) *interactions among intentional actors*.

ACI does not restrict its understanding of institutions to formal legal regulations resorting under the regulative pillar, as traditional rational choice institutionalists would do (Scott, 1995). Also (informal) social norms, imposing normative pressure on actors, are deliberately taken into consideration. Yet, on the other hand, ACI is in some respects more restrictive than sociological neo-institutionalism in deliberately leaving concepts as 'daily routines' or culture out of the definition. This restriction is an important step in realizing the core assumption that institutions structure, but not fully determine behaviour (Van Lieshout, 2008).

[8] This screening of the evaluation literature for our research purposes has been more extensively described in De Peuter & Pattyn (2009).

Applied to evaluation activeness, we will scrutinize the relevance of six different institutional attributes, which 'frame' the interactions between intentional actors: (1) the size of the organization; (2) the extent of formal autonomy that an organization enjoys; (3) the legal status of the organization, referring to the fact whether an organization has the 'department' or 'agency' status; (4) the existence of an evaluation unit, formal or informal; (5) the engagement of staff in the evaluation community (operationalized as the participation in training and networking on evaluation); (6) the presence of requirements for evaluation in regulations to which the organization is subjected.

With regard to actors, the framework explicitly allows the analysis of 'composite actors', composed of a multitude of individuals. The key for this theoretical justification lies in the institutional setting that 'frames' the context of action for individual actors (Scharpf, 1997). Actors are characterized by specific capabilities on the one hand, and specific cognitive and normative orientations on the other. These *capabilities* are understood as all action resources allowing an actor to influence a certain outcome in certain respects and to a certain degree. Capabilities can be of all kinds and for instance refer to personal properties; physical resources; technological capabilities, etc. The importance of capabilities is evident. Without, actors will not be able to realize their orientations.[9] This does not imply, however, that the relationship of capabilities and evaluation activities is straightforward. The mere existence of evaluation capabilities will not automatically be translated into evaluation practices. Other factors can be assumed to matter (Fierro, 2011).

To realize evaluation activities, the following capabilities return as relevant in the evaluation literature, and seem worth to scrutinize in more depth: (1) skills to conduct internal evaluations; (2) skills to steer external evaluations; (3) budgetary resources to conduct evaluations; (4) availability of experienced staff to conduct evaluations; (5) availability of an external evaluation community with expertise on the themes of the organization; (6) extent of development of a performance monitoring system.

[9] As can be remarked, capabilities can –amongst other elements- encompass a structural component. Although clear in theoretical terms, the operational delineation from the institutional setting is often complicating, especially since composite actors (its composition being defined by certain rules) are explicitly recognized as actors in an ACI perspective. This difficulty is also signaled by Witte (2006) in her application of ACI to the Europeanization of higher education systems. For instance: many scholars would intuitively consider budgetary resources as part of the institutional/structural setting. However, in accordance with ACI, these resources are classified as belonging to the 'actor category', as they are of relevance to influence the policy outcome.

Besides capabilities, actors are having certain action *orientations*, which can in turn be disentangled into preferences and perceptions. Here, Scharpf makes again a simplifying methodological assumption, of major significance for conducting empirical research. Actor perceptions should be considered as empirically correct, no matter their theoretical and objective correctness. The perception will constitute the basis of their actions.

Operationalized in an evaluation setting, we can identify three major types of actor categories which are typically involved: politics and administration; science and research and the public sphere (Widmer, 2005). All have their own value claims and rationalities vis-à-vis evaluation. In the present research, we focus on the influence of demand for evaluation from (1) organizational management; (2) Minister(ial Cabinet); (3) Parliament; (4) civil society organizations; (5) other organizations active within the same policy field; and (6) organization-wide support for evaluations.

From a causal perspective, both explanatory categories, actors and institutions, are to be treated differently. According to ACI, "actors and their interacting choices, rather than institutions, are the *proximate* causes of policy responses whereas institutional conditions, to the extent that they are able to influence actor choices, are conceptualized as *remote* causes" (Scharpf, 1997). 'How actors are influenced by institutions' is being stressed, rather than 'how institutions influence actors' (Witte, 2006).

Most descriptions of ACI only focus on these two categories. Doing right at Scharpf's contribution to policy analysis, two extra mediating factors should also be taken into account, *policy issue related variables* and *path dependencies* (Schmidt, 1993). Whether the development of policy evaluation practices is considered necessary and possible, can be presumed to depend on the "fit" with the issue-specific characteristics of the policy organization itself. Scharpf (2000) talks about "specific patterns of vulnerabilities". We will investigate the impact of the extent of (1) salience of the tasks of the organization (proxies: media attention and parliamentary attention), (2) competition on tasks of the organization; (3) the perceived measurability of the organizational output; (4) the perceived measurability of the organizational outcome.

Also the 'path' of the organization will be decisive. As Scharpf states: "Not everything can be changed at the same time. In any one policy area, the body of existing policy must mostly be

considered an invariant environment of present policy choices" (Scharpf, 2000). Whether organizations will currently conduct policy evaluations, will likely be influenced by their experiences with this policy supporting instrument in the past. In the present paper, we will examine the influence of (1) the extent in which the predecessor of the organization (prior to the NPM inspired reshuffling) conducted evaluations; (2) organizational stability; and (3) ministerial stability.

These two mediating categories, policy issue related variables and path-dependencies are also logically situated at a more remote distance from the outcome than the actor related factors. (Schneider & Wagemann, 2003). The explicit distinction between both types of factors at two causal levels offers a powerful analytical tool. Scharpf has also emphasized this, by suggesting to consider them as different steps in an explanatory exercise. Scharpf advices to follow Lindenberg's method of 'increasing closeness to reality' (also known as the 'decreasing method of abstraction'). If the remote conditions, which form the setting in which the actor interactions take place, can already provide a satisfactory explanation, there is in principle no analytical need to descend to a more proximate level of analysis (Scharpf, 1997). Also in QCA, this 'two-step approach' has increasingly gained ground as a tool to come to more subtle, and parsimonious explanations (Mannewitz, 2011). Convinced of its potentials, the present paper will follow this two-step logic (see section 6.3. in particular).

The overview of conditions finally included in the research, can be consulted in appendix 1.

4. A configurational comparative approach

As suggested earlier, considering the contextual nature of evaluation activeness, it is highly unlikely to expect a permanent universal effect of a certain condition on the outcome under investigation. Searching for the 'net effect' of a condition from the outset does not make much sense. It seems more plausible to assume that the effect of the presence/absence of a condition might differ depending on the wider context *(assumption 1)*. In the same vein, we expect that a combination of conditions will lead to a certain evaluation profile (conjunctions) *(assumption 2)*, and that different configurations might "produce" the same outcome ("equifinality") *(assumption 3)* (Rihoux & Ragin, 2009). These broad assumptions constitute the methodological working hypotheses which inspired the approach applied throughout the research. They involve a different notion of causality, known as "multiple conjunctural

causation". As a matter of fact, also Scharpf has stressed his conviction and belief in 'complex conjunctural causation', as an argument behind ACI (Scharpf, 1997).

Rather than decomposing a case in a set of variables with independent linear effects, complex conjunctural causation requires social phenomena to be treated as 'whole entities'. A case-based approach ideally serves these needs. That does, however, not necessarily imply that 'outcomes' are merely the result of specific and unique instances. On the contrary, a certain extent of generalization may well be achieved, especially if approached from an explicit comparative perspective (Byrne, 2009). With this ambition, this research strategically opted to study a medium (intermediate) number of cases. Citing Ragin's words, this strategy "aims at meeting the needs to gather in-depth insight into different cases and to capture their complexity, while still attempting to produce some form of generalization" (Ragin, 1987).

To keep the overview over the intermediate number of cases without falling in 'weak' comparisons, we make use of CCM and different related techniques (i.c. MSDO/MDSO and csQCA). The techniques we apply all rely on Boolean or set-theoretical algebra. This involves a coding of cases in terms of combinations of binary (values: 0/1) conditions and outcomes.[10] With this input, a truth table can be drawn up, which basically presents the data as a list of configurations. A configuration refers to a given combination of conditions and an outcome, with the possibility that a similar configuration may correspond to various empirical cases (Rihoux & Lobe, 2009). Having this truth table at disposal allows proceeding to the heart of the analysis, which involves a systematic and pairwise comparison of the configurations. Mill's (1973) famous canons of agreement and difference constitute the analytical foundations of these methods (Berg-Schlosser, De Meur, Rihoux, & Ragin, 2009). By a deliberate simplification of the complex reality, we thus strive to identify general trends. Moreover, the dichotomisation also enables to compensate for the slight differences in subjective assessment between respondents, common to interview & survey data.

[10] Conventional QCA, or crisp set QCA (csQCA) is dichotomous. Recent advances in the methodology, fuzzy set QCA, allow a more ordinal approach of sets, and permit membership in an interval between 0 and 1. For our research, in which perceptions constitute the basis for many codings, we consider it less appropriate to scale in this ordinal way. We for instance noticed that respondents sometimes use different nuances throughout time to assess the same situation, but are consistent in the general pattern. As we are mainly interested in broad trends, and as we consider a focus on these trends as more reliable in our research, we prefer to rely on csQCA.

5. Data and dichotomization

In order to get a comprehensive and reliable picture of each organization, we applied a triangular data collection approach via the combination of document analysis, interviews and a survey. The survey was sent to these respondents which were in a prior stage already interviewed. Interviews with ministerial cabinets were set up as a double-check with the data received from the organizations themselves. The document analysis mainly involved the verification of studies which were reported as 'policy evaluations' by the different organizations. Only these organizations which participated in the interviews *and* in the survey were kept in the analysis. The survey, which dominantly consisted of closed questions constituted the primary source of reference for the binary coding of the data. Appendices 1 and 2 respectively describe the thresholds used for the crisp set coding of the variables, as well as the resulting data table.[11]

6. Data analysis

Our analysis proceeds in three major steps. Each of the steps generates a different but complementary output, of relevance for both analytical and practical purposes.

6.1. Identification of the necessary conditions

Necessary conditions represent "a core part of social science research" (Goertz, 2003, quoted in Keading, 2006). A variable found as necessary implies that it exerts its influence independent of its accompanying variables. The evaluation literature identifies various factors which may account for differences in evaluation activeness. As yet, it is unclear, however, which of these factors are necessary for each of the outcomes of explanation.

One can think of an infinite number of necessary conditions for any social science phenomenon. (Goertz, 2006). Crucial question is thus how we can identify these necessary factors which are relevant- read: *not trivial*? QCA provides the technical tools to reveal these conditions, and especially to indicate their relevance (Ragin, 2006).[12] Two measures of fit play a crucial role in this respect.

[11] QCA 'best practices' in principle also require the presentation of the truth table. Yet, given the large number of conditions included, every case is expressed by a single configuration. The truth table and data table are thus identical.

[12] For the analysis below, we used the necessity analysis tool of the fsQCA.2.0 software

First of all, the qualification that a condition is necessary is a strong statement, which should be applied with caution. Only when the *consistency*[13] is 100%, this signals perfect necessity, in which the condition seems to 'enable' the outcome (Mannewitz, 2011). To open up the possibility to take e.g. measurement errors, chance, randomness and other "troubling aspects of social data" into account (Ragin, 2000), we use a necessity threshold of 90%.

The consistency value only gives a partial assessment of the relevance of a particular condition. To have a more accurate image, one should also have eye for the *'coverage'* measure. Simply put, this refers to the extent in which the consistent necessary condition is shared by cases across two values of the outcome.[14] If a particular necessary condition is featuring widely in both the presence and the absence of the outcome, the necessity qualification can be considered trivial. In our research, we consider a necessary condition as trivial if it indeed appears in both the presence and the absence of the outcome, *or* when its coverage value is lower than 50%.

Table 1 offers the overview of necessary conditions for the entire pool of conditions, which do have a minimum consistency value of 90%. The coverage score of each of these conditions is mentioned between brackets. Conditions found as necessary but trivial from a coverage point of view, are written in *italics*. We perform the necessity analyses for both the presence (1) as well as the absence (0) of the outcomes, in line with the assymetric assumption of causality, common to CCM. Note that the evidence regarding the absence of the outcomes (Do=0 and Non=0) does also concern these cases with plans to conduct evaluations, but which have not yet implemented these (Plan=1).

[13] Consistency examines to what extent the empirical evidence is matching the statement of necessity . Technically, this can be computed as "the number of cases with a (1) value on the condition AND a (1) outcome, divided by the total number of cases with a (1) value on the outcome" (Rihoux & Ragin, 2009). Consistency will be 100% if the necessary condition is shared by *all cases* with a particular outcome value.

[14] Coverage of a necessary condition can be calculated as the "number of cases with a (1) value on the condition AND on the outcome, divided by the number of cases with a (1) value on the condition". If the necessary condition is unique for all cases with a particular value, the coverage is 100%.

Table 1: Overview of necessary conditions, with min. consistency of 90%

	Do=1	Do=0 [= Non=1 OR Plan=1]	Non=1	Non=0 [=Do=1 OR Plan=1]
Category A: Capabilities of the organization				
SKINT		**○(0.63)**		
SKEXT	**●(0.69)**			**●(0.77)**
BUDG				
STAFF				
EVCOMM				
MONIT				
Category B: Actor orientations				
DEMMAN				**●(0.90)**
DEMMIN		**○(0.67)**	**○(0.60)**	
DEMPARL				
DEMCSO				
DEMIO				
SUPP				
Category C: Conditions related to the institutional setting				
SIZE				
AUTO		*○(0.45)*		
STATUS				
ANCH				
EPIST		*○(0.40)*	*○(0.32)*	
REGMA			*○(0.40)*	
REGFL		*○(0.38)*	*○(0.31)*	*○(0.69)*
REGINT		*○(0.42)*	*○(0.33)*	
Category D: Policy issue characteristics				
SAL				
COMP				
MEASOP		*●(0.42)*	*●(0.38)*	
MEASOC				
Category E: Conditions related to the path of the organization				
LEG		**○(0.79)**	**○(0.64)**	
ORGSTAB				
MINSTAB				

Symbols within the cells refer to:
●= Necessary condition present, min. 90% consistency
○= Necessary condition absent, min. 90% consistency

These two measures of fit combined, gives an interesting picture. The distribution of the necessary conditions across the various categories does seem in line with the *actor centered* neo-institutionalist approach. The proximate categories of explanation by far account for more non-trivial necessary conditions than the categories situated at a more remote distance from the outcomes of explanation.

Having the skills to outsource an evaluation (Skext=1) appears to be an absolute minimum for these organizations active in evaluations (Do=1) or who have concrete plans to develop them (Plan=1). This may indeed not surprise. Remarkable is that none of the other factors

traditionally identified by evaluation capacity builders turns out to be strictly necessary for the conduct of evaluations. Interestingly, however, evaluation demand from management proves necessary for these cases either conducting or planning evaluations (Non=0), but it does not meet the consistency threshold when only considering the organizations conducting evaluations (Do=1).

While the presence of managerial demand for evaluation seems to be facilitating the planning or conduct of evaluations, the absence of ministerial demand is a major impeding factor to conduct evaluations. Other obstacles revealed are the absence of skills to conduct internal evaluations and the lack of evaluation experience in the pre-NPM era. The latter is the only non-trivial necessary condition of more 'remote' nature. Although not presented in the table, as it only achieved a consistency threshold of 83%, it is worth mentioning that *all* cases which conducted evaluations prior to the implementation of the NPM framework, are still doing so (coverage: 100%). In other words, having previous evaluation experience seems to be a major trigger to continue with evaluations after the introduction of the NPM reforms. In this respect, the implementation of the NPM framework has not had any significant impact, no matter the amount of evaluation practice that the organization's predecessor conducted. From a perspective of evaluation capacity building, this is encouraging. Once an organization has 'tasted' from evaluation practice, this apparently stimulates the hunger to proceed on this track.

The table also lists a large number of trivial necessary conditions, which either scored below the coverage threshold, or which appeared as necessary for both the presence and the absence of a particular outcome. The position of the institution related category, C, is remarkable in this respect. It disproportionally covers most trivial conditions. Especially the lack of participation in evaluation training and networking (Epist), the absence of regulatory requirements for evaluation at Flemish (Regfl), and at international level (Regint) return as trivial. At least in Flanders, these factors will not play a major explanatory role for differences in the extent of evaluation activeness.

For evaluation capacity builders, these findings provide relevant guidelines about the key elements which should inevitably be focused upon. The fact that most necessary conditions are of actor oriented nature is particularly stimulating as these can most easy be changed. Also the trivial conditions are in principle worthwhile to consider in evaluation capacity building.

While they do not have explanatory power to discriminate between organizations conducting evaluations or not, they do nevertheless tell us something about the broader setting in which capacity building can occur.

Note that at this stage, we cannot conclude that the 'non-necessary conditions' have no practical or explanatory value at all. They can still be of relevance from a perspective of MSDO/MDSO, or from a stance of sufficiency.

6.2. Towards a reduced list of conditions with MSDO/MDSO

We deliberately chose to depart from a large set of potentially relevant conditions. Assessing all these conditions would create a risk of coming to just 'individualized' explanations per case, not bringing much analytical insights. A selection of conditions is thus advisable. To make a justified choice of conditions, we rely on the Most Similar Different Outcome/Most Different Similar Outcome (MSDO/MDSO) technique. The latter was originally developed by G. De Meur (1996) as a systematic comparative technique to reduce the complexity of a large data set on the "Inter War Europe Crisis" (see inter alia De Meur and Berg-Schlosser, 1994; De Meur, 1996 for the application of the method on that particular subject). Similar to QCA, its underlying logic is based on the above-mentioned canons of J.S. Mill (1973 [1843]). But rather than focusing on similar/different cases which differ/share *only one* condition, MSDO/MDSO takes a more realistic approach by comparing *'most similar'* and *'most different'* cases (De Meur, 1996; De Meur and Gottcheiner, 2009). As its name suggests, on the one hand, it seeks to identify the relevant variables (and categories of variables) which are capable to explain why 'most similar' cases nevertheless correspond with a different outcome value (e.g. Do=1 versus Do=0). On the other hand, it seeks to point at the relevant variables (and categories of variables) which are capable to explain why 'most different' cases nevertheless correspond with the same outcome value (De Meur, Bursens & Gottcheiner, 2009). Variables identified can be considered as likely candidates with key explanatory potential that can be further examined in subsequent analyses, as csQCA (Rihoux and Ragin, 2009). As such, the technique offers a valuable strategy to deal with the typical 'many variables, small N' dilemma (De Meur & Berg-Schlosser, 1996).

Particularly interesting is that MSDO/MDSO enables a simultaneous consideration of both the depth (intensity) within which pairs of cases are (dis)similar, as well as the extension of

this (dis)similarity (De Meur and Berg-Schlosser, 1994; De Meur, Bursens and Gottcheiner, 2006). After all, cases can be similar in one dimension/category, but dissimilar in another. The method explicitly takes into account these various dimensions, by first establishing similarities and dissimilarities category per category, and then by aggregating these insights. Each category is assigned equal weight, no matter the number or type of variables it contains (De Meur, 1996). In line with our theoretical framework, we logically proceeded with the five categories (A to E), identified above (see also appendix 1).

To come to a comprehensive judgment on (dis)similarity, the technique requires the application of several steps. Within the scope of this paper, we restrict ourselves to a concise presentation of the core tenets. Technical details of the procedure have been more extensively described in e.g. De Meur, 1996; De Meur and Berg-Schlosser, 1996; De Meur, Bursens and Gottcheiner, 2006; De Meur and Gottcheiner, 2009.

For each outcome variable, three pairwise analyses are conducted in parallel, and compared with each other: (1) A comparison of most different cases sharing the 'presence' of a specific outcome (MDSO for Outcome=1), (2) A comparison of most different cases sharing the 'absence' of a specific outcome (MDSO for Outcome=0), (3) A comparison of the cases which are most similar but nevertheless result in different values on the outcome variable (MSDO).

To identify the cases which should be compared, we need to calculate distances (for MDSO) and similarities (MSDO). As a measure of distance, the technique relies on the ‘Boolean distance’. The binary data table (appendix 2) hereto provides the raw material. The distances are simply the number of variables for which two cases differ from each other (per category). Having the Boolean distances computed makes it possible to identify the minimum distance for pairs of cases with different outcomes (MSDO) and the maximum distance for pairs with the same outcome (MDSO) (Rihoux and Ragin, 2009).

With these data at disposal, levels of distance/proximity can be calculated, category per category. These cases which differ most from each other are said to differ at 'level 0' D(0) from each other. ‘Level 1’ D(1) is 1 away from ‘level 0’, whether there is a pair with this value of distance or not. The total number of levels is determined by the threshold level (k), differentiating proximity from remoteness. It is commonly accepted that this threshold is situated at half the number of variables per category (De Meur, 1994). The inverse reasoning

should be followed to calculate the similarity levels S(0) to S(k) between cases for MSDO. Applied to our research, the threshold (k) for (dis)similarity for the different categories are respectively 3 (category A); 3 (category B); 4 (category C); 2 (category D) and 1.5 (category E).

Appendix 3 exemplary presents the different levels of (dis)similarity per type of analysis, for outcome Do. The Boolean distances per pair of cases can be compared with these levels of (dis)similarity, on which basis an aggregated overview can be composed, which presents the levels for the various categories simultaneously. Figure 1 illustrates the levels of (dis)similarity for each pair of cases across the five categories for the analysis of Do. The figure is divided in three zones, which comply with our three analyses mentioned above. Consider for instance the pair of cases 2 and 3 (-13--). For category B, the pair differs at level 1, for category C at level 3. For categories A, D and E the pair can be said not to differ substantially (at least half of the variables are same-valued). The latter categories are marked by a dash (-).

<Figure 1>

For our research purposes, we interpreted the labels 'most similar' and 'most different' rather restrictive, and therefore decided to only single out these pairs of cases which reach levels S(0)/D(0) and/or S(1)/D(1) for the highest number of 'categories' (h). By not to taking into account all possible levels of (dis)similarity (see appendix 3) and by only focusing on the highest number of categories (h), we thus concentrated our analysis on these pairs which are most similar/different in depth, but also in breadth on the highest levels. We as such reduced our focus to the most remarkable pairs, in an attempt to come to the most powerful explanatory variables. Meanwhile, by also including level S(1) and/or D(1), we wanted to take account of possible "troubling aspects of social data" (see above: Ragin, 2000). Our approach is in this respect more conservative than other applications of MSDO/MDSO (e.g. De Meur and Berg-Schlosser, 1994), but in line with several others (e.g. Bursens, 1999; Benijts, 2005).

Once the relevant pairs of cases and categories are selected, we can look for the conditions which matter most in the categories identified. In case of zone 1 (outcome value=1), we look for these conditions on which the most different cases achieve the same value. The same is

done for zone 2 (outcome value=0). In case of zone 3 (outcome values 1 versus 0), we are especially interested in these conditions for which the cases achieve a different value. Appendix 4 presents the overview of pairs of cases, categories and conditions, complying with these selection criteria for outcome Do.

Not all conditions are equally relevant, though. With Bursens (1999) we consider conditions which are mentioned several times within one analysis as more relevant than those only mentioned once. We further consider conditions mentioned in several analyses and matching with the same configuration as more interesting than those mentioned in one analysis; or mentioned in several analyses but matching with different (contradictory) outcomes. We in this respect follow a middle-way approach, between a strict pairwise comparison, and a multiple comparison between various pairs of cases at once (for an example of the latter, see De Meur and Berg-Schlosser, 1994; De Meur and Gottcheiner, 2009). True, a real multiple comparison would yield a more selective set of conditions, yet, it also risks to neglect potentially interesting conditions, which are not shared by all relevant cases. A pairwise comparison more allows the possible existence of multiple paths towards the outcome, which is after all one of the assumptions behind this research.[15]

Table 2 lists the conditions, identified by the MSDO/MDSO analyses for our two outcomes of interest. Note that conditions giving contradictory outcomes might nevertheless be relevant from an analysis of sufficiency perspective, which focuses more on *combinations* of conditions (see below) (Bursens, 1999). We mention these 'ambiguous conditions' separately. The table raises some interesting observations. Starting with the categories' level, the position of the actor-related conditions which featured prominently in the necessity analyses, should somehow be put into perspective. Whereas actor related categories A and B relate more closely to the outcomes of investigation, and are accountable for more necessary conditions, they do not appear most frequently in the MSDO/MDSO analysis. Instead, the categories relating to the path of the organization (category E), and the nature of the particular tasks of the organization (category D) seem better able to discriminate between 'most similar'

[15] When merely focusing on pairwise comparisons, a graphical representation of the mutual relationship between various pairs of cases is not of that importance, as it is the case for multiple comparisons (for an illustration of such graphical representations, see again for instance De Meur and Berg-Schlosser, 1994; De Meur and Gottcheiner, 2009).

organizations, or are better capable to explain why 'most different' cases nevertheless correspond with the same outcome value.

Table 2: Overview of most relevant conditions, as identified in the MSDO/MDSO analyses, for outcomes Do and Non

Identification of the most relevant conditions, resulting in Do=1		
	Category	**Necessary condition?**
SKEXT (1)	A	Yes (coverage: 0.69)
SKINT (1)	A	No
DEMCSO (1)	B	No
DEMIO (0)	B	No
DEMMAN (1)	B	No
ANCH (1)	C	No
STATUS (1)	C	No
COMP (1)	D	No
MEASOP (1)	D	No
LEG (1)	E	No
Identification of the most relevant conditions, resulting in Do=0		
ANCH (0)	C	No
STATUS (0)	C	No
LEG (0)	E	Yes (coverage: 0.79)
Conditions leading to ambiguous configurations for outcome Do		
BUDG	A	No
EVCOMM	A	No
DEMMIN	B	Yes, demmin=0 for do=0 (coverage: 0.67)
MEASOC	D	No
MINSTAB	E	No
Identification of the most relevant conditions, resulting in Non=1		
STAFF (0)	A	No
DEMMIN (0)	B	Yes (coverage: 0.60)
STATUS (0)	C	No
MEASOP (1)	D	Yes (coverage: 0.38)
LEG (0)	E	Yes (coverage: 0.64)
Identification of the most relevant conditions, resulting in Non=0		
SKEXT (1)	A	Yes (coverage: 0.77)
DEMCSO (1)	B	No
STATUS (1)	C	No
LEG (1)	E	No

When zooming into the individual conditions level, the picture gets clearer. First, as for category E, only the pre-NPM evaluation experience of the organization matters. Organizational and ministerial stability have not been identified as most relevant from an MSDO/MDSO perspective. Remarkably further is that the possession of former evaluation experience returns as relevant for the two outcomes, in both their presence and their absence. This is certainly not evident from a configurational point of view, with its asymmetric assumption of causality. This particular situation of symmetry can also be observed for the condition 'status', which refers to the fact whether an organization is a department or an agency.

The importance of both conditions, the status of the organization and its previous evaluation experience, gives an interesting situation. While the necessity analysis made us conclude that the impact of the NPM oriented reforms was limited in the sense that it not had a major impact for these organizations which evaluated before, it constituted an important stimulus to start with evaluations for these organizations that acquired the department status. The sufficiency analysis will more in depth reveal the relative role of both conditions, also in interaction with the others.

The MSDO/MDSO analyses further draw attention to some other conditions with analytical power, which were not identified as relevant in the necessity analysis.[16] The sufficiency analysis should reveal whether these are rightly kept in the analysis. Similarly, the value of the contradictory 'C' conditions, which correspond with ambiguous outcomes, should then become clear. Not all conditions found as necessary *and* non-trivial do play a major role in explaining differences/similarities in evaluation activeness. The absence of evaluation demand from management, which proved necessary for these cases not conducting evaluations (Non=0), and the absence of skills to implement internal evaluations, earlier found necessary for Do=0, are not of prime importance from an MSDO/MDSO point of view.

In sum, comparing the conditions yielded by the MSDO/MDSO analysis of Do and Non gives a partially overlapping picture, but definitely not entirely. This supports us in the choice for a multiple sets approach rather than a single set.

[16] It concerns the demand of civil society organizations for Do=1 and for Non=0, the absence of evaluation demand of other organizations for Do=1, and the perceived measurability of the output of the organization for Do=1 and Non=1.

For practical purposes, the results of the MSDO/MDSO analysis give complementary information to the necessity analysis. Where the latter highlighted these factors which should in principle always be present/absent in a favourable evaluation setting, the MSDO/MDSO analysis gives information on the factors which are *most critical in explaining variation*. Factors which are both necessary ánd appear in the MSDO/MDSO analysis deserve most priority.

6.3. Analysis of sufficiency through a two-steps csQCA approach

So far, our analyses focused on the role of single conditions for the different outcomes, which is yet somewhat at odds with the configurational underpinnings of QCA. We precisely assume that it is the combination of factors which give cases their unique nature, and that different combinations might produce the same outcome ('equifinality'). While the MSDO/MDSO analyses identified the building blocks of likely explanatory relevance per outcome, we are uncertain, though, which combinations can be made of them, that can account for a sufficient explanation. For evaluation capacity building interests, having this information is most fruitful to know which possible 'recipes' can account for a successful/failing outcome. The reduced data tables per outcome form the pivot to proceed to this sufficiency analysis, or minimization process.

The logic of the minimization is again strongly inspired by Mill's method of pairwise difference, which assumes that "if two configurations differ only in one condition but show the same outcome, this distinguishing condition is irrelevant and can be eliminated" (Ragin, 1987). QCA continually applies the pairwise comparison of configurations until the point is reached at which no further minimization can take place. The result of the minimization process is an overview of prime implicants, which can 'imply' the entire set of configurations.

The strength of the 'minimisation' will to large extent depend on the simplicity of the resulting 'solutions'. The latter will in turn strongly be influenced by the empirical diversity that the research has observed. Yet, much social science research is naturally faced by limited diversity, which makes it nearly impossible to find empirical evidence for all logically possible combinations of conditions (Rihoux & Ragin, 2009; Ragin & Sonnett, 2004). This

problem increases (exponentially) with the number of conditions that a research tackles.[17] Nevertheless to achieve more parsimony, QCA provides the possibility to make simplifying assumptions about non-observed remainders of the truth table, for which we hypothesize (through counterfactual analysis) the likely outcome. Logical remainders might differ in theoretical and empirical plausibility, though (Ragin & Sonnett, 2004). We will in particular only include these remainders that are consistent with our knowledge on necessary conditions.[18] Given the embryonic state of affairs of theorizing about differences in evaluation activeness, this position seems most justified. Our main source of reference for the analysis will thus be the type of 'intermediate' solutions, which correspond with a medium level of inference.[19] (Ragin and Sonnett, 2004).

Although we will in the first place rely on these intermediate solutions, the three types of solutions, and their related underlying assumptions, offer us the interesting possibility to make a causal distinction between *core* causal conditions and *peripheral* conditions (Fiss, 2011; Ragin, 2008). With Fiss, we conceive causal coreness in terms of strength of the evidence in relation to the outcome (Fiss, 2011). Core conditions are these that are shared by all three solutions. They can be surrounded by peripheral conditions, which correspond with these conditions that are eliminated in the parsimonious solution, but which appear in the intermediate one. Removing the peripheral conditions would imply the inclusion of 'difficult' counterfactuals, which are more distant from our empirical observations.

As suggested before, we conduct the csQCA analysis in a *two-steps way* (Mannewitz, 2011, Schneider and Wagemann, 2003). The first step will exclusively include the remote conditions. Only if these 'framing' conditions cannot provide a fully sufficient[20] picture, we 'descend' towards the actor-related conditions. Within the scope of this paper, we will only

[17] E.g. with 3 binary conditions, there are 2^3 logically possible combinations (=8). 4 conditions correspond with 2^4 logically possible combinations (=16). 8 conditions with 2^8 possible combinations (=256) etc.
[18] Trivial necessary conditions included.
[19] Recent advances in the QCA software (fsQCA 2.0 in particular) assist the researcher in making its assumptions on the plausibility of the remainders explicit. Based on this input, the software automatically generates three kinds of solutions: complex; intermediate and parsimonious.
[20] As we conceive QCA as a dominantly qualitative approach, we consider it necessary to explain *every* observed case, including possible outliers. In QCA terminology, we strive for a sufficient solution with total consistency and coverage of 100%. In csQCA *consistency* of sufficient conditions refers to the proportion of cases with a certain value that also cause the outcome, in relation to all cases that share the same value. A solution consistency of 1 implies that all cases that display the value display the outcome too (Mannewitz, 2011, Ragin, 2008). *Coverage* is related to the empirical relevance of a solution, and refers to the extent of the cases that is explained by a solution. A perfect solution coverage score of 1 means that all empirical cases are covered by the entirety of solutions.

present the configurations leading to the conduct of evaluations (Do=1) and the non-conduct of evaluations (Non=1), as these are our primary outcomes of interest.[21]

6.3.1. Sufficient paths for the conduct of evaluations (Do=1)

Table 3 visualizes the results of the minimization process, based on the conditions identified in the MSDO/MDSO analysis[22]. The analysis indicates the existence of four distinct configurational groupings corresponding with four different (combinations of) core conditions.[23] This confirms the assumption of equifinality, which emphasizes the idea of several causal paths leading to the outcome. For evaluation capacity building exercises, this implies that there are several scenarios which can all lead to a successful outcome. Equifinality can not only be perceived at the level of the core conditions. Considering solution 1 for instance, we can observe several constellations of peripheral conditions that surround the core condition (1A to 1H). This enables the consideration of various *neutral permutations* within a certain causal path. No matter the particular constellation of peripheral conditions, they all lead to the same outcome. To mark these different levels of substitutability, Fiss (2011) labels them as *first order* and *second order equifinality*.

Although theoretically equivalent, the raw and unique coverage[24] of the configurations are largely different, expressing variety in empirical relevance. The most empirical relevant path, matching with 83% of the cases, revolves around the core condition 'leg', which also featured prominently in the MSDO/MDSO and necessity analyses. Yet, just having the evaluation experience is by itself not sufficient to get a comprehensive explanation. Around the core condition leg we can identify eight different solutions, all of which can account for the conduct of evaluation. Several tendencies appear.

[21] The analyses of the inverses, Do=0 and Non=0 can be requested from the author.
[22] Within the scope of this paper, we only list the 'output' of the minimization process. The full analyses can be requested via the author. Contradictory simplifying assumptions were solved according to the approach presented by Delreux & Hesters (2010).
[23] The different core constellations are indicated with a different number above the columns.
[24] *Raw coverage* refers to the proportion of empirical cases that is covered by a given solution. *Unique coverage* concerns the proportion of cases that are uniquely covered by a given solution (no other solutions cover those cases) (Rihoux & Ragin, 2009)

Table 3: Results of minimization of Do=1. Remote conditions[25]

ANALYSIS OF REMOTE CONDITIONS												
Configurations for the presence of Do (n=18)	Necess. Cond.?	1A	1B	1C	1D	1E	1F	1G	1H	2	3	4
Category C: Conditions related to the institutional setting of the organization												
ANCH		⊖	⊖	•	•	•	•	⊖	•	●	○	○
STATUS		⊖	⊖	•	⊖	•	•	•	⊖	⊖	●	●
Category D: *Conditions related to the tasks of the organization*												
COMP		⊖	⊖		⊖	⊖	⊖	⊖	•	⊖	⊖	●
MEASOP				⊖	•	•	•	•	•	•	•	•
MEASOC		⊖	•	⊖		•	•	•	•	●	⊖	⊖
Category E: Conditions related to the path of the organization												
LEG		●	●	●	●	●	●	●	●			⊖
MINSTAB		•	⊖	⊖	⊖	⊖	⊖	•	•	⊖	○	•
Measures of fit												
Consistency		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Raw coverage		0.22	0.11	0.11	0.11	0.11	0.06	0.06	0.06	0.06	0.06	0.06
Unique coverage		0.22	0.11	0.11	0.11	0.11	0.06	0.06	0.06	0.06	0.06	0.06
Raw coverage per shared core configuration					0.83					0.17	0.06	0.06
Unique coverage per shared core configuration					0.72					0.06	0.06	0.06
Solution coverage: 1.00/ Solution consistency: 1.00												

[25] The notation is based on Fiss (2011) and Ragin (2008). *Full circles:* presence of a condition; *Crossed-out circles:* absence of a condition; *Large circles* indicate core conditions; *Small circles* refer to peripheral conditions; *Blank spaces* refer to a "don't care" situation in which the causal condition may be either present or absent. The second column presents the necessary conditions that were identified in the necessity analysis (if applicable).

First, the extent of competition that the organization experiences from other organizations returns as a key characteristic. All but two configurations are marked by the absence of competition. The cases characterized by a low degree of competition, can be roughly classified in two categories, in which the existence of an evaluation unit seems to matter importantly. These cases having an evaluation cell conduct evaluations, even if there are frequent minister swaps (1C,1D,1E,1F) and on condition that at least the output of the organization is perceived as measurable. Where there is no formal evaluation unit, policy evaluations may still be conducted, where the 'outcome' of the organization is either perceived as measurable, or in a situation of ministerial stability. The latter can be considered as neutral permutations (compare paths 1A and 1B).

Note that in the fewer cases where competition is nonetheless felt (1C and 1H), anchorage of the evaluation function again plays an important role. For agencies in this situation, the conduct of evaluations is facilitated by measurable outputs and outcomes and a stable ministerial environment. The other configurations 2, 3 and 4 do not necessarily require the experience of having conducted policy evaluations. This situation is rarer than the first scenario, as it only accounts for a raw coverage of 29% and a unique coverage of 18%.

As for the second configuration, two core conditions make agencies evaluate, also in a situation of ministerial instability. Again, the establishment of an evaluation unit with the NPM reshuffling seems an essential element in this regard. This establishment is on condition of measurable outputs and outcomes.

Paths 3 and 4 sketch two other, less frequent, alternatives making organizations evaluate without pre-NPM evaluation expertise. The paths apply to departments without an evaluation unit. As suggested before, whereas the introduction of the NPM setting did not have major impact on these organizations which did already evaluate before, it nevertheless seemed an important triggering factor for departments without this experience, which are situated in a setting of measurable outputs and either (a) the absence of ministerial stability as core condition with the absence of competition as peripheral condition and (b) the presence of competition as core condition combined with the presence of ministerial stability as peripheral condition. The fact that they consider their outcomes as difficultly measurable does not seem to make a difference.

To sum up, each of the studied 'remote' categories C, D and E, involves a core condition, which confirms their explanatory relevance. Interestingly, the remote conditions do have sufficient explanatory power to account for a 100% consistent explanation for Do=1. Strictly analytically, there is in principle no need to proceed to the analysis of proximate conditions. It is nevertheless important to recall the necessity of having skills to steer external evaluations as a 'condition sine qua non' for the conduct of evaluations.

6.3.2. Sufficient paths for these cases not conducting evaluations and neither having plans to do this (Non=1)

Proceeding then to the explanations that distinguish these organizations without any intention to evaluate from these that at least plan to evaluate, gives the following minimisation tables (Table 4 &

Table 5: Results of minimization of Non=1. Proximate conditions

STAGE 2: ANALYSIS OF PROXIMATE CONDITIONS			
Configurations for the presence of Non (n=9)	Necess. Cond.?	1A	1B
Category A: Actor capabilities			
SKEXT		○	•
STAFF		⊖	
Category B: Actor orientations			
DEMCSO			○
DEMMIN	○ (0.60)	⊖	○
Measures of fit			
Consistency		1.00	1.00
Raw coverage		0.33	0.67
Unique coverage		0.33	0.67
Raw coverage per shared core configuration		0.33	0.78
Unique coverage per shared core configuration		0.22	0.67
Solution coverage: 1.00/ Solution consistency: 1.00			

5).

Table 4: Results of minimization of Non=1. Remote conditions

STAGE 1: ANALYSIS OF REMOTE CONDITIONS		
Configurations for the presence of Non (n=9)	Necess. Cond.?	1
Category C: Conditions related to the institutional setting of the organization		
STATUS		
Category D*: *Conditions related to the tasks of the organization		
MEASOP	● (0.38)[26]	●
Category E: Conditions related to the path of the organization		
LEG	○ (0.64)	○
Measures of fit		
Consistency		0.69
Raw coverage		1.00
Unique coverage		1.00

[26] The number between brackets refers to the coverage score of the necessary conditions.

Table 5: Results of minimization of Non=1. Proximate conditions

STAGE 2: ANALYSIS OF PROXIMATE CONDITIONS			
Configurations for the presence of Non (n=9)	Necess. Cond.?	1A	1B
Category A: Actor capabilities			
SKEXT		○	•
STAFF		⊖	
Category B: Actor orientations			
DEMCSO			○
DEMMIN	○ (0.60)	⊖	○
Measures of fit			
Consistency		1.00	1.00
Raw coverage		0.33	0.67
Unique coverage		0.33	0.67
Raw coverage per shared core configuration		0.33	0.78
Unique coverage per shared core configuration		0.22	0.67
Solution coverage: 1.00/ Solution consistency: 1.00			

Two general observations can be made: On the one hand, there are fewer paths available and fewer elements at play in the explanation of the non-conduct of evaluation than for explaining the conditions behind the actual conduct of evaluations. On the other hand, in contrast with the situation for Do=1, the remote conditions which we identified in the MSDO/MDSO analysis cannot account for a 100% consistent explanation for Non=1.

Nine cases constitute the observed empirical arena for Non=1. Comparing the values of the cases on the remote conditions yields a single path, marking the absence of evaluation experience of these organizations characterized by easy measurable outputs. These attributes are not unique, however, to these cases, as also expressed by the suboptimal consistency score of 0.69. Only the remote conditions can thus not provide a fully satisfying answer.

To increase the consistency to the ideal level of 100%, the inclusion of our proximate causal conditions is essential. The MSDO/MDSO analysis identified four proximate elements likely at play in the explanation of Non, out of which one with perfect necessity: the absence of any ministerial demand for evaluation. Two paths can be discerned if we compare the proximate conditions in the remote setting identified above. Most empirically important seems to be a combination of the absence of significant ministerial demand with the absence of demand

from civil society organizations active in the field of the organization, even in case the organization is having the skills to outsource and steer external evaluations. A second explanation, capable to explain 33% raw coverage, revolves around the absence of skills to steer external evaluations as core condition, combined with the absence of ministerial demand and staff capable to conduct evaluations as periphery conditions.

7. Conclusion

Comparing the outputs of our analyses, the statement of 'contingent evaluation practices' should be put in perspective. If contingency means that differences in the extent of policy evaluation activeness are context dependent, this could definitely be confirmed. But we could nonetheless discern several cross-case patterns.

First of all, our necessity analyses revealed various necessary conditions with independent effects on the outcome, no matter the particular combination of other conditions with which they are combined. Overall, actor related characteristics (capabilities and orientations) feature as more 'necessary' than characteristics related to the institutional framework surrounding the organization or attributes of the tasks of the organization. As for the more remote conditions, the 'path' of the organization deserves a special mention. These organizations which conducted evaluations prior to NPM, all continued on this élan. For the development of future capacity building strategies this information can be useful. The necessary conditions provide guidelines on the factors that should inevitably be taken into account.

How these necessary conditions should be embedded in a broader strategy has been revealed in our analysis of sufficiency, which focused on the interplay between conditions. Not all conditions were incorporated in the sufficiency analysis. Whereas it has been a deliberate choice to depart from a comprehensive approach to understand differences in evaluation activeness, a systematic pairwise comparison of cases via the MSDO/MDSO technique helped us identifying the factors that matter most in explaining why most similar organizations nevertheless correspond with a different outcome, and different organizations with similar outcomes. From this point of view remote conditions appeared overall more decisive. The MSDO/MDSO analysis reconfirmed the importance of the pre-NPM evaluation experience of the organization, but also brought the status of the organization (department or agency) in the picture. Whereas the implementation of the NPM framework has not had major

impact on the extent of evaluation activeness for these organizations already active in evaluation before, it has nonetheless served as an important trigger for departments without that expertise. In that respect, the influence NPM had on the extent of evaluation activeness is mixed. From the actor related conditions, evaluation demand of civil society organizations, inter-organizational demand, demand of the minister and organizational management, internal organizational skills to evaluate, and the availability of staff for evaluation appeared as relevant from an MSDO/MDSO point of view.

Yet, none of these conditions proved sufficient on its own to account for a full explanation. The analyses of sufficiency revealed several paths for each outcome of interest. These can be read as potential recipes for evaluation capacity builders. Overall, we could identify much more scenarios which explain why organizations evaluate, than why do they not evaluate. On the other hand, though, for these cases which do conduct evaluations, a fully consistent explanation could already be achieved on the mere basis of the remote conditions. The contingent reality of evaluation activeness could thus be reduced to a selected list of configurations. Moreover, for each of the outcomes, Do=1 and Non=1, we could find a single path capable to explain more than 75% of the variation.

The research strived for a medium level of generalization, by taking into account the logical remainders consistent with our necessary conditions. Future research should investigate which other, more difficult, remainders can potentially be included. Similarly, also the scope of external validity of our findings should be assessed.

Overall, the analysis is a first attempt to bring more systematic insights on the dynamics behind organizational differences in evaluation activeness on the basis of configurational comparative methods. For this reason, we chose for a rather basic conceptualization of the extent of evaluation activeness (the mere presence/absence of evaluation practices), and opted to depart from a large set of potentially interesting conditions. Further studies preferably also include more complex aspects of evaluation activeness, and scrutinize a selected number of conditions more in depth.

Acknowledgements:

I would like to thank Eva Platteau, Christopher Pollitt, Maarten Vink, and an anonymous reviewer for their useful comments. A previous version of this paper was presented at the 2011 American Evaluation Association Conference, Anaheim, with support of Academische Stichting Leuven.

Bibliography

Benijts, Tim (2005). *De keuze van beleidsinstrumenten: Een vergelijkend onderzoek naar duurzaam sparen en beleggen in België en Nederland.* Proefschrift ter verkrijgen van de graad van Doctor in de Sociale Wetenschappen. K.U.Leuven: Instituut voor de Overheid.

Berg-Schlosser, D., De Meur G., Rihoux B., & Ragin C.C.(2009). Qualitative Comparative Analysis (QCA) as an approach. In B. Rihoux & C.C. Ragin (Eds.), *Configurational Comparative Methods*, London: Sage.

Bressers, J. Th. A., & Hoogerwerf, A. (1991). Inleiding tot de beleidsevaluatie. In J. Th. A. Bressers & A. Hoogerwerf (eds.) *Beleidsevaluatie.* 2de druk. Alphen aan den Rijn: Samson H.D. Tjeenk Willinnk.

Bursens, P. (1999). *Impact van instituties op besluitvorming. Een institutioneel perspectief op besluitvorming in de communautaire pijler van de Europese Unie.* Proefschrift voorgelegd tot het behalen van de graad van Doctor in de Politieke en Sociale Wetenschappen. Antwerpen: UA.

Byrne, D. (2009). Introduction. In D. Byrne, & C.C. Ragin (Eds.), *The Sage Handbook of Case-Based Methods*. London: Sage.

Delreux, T. & Hesters, D. (2010*).* Solving contradictory simplifying assumptions in QCA: presentation of a new best practice. *COMPASSS Working Paper, 58.* Retrieved from www.compasss.org

De Meur, G. (1996). La comparaison des systèmes politiques: recherche des similarities et des differences. *Revue Internationale de Politique Comparée, 3*(2), 405-437.

De Meur, G. & Berg-Schlosser, D. (1994). Comparing political systems: Establishing similarities and dissimilarities. *European Journal of Political Research, 26*, 193-219.

De Meur, G. & Berg-Schlosser, D. (1996). Conditions of Authoritarianism, Fascism and Democracy in Inter-War Europe: Systematic Matching and Contrasting of Cases for "Small N" Analysis. *Comparative Political Studies, 29(*4*):* 423–468 .

De Meur, G., Bursens P., & Gottcheiner, A. (2006). MSDO/MDSO Revisited for Public Policy Analysis. In B. Rihoux & H. Grimm (Eds.), *Innovative Comparative Methods for Policy Analysis. Beyond the Quantitative-Qualitative Divide*. New York: Springer.

De Meur, G. & Gottcheiner, A. (2009). The logic and assumptions of MDSO/MSDO designs. In D. Byrne & C.C. Ragin (Eds.), *The Sage Handbook of Case-Based Methods*, London, Sage, pp. 208-221.

De Peuter B. & Pattyn V. (2009). Evaluation capacity: enabler or exponent of evaluation culture. In A. Fouquet & L. Méasson (Eds.), *L'évaluation des politiques publiques en Europe. Cultures et futures.* Paris: l'Harmattan.

Eliadis P., Furubo J.E. & Jacob S. (Eds.) (2011). *Evaluation. Seeking truth or power? , Comparative Policy Evaluation Volume 17.* New Brunswick: Transaction Publishers.

Fierro, L. (2011). *Making Sense of "Capacity" in the Evaluation Process by Leveraging Existing Theories and Frameworks*. Presentation held at the American Evaluation Association Conference. California: Anaheim.

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

Guerrero, P. (1999). Evaluation capacity development. Comparative insights from Colombia, China, and Indonesia. *ECD Working Paper Series, 5*. World Bank: Operations Evaluation Department.

Goertz, G. (2006). Assessing the trivialness, relevance, and relative importance of necessary or sufficient conditions in social science. *Studies in Comparative International Development* 41(2), 88-109.

Gray, A. & Jenkins, B. (2011). Policy and evaluation: Many powers, many truths. In P. Eliadis, J.E. Furubo, & Steve Jacob (Eds.), *Evaluation. Seeking truth or power?, Comparative Policy Evaluation Volume 17*, New Brunswick: Transaction Publishers.

Henry, G. T. & Mark, M. (2003). Toward an agenda for research on evaluation. *New Directions for Evaluation, 97*, 69-80.

IAVA (2007). *Jaarverslag van het Auditcomité en het Agentschap Interne Audit van de Vlaamse Administratie*. Retrieved from http://www2.vlaanderen.be/doelbewustmanagement/jaarverslag_2007.pdf.

Kaeding, M. (2006). In good times and bad: Legal transposition in the European Union. Assessing correlational and necessary/sufficient causation. *German Working Papers in Law and Economics, 29*.

Mackay, K. (1999). Evaluation capacity development: A diagnostic guide and action framework. *ECD Working Paper, 6*. World Bank: Operations Evaluation Department.

Mannewitz, T. (2011). Two-level theories in QCA: A discussion of Schneider and Wagemann's Two-step approach. *COMPASSS working paper, 64*. Retrieved from www.compasss.org

Meyer, W. & Stockmann, R. (2007). Comment on the paper: An evaluation tree for Europe. In G.J. Peterson & O.K. Vestman (Eds.), *Conceptions of evaluation, Rapport 08/2007*. Härnösand: NSHU.

Mill, J.S. (1973 [1843]). Of the Four Methods of Experimental Inquiry, chapter VIII. In *The Collected Works of John Stuart Mill* (Vol. VII - A System of Logic Ratiocinative and Inductive). London: Routledge and Kegan Paul.

Ragin, C.C. (1987). The comparative method. Moving beyond qualitative and quantitative strategies. London: University of California Press.

Ragin, C.C.(2000). *Fuzzy set social science*. Chicago: University Chicago Press.

Ragin, C.C. (2006). Set relations in social research: evaluating their consistency and coverage. *Political Analysis 14*(3), 291-310.

Ragin, C.C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond.* Chicago: University Chicago Press.

Ragin, C.C. & J. Sonnett (2004). Between Complexity and Parsimony: Limited Diversity, Counterfactual Cases and Comparative Analysis. In S. Kropp & M. Minkenberg (Eds.), *Vergleichen in der Politikwissenschaft*. Wiesbaden: VS Verlag fur Sozialwissenschaften.

Rihoux, B. & Lobe, B. (2009). The Case for Qualitative Comparative Analysis (QCA): Adding Leverage for Thick Cross-Case Comparison. In D. Byrne & C.C. Ragin (Eds.), *The Sage Handbook of Case-Based Methods*. London: Sage.

Rihoux, B. & Ragin, C.C. (2009). *Configurational comparative methods. Qualititative comparative analysis (QCA) and related techniques.* Thousand Oaks and London: Sage.

Scharpf, F.W. (1997). *Games real actors play: actor centered institutionalism in policy research*. Oxford: Westview Press.

Scharpf, F.W. (2000). Institutions in comparative policy research. *Max Planck Institut fur Gesellschaftsforschung WP 00/3.*

Schmidt, V. (1993). The boundaries of 'bounded generalizations: Discourse as the missing factor in actor-centered institutionalism. In R. Mayntz & W. Streeck (Eds.), *Die Reformierbarkeit der Demokratie: Innovationen und Blockaden: Festschrift für Fritz W. Scharpf,*. Frankfurt: Campus.

Schneider, C.Q. & Wagemann C. (2003). Improving Inference with a Two-step Approach: Theory and Limited Diversity in fs/QCA. *EUI Working Papers. 2003/7.* European University Institute: San Domenico di Fiesole.

Scott, R. W. (1995). *Institutions and organizations.* Thousand Oaks: Sage.

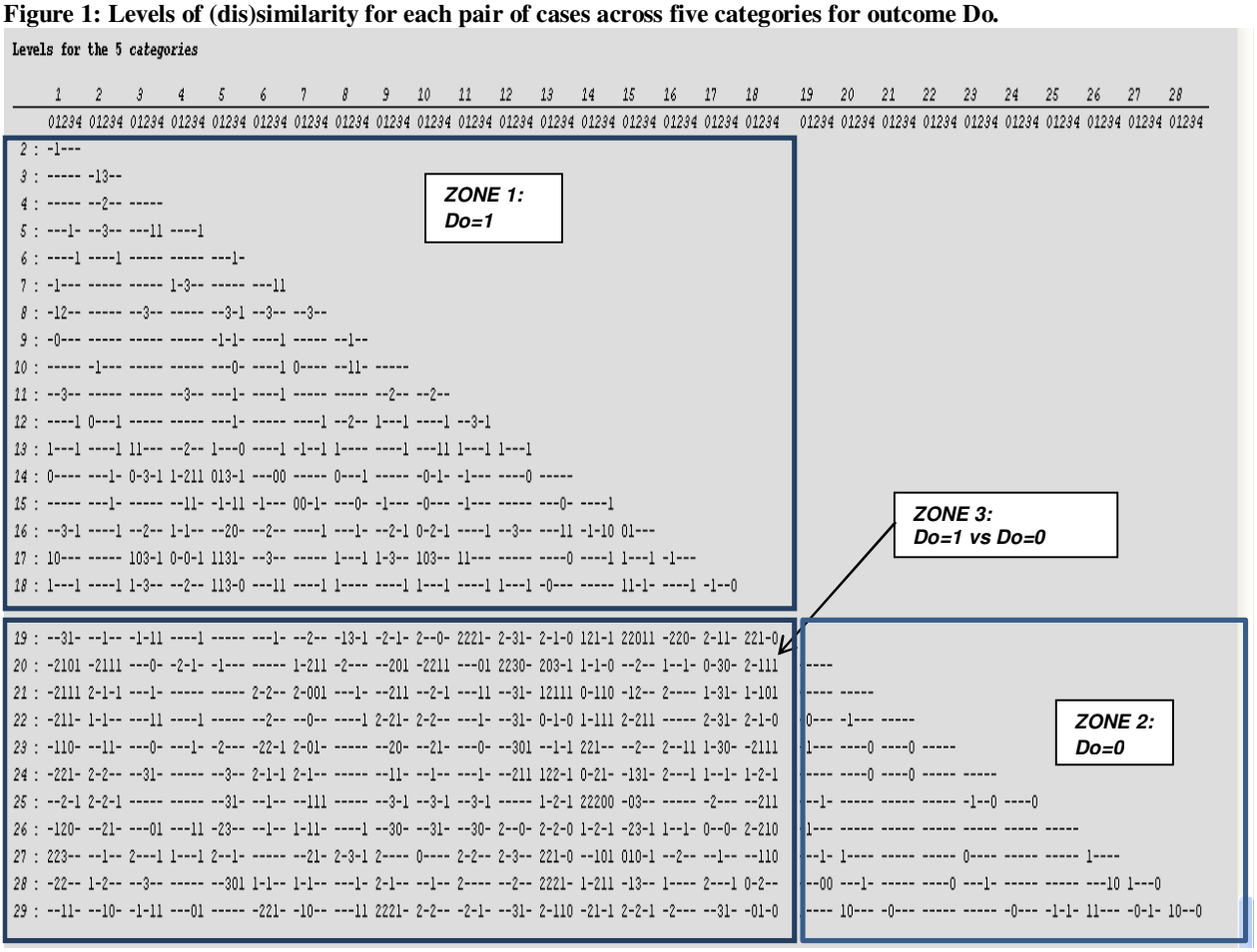
Segerholm, C. (2003). Researching evaluation in national (state) politics and administration: A critical approach. *American Journal of Evaluation, 24,* 357-372.

Van Lieshout, H.A.M. (2008). *Different hands. Markets for intermediate skills in Germany, the U.S. and the Netherlands.* Proefschrift ter verkrijging van de graad van doctor. Utrecht: Universiteit Utrecht.

Widmer, T. (2005). Instruments and procedures for assuring evaluation quality: A Swiss perspective. In R. Schwartz & J. Mayne (Eds.). *Quality matters. Seeking confidence in evaluating, auditing and performance reporting, Comparative Policy Evaluation, Volume XI*, New Brunswick & London: Transaction Publishers.

Witte, J. (2006). *Change of degrees and degrees of change. Comparing adaptations of European Higher Education Systems in the context of the Bologna Process*, Dissertation to obtain the doctor's degree. Twente: University of Twente.

Figure 1: Levels of (dis)similarity for each pair of cases across five categories for outcome Do.



Output generated by the MSDO/MDSO software (version 8/7/2006), developed by G. De Meur (available via http://www.jchr.be/01/beta.htm)

Appendix 1: Tresholds used for the crisp set coding of the conditions

List of conditions	Code	Tresholds for binary coding
Category A: Capabilities of the organization		
Internal skills to evaluate	SKINT	*-Code 0: Totally insufficient; Rather insufficient* *-Code 1: Rather sufficient, Fully sufficient*
External skills to evaluate	SKEXT	*-Code 0: Totally insufficient; Rather insufficient* *-Code 1: Rather sufficient, Fully sufficient*
Budget	BUDG	*-Code 0: Strong budgetary hindrance; Rather budgetary hindrance* *-Code 1: Rather no budgetary hindrance; No budgetary hindrance*
Availability of staff	STAFF	*-Code 0: Totally insufficient; Rather insufficient* *-Code 1: Rather sufficient, Fully sufficient*
Availability of evaluation community	EVCOMM	*-Code 0: Totally insufficient; Rather insufficient* *-Code 1: Rather sufficient, Fully sufficient*
Extent of development of monitoring system	MONIT	*-Code 0: Totally insufficient; Rather insufficient* *-Code 1: Rather sufficient, Fully sufficient*
Category B: Orientations		
Evaluation demand from management	DEMMAN	*-Code 0: No demand; Hardly any demand* *-Code 1: Sometimes demand, Frequent demand*
Evaluation demand from minister(ial cabinet)	DEMMIN	*-Code 0: No demand; Hardly any demand* *-Code 1: Sometimes demand, Frequent demand*
Evaluation demand from parliament	DEMPARL	*-Code 0: No demand; Hardly any demand* *-Code 1: Sometimes demand, Frequent demand*
Evaluation demand from civil society organizations	DEMCSO	*-Code 0: No demand; Hardly any demand* *-Code 1: Sometimes demand, Frequent demand*
Evaluation demand from other organizations	DEMIO	*-Code 0: No demand; Hardly any demand* *-Code 1: Sometimes demand, Frequent demand*
Organizational support for evaluations	SUPP	*-Code 0: Not at all; To limited extent* *-Code 1: To major extent; To large extent*
Category C: Conditions with regard to the institutional setting		
Size of the organization	SIZE	The indicator concerns both financial material weight (weight: 50%) and material weight with regard to personnel (weight: 50%). -For financial material weight, the following scales are used [in 10 000EUR]: (1) very low material weight: 0-50000; (2) low material weight: 10000-50000; (3) average material weight: 50000-100000; (4) high material weight: 100000-500000; (5) very high material weight: 500000. -As for material weight with regard to personnel, in staff numbers per organization: (1) very low: 0-100; (2) low: 101-200; (3) average: 201-400; (4) high: 401-900; (5) very high: 900. We calculated the average of these two dimensions for the years 2007-2008-2009. *-Code 0: Very low or low material weight* *-Code 1: At least average material weight*
Autonomy of the organization	AUTO	*-Code 0: No legal personality* *-Code 1: Legal personality*
Status of the organization	STATUS	*-Code 0: Agency* *-Code 1: Department*
Anchorage of evaluation function	ANCH	*-Code 0: No evaluation unit* *-Code 1: Formal/de facto evaluation unit*
Participation in epistemic evaluation community	EPIST	Proxies: (1) Extent of engagement in formal networks with a focus on M&E (2) Extent of engagement in training activities with a focus on M&E *-Code 0: No engagement in evaluation networks/training* *-Code 1: When minimally 'sometimes' participating at sectoral trainings/networking on evaluation.*
Regulatory evaluation requirements[27]	REGFL; REGINT;	*-Code 0: No evaluation requirements* *-Code 1: Evaluation requirements*

[27] It was deemed relevant to examine the specific influence of three types of regulatory evaluation requirements: (1) REGFL: These requirements stipulated in regulation or decrees at Flemish level – management agreements

	REGMA	
Category D: Conditions concerning policy issue characteristics		
Salience	SAL	Proxies: Maximum value of proxy 1 (Media attention for tasks of the organization) and proxy 2 (Parliamentary attention for tasks of the organization). *-Code 0: Not at all; Limited; Rather limited* *-Code 1: Rather much, much; very much.*
Competition on tasks of the organization	COMP	-Code 0: Not at all; Limited; Rather limited -Code 1: Rather strong; Strong; Very strong
Perceived measurability of outputs	MEASOP	Proxy: average score of measurability on a scale of 1 (very difficult to measure) to 5 (very easy to measure) of the three most important outputs of the organization. *-Code 0: Average score ≤3 and/or qualification:* *Very difficult; Difficult; Rather difficult* *-Code 1: Average score ≥3 and/or qualification: Very easy; Easy; Rather easy*
Perceived measurability of outcomes	MEASOC	Proxy: average score of measurability on a scale of 1 (very difficult to measure) to 5 (very easy to measure) of the three most important outcomes of the organization. *-Code 0: Average score ≤3 and/or qualification:* *Very difficult; Difficult; Rather difficult* *-Code 1: Average score ≥3 and/or qualification: Very easy; Easy; Rather easy*
Category E: Conditions characterizing the 'path' of the organization		
Pre-NPM evaluation experience	LEG	*-Code 0: No/Seldom evaluation practice prior to the NPM inspired reforms* *-Code 1: Sometimes/Frequent evaluation practice prior to the NPM inspired reforms*
Organizational stability	ORGSTAB	Four subcriteria constitute this indicator. Three of them relate to the impact of the NPM-oriented reforms (which account for 60% of the indicator in total): (1) changes in the form of management/steering of the organization; (2) changes with regard to the composition of the public entity; (3) changes with regard to the organization of the management support services. The remaining 40% of the indicator refers to changes independent of the NPM reforms. Based on the sum of these subcriteria, a scale can be composed ranging from 0.1 to 0.5, with 0.5 standing for these organizations which underwent a large number of changes; 0.3 for these which underwent a medium number of changes and 0.1. for these organizations which can be characterized by large stability. Data have been retrieved from IAVA, 2007. *-Code 0: Organizations which underwent medium or large changes* *-Code 1: Organizations which underwent no or small changes*
Ministerial stability	MINSTAB	*-Code 0:* *≥ 1 Minister change since the introduction of the NPM inspired reforms* *-Code 1: No minister change since the implementation of NPM*

excluded- (2) REGINT; requirements for evaluation stipulated in international regulation (EU; OECD…) (3) REGMA requirements for evaluation stipulated in organization's management agreements.

Appendix 2: Binary data table

CODE CASES[28]	SKINT	SKEXT	BUDG	STAFF	EVCOMM	MONIT	DEMMAN	DEMMIN	DEMPARL	DEMCSO	DEMIO	SUPP	SIZE	AUTO	STATUS	ANCH	EPIST	REGMA	REGFL	REGINT	SAL	COMP	MEASOP	MEASOC	LEG	ORGSTAB	MINSTAB	DO	NON
INTA7	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1
INTA6	0	1	1	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1
INTA1	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	1
EXTA4	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1	1	0	0	1	0	0	1
INTA3	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1
INTA8	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1
DEPT4	0	1	1	1	1	1	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	1	1	0	0	0	0	0	1
EXTA5	1	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	1	0	1	1	0	1
INTA4	0	0	1	0	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	1
DEPT10	0	1	0	1	1	0	1	0	0	1	1	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
DEPT9	0	1	0	0	0	0	1	1	0	0	0	0	1	0	1	0	0	1	0	0	1	0	1	0	0	0	1	0	0
INTA2	1	1	1	1	0	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	1	0	1	0	1	1	0
DEPT7	1	1	1	0	1	0	1	1	1	1	0	0	1	0	1	0	0	0	0	0	1	0	1	1	1	0	1	1	0
INTALP3	1	1	1	1	0	1	1	0	0	1	1	1	1	1	0	1	0	1	0	1	1	0	1	0	1	0	0	1	0
INTA9	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0	0	1	1	0	1	1	0	1	1	1	0	0	1	0
ILTALP1	1	1	1	1	0	1	1	1	0	0	0	1	0	1	0	1	0	1	0	0	0	1	1	1	1	1	1	1	0
EXTA6	1	1	0	0	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	0	1	0	0	1	1	1	0	1	0
INTA5	1	1	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	1	0
DEPT3	1	1	1	1	0	1	1	1	0	1	1	0	0	0	1	1	1	0	0	1	0	0	1	1	1	0	0	1	0
EXTA2	1	1	1	0	1	1	1	1	1	1	1	0	1	1	0	0	0	1	0	0	1	0	1	0	1	0	1	1	0
EXTA3	0	1	1	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1	0	0	1	0	0	0	1	0	1	1	0
DEPT11	1	1	0	1	1	1	1	1	0	1	1	1	0	0	1	1	0	0	0	1	1	0	1	0	1	0	1	1	0
ILTALP2	0	1	0	1	0	1	1	1	0	1	0	1	1	1	0	1	0	1	0	0	1	0	1	0	1	1	0	1	0
EXTA1	0	1	1	0	1	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	1	0
DEPT1	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	0	1	1	0	0	0	1	1	0
DEPT5	0	1	1	1	1	1	0	0	0	1	0	0	1	0	1	1	0	0	0	0	1	1	0	0	1	0	0	1	0
DEPT2	1	1	0	0	0	0	1	1	0	1	1	1	1	0	1	1	1	0	1	0	1	0	0	0	1	1	0	1	0
DEPT8	0	1	0	0	0	0	0	1	1	1	0	0	1	0	1	1	0	0	1	0	1	0	1	0	1	1	1	1	0
DEPT6	1	1	0	0	1	0	1	0	1	1	1	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0

[28] The *codes* are composed by a letter combination and a number. The letter code refers to the different statutes of the organization as prescribed by the Framework Decree of the NPM oriented reforms. DEPT stands for Department; EXTA for external autonomized bodies of public nature; INTA for internal autonomized bodies without legal personality; INTALP for internal autonomized bodies with legal personality. The accompanying identification number has been at random given to each entity.

Appendix 3- Figure 1a: Levels of dissimilarity for pairs of cases sharing outcome Do=1

Level D(0)	
5 *(cat. A)*	4 *(cat. D)*
5 *(cat. B)*	3 *(cat. E)*
8 *(cat. C)*	

Level D(1)	
4	3
4	2
7	

Level D(2)	
3	2
3	2*[29]
6	

Level D(3)	
3*	2*
3*	2*
5	

Level D(4)	
3*	2*
3*	2*
4	

Appendix 3- Figure 1b: Levels of dissimilarity for pairs of cases sharing outcome Do=0

Level D(0)	
5	4
5	2
-	

Level D(1)	
4	3
4	2*
-	

Level D(2)	
3	2
3	2*
-	

Appendix 3- Figure 1c: Levels of similarity for pairs of cases with different outcomes on outcome Do

Level S(0)	
0	0
0	0
1	

Level S(1)	
1	1
1	1
2	

Level S(2)	
2	2
2	1*
3	

Level S(3)	
3	2*
3	1*
4	

[29] If the threshold level in a particular category is already obtained at an earlier stage, this value is retained for the subsequent levels (De Meur and Berg-Schlosser, 1994).

Appendix 4: Overview of pairs of cases, categories and conditions complying with the MSDO/MDSO selection criteria, for outcome Do

Pairs number[30]	Organisations	Level 0	Level 1[31]	Level 0	Level 1
		Categories	Categories	Conditions	Conditions
ZONE 1: Do=1					
Level D(0)&D(1)					
7,15	INTA5- DEPT5	A,B	D	SKEXT (1), DEMCSO (1)	MEASOC (0)
4,17	INTA9-DEPT8	A,C	E	SKEXT (1)	LEG (1)
Level D(0)					
6,14	EXTA6-EXTA4	D,E		/	
Level D(1)					
4,14	INTA9- DEPT1		A,D,E		SKEXT (1), EVCOMM (1), MEASOP (1), ORGSTAB (0)
5,14	INTALP1-DEPT1	A	B,E	SKEXT (1)	DEMMAN (1), DEMIO (0), MINSTAB (1)
5,15	INTALP1- DEPT5		B,D,E		DEMPARL (0), DEMIO (0), COMP (1), LEG (1)
14,16	DEPT1- DEPT2	E	B,D	/	DEMMAN (1), DEMCSO (1), MEASOC (0)
3,17	INTALP3- DEPT8	B	A,E	DEMCSO (1)	SKEXT (1), LEG (1)
5,17	INTALP1-DEPT8		A,B,D		SKEXT (1), EVCOMM (0), DEMMIN (1), DEMIO (0), MEASOP (1)
5,18	INTALP1- DEPT6	E	A,B	/	SKINT (1), SKEXT (1), DEMMAN (1), SUPP (1)
15,18	INTALP1- DEPT6		A,B,D		SKEXT (1), EVCOMM (1), DEMMIN (0), DEMCSO (1), MEASOC (0)
ZONE 2: Do=0					
Level D(0)&D(1)					
28,29	EXTA5- INTA4	B,E	A	DEMMIN (0), LEG (0)	STAFF (0), EVCOMM (1)
Level D(0)					
19,28	DEPT10- EXTA5	D,E		LEG (0)	

[30] Corresponding with Figure 1.

[31] To have a full understanding of these cases which reach the highest level of (dis)similarity D(0) or S(0) for the highest number of categories, we also took into account these categories on which these particular cases reached level D(1) or S(1).

ZONE 3: Do=1 vs Do=0					
Level S(0)&S(1)					
14,21	DEPT1- INTA7	A,E	C,D	/	STATUS (1) vs STATUS (0), COMP (1) vs COMP (0)
15,27	DEPT5- DEPT4	A,C	B,E	/	DEMCSO (1) vs DEMCSO (0), LEG (1) vs LEG (0)
Level S(0)					
17,20	DEPT8- DEPT9	A,D		/	
7,21	INTA5- INTA7	C,D	E	/	LEG (1) vs LEG (0)
13,22	EXTA1- INTA6	A,E	C	/	ANCH (1) vs ANCH (0)
14,25	DEPT1- INTA3	D,E		/	
17,26	DEPT8- INTA8	A,D		/	
18,29	DEPT6- INTA4	B,E	C	/	STATUS (1) vs STATUS (0)
Level S(1)					
13,21	EXTA1- INTA7		A,C,D,E		BUDG (1) vs BUDG (0), ANCH (1) vs ANCH (0), MEASOC (1) vs MEASOC (0), MINSTAB (0) vs MINSTAB (1)
18,21	DEPT6- INTA7	D	A,C,E	/	SKINT (1) vs SKINT (0), STATUS (1) vs STATUS (0), MINSTAB (0) vs MINSTAB (1)
14,22	DEPT1- INTA6		A,C,D,E		BUDG (0) vs BUDG (1), STATUS (1) vs STATUS (0), SAL (0) vs SAL (1), MINSTAB (1) vs MINSTAB (0)

The same overview for outcome Non is available from the author upon request.