Explaining organisational variety in evaluation quality assurance. Which conditions matter?

This is the version of the article that was accepted for publication by the journal. The final, definitive article has been published in the International Journal of Public Administration by Taylor and Francis, all rights reserved.

Pattyn, V., Brans, M. (2014). Explaining organisational variety in evaluation quality assurance. Which conditions matter? *International Journal of Public Administration*, 37 (6), 363-375.

Abstract:

The growth in evaluation practice has not always gone hand in hand with an increase in the actual quality of these evaluations. This article addresses this concern, by examining the factors that explain organisational variety in the application of evaluation quality assurance measures. We present the results of an analysis of eighteen Flemish (Belgian) public sector organisations. To unravel different explanatory trajectories, we combine two comparative methods that rely on Boolean analysis: the most similar different outcome/most different similar outcome technique on the one hand, and crisp set qualitative comparative analysis on the other hand.

Key words:

policy evaluation, quality assurance, QCA, Boolean methods

Acknowledgements:

The authors would like to thank the two anonymous reviewers for the helpful comments.

The growth in evaluation practice worldwide has not been matched by an equal growth in the quality of evaluations. Several studies are pessimistic about the quality of evaluation reports (e.g. Forss et al., 2008; Johnsen et al., 2001; Borrman & Stockmann, 2009). Evaluative information would even face a credibility crisis (Schwartz & Mayne, 2005). This negative diagnosis is surprising, especially in light of the many efforts to develop evaluation standards and quality assurance models. To be sure, evaluation quality assurance measures are not everywhere applied to the same extent. Yet, little is known about the conditions that promote or impede the application of evaluation quality assurance mechanisms (Schwartz & Mayne, 2005). The literature offers fragmented insights only, that are mostly of anecdotal nature and have seldom been tested systematically. This article aims to fill this empirical void, via systematically scrutinizing the evaluation quality assurance practices of eighteen public sector organizations in Flanders (Belgium).

It pursues two objectives. First, we test the explanatory relevance of a large number of conditions that might possible explain organisational variety in the extent of evaluation quality assurance. We identified these conditions in the evaluation literature and through explorative interviews . Second, we unravel the combinations of conditions that can explain high evaluation quality assurance at organisational level. We consider it not very plausible to come to any universal law-like formula for explaining variety in evaluation quality control. But we assume that it is nonetheless possible to find patterns behind similarities and

differences. Systematic insights into the conditions of the usage of evaluation quality arrangements can moreover be useful for practice that seek professional development, and advocate better evaluation policies (Cooksy & Mark, 2011; Cooksy & Caracelli, 2005). The quality with which evaluations are conducted is after all considered to be one of the determinants of their use (Leviton & Hughes, 1981; Cooksy & Caracelli, 2005; European Commission, 2012; Schwartz & Mayne, 2005). In the following section, we explain how we conceive the outcome variable of this research, i.e. the extent of evaluation quality assurance by public sector organizations. Next, we present the state of evaluation praxis of the Flemish public sector, our case area. In a subsequent part, we discuss the methodology. We present the data collection strategy, we clarify the choice for an explorative approach, and we introduce the two Boolean methods used (Most Similar Different Outcome/Most Different Similar Outcome technique (MSDO/MDSO) and Qualitative Comparative Analysis (QCA)). The remainder of the article, then, is organized around the two research questions. With the MSDO/MDSO technique, we identify the conditions that are most relevant to explain organizational variety in evaluation quality assurance. With QCA, we examine how the conditions, identified as having most explanatory potential, interact. The conclusion reiterates our findings and formulates suggestions for further research.

Evaluation quality assurance by public sector organizations

From the many possible conceptualizations of evaluation, the present research deals with ‘policy’ evaluations. We define these 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.)’ (based on Scriven, 1991). With the notion scientific, we refer to the application of science-based methods. The definition does not imply that an evaluation should be conducted by scientists. Rather, it stresses the research-

based link of evaluation. It highlights the fact that evaluation provides a systematic and transparent assessment of an object. The evaluations of our interest pertain to a large array of possible evaluanda, including evaluations on the content of the policy, the process, or the effect, and impact of the policy.

While evaluators share a commitment to high quality evaluations, they do not always agree on what this exactly entails. As Benson, Hinn, & Lloyd (2001: ix) stated: “There may be as many definitions of quality as there are interpretations of experience”. Evaluators work in a large variety of settings, and often have different values and beliefs about what is important in their work (Cooksy & Mark, 2011). As a result of this, the evaluation literature is full of guidelines and checklists (a selective overview can be consulted at Mande, 2011), often highlighting different elements (Widmer, 2005). Cooksy & Caracelli (2005: 34-35) compared various evaluation standards and made an overview of the large range of quality criteria that circulate in the evaluation literature. They identified criteria as utility, feasibility, propriety, accuracy, transparency, relevance, balance, credibility, legitimacy, cultural competence, ... all given a different emphasis due to different expectations of an evaluation. As Benson et al. (2001: ix) stated: “A complex array of contingencies- in both local contexts and larger socio-political systems- shape stakeholders’ particular experiences of program quality”.

In this article, we approach evaluation quality from an organisational stance. Rather than assessing the quality of individual evaluation reports, we focus on the organisational system within which the evaluation process occurs. By approaching quality from the perspective of the evaluation quality assurance system of the organisation, we emphasize the *intention* to deliver high quality reports, rather than the actual quality. The eventual quality after all strongly depends on numerous intervening factors, such as for instance the willingness and ability of external evaluators to produce high quality reports. In this contribution, we will focus on a set of quality criteria that we assume to be always important,

irrespective of context and purpose of the evaluation, and of the evaluation model. In the methodology section of this article, we will explain the specific quality criteria used.

Policy evaluation in Flanders

Furubo and Sandahl (2002) identified two different waves of diffusion of evaluation in Europe, respectively taking place in 1960's -1970's and end 1990's. The first group of countries (including Sweden, Germany, UK) adopted evaluation mainly as a package following internal and domestic pressures. For the larger, second group of countries, in turn, external factors were the main motor for diffusion. The introduction of New Public Management (NPM), with the associated promotion of policy evaluation as an accountability tool (Van de Meer, 2007), and the pressure of international institutions as the European Union were major catalysts for change. Belgium is clearly a member of this second group of countries (Jacob, Varone, & De Winter, 2005).

The present study focuses on the Flemish public sector, i.e. the administration of the Flemish Region and the Flemish Community. Whereas the Flemish Region has powers in territory-related fields (such as environment and agriculture), the Flemish Community has language-based competencies (such as education and culture). The Flemish administration is a particularly interesting research object, as it relatively recently (in 2006) underwent a major reshuffling following the principles of NPM. With these reforms, policy evaluation became formally anchored in the task prescription of departments and agencies. Departments are charged with the evaluation of instruments, outputs and outcomes of policy implementation. Agencies, in turn, are required to provide the necessary input for these evaluations (Vlaams Parlement, 2003). Triggered by the reforms, evaluation is now widely practiced by departments and agencies, albeit still major discrepancies exist between policy fields. For our purposes, the study of Flanders is relevant, as it can shed light on the conditions that foster or

impede the implementation of quality assurance in public sector organizations modelled along the NPM logic. Our findings can be instructive for other second wave countries, in which evaluation practices are also recently introduced. Besides, our analysis offers an alternative to the studies on evaluation quality assurance that exclusively focus on the world leaders in the practice of evaluation (e.g. Schwartz & Mayne, 2005).

Methodology

Data collection

As for the selection of the specific public organisations, we approached the departments and autonomized agencies of all policy fields, with a vertical focus: (1) education; (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. We first determined which entities were active in evaluation, based on interviews and document analysis (i.e. the scrutiny of evaluation reports). Organisations that met this criterion were asked to complete a survey. As explained above, we study the extent to which an organization applies evaluation quality measures. In Flanders, employees dealing with policy evaluation are situated in a wide scale of positions in the organization chart. We therefore contacted the heads of the organizations asking who we could best approach for answering questions on evaluation quality assurance by the organization. Our respondents occupied positions such as scientific attaché; head of the units ‘management and strategy’, ‘policy and finance’ or ‘planning and organization’; and director of the ‘research unit’. Only those organisations that participated in the interview and in the survey were selected as cases. The preceding interview was necessary to discuss our conceptualization of policy evaluation in depth, and to gather detailed narrative information from the organisation. The stringent selection exercise resulted in eighteen cases. They

concern a large variety of departments and agencies . Examples are the Public Transport Agency, the Social Housing Company, the Employment Service, the Public Waste Agency, the Energy Agency, the Agency of Youth Welfare. To come to a valid assessment for each case, interviews with special advisors of the different sector ministers were planned to double-check the data received from the organisations themselves.

An open approach

As for the selection of conditions, we purposefully apply an eclectic and open approach. Rather than delimiting the analysis to a predefined selection of conditions, we analyze the explanatory value of a rich variety of organisational characteristics that are potentially relevant to understand evaluation quality assurance. We share the point of view of Katzenstein and Okawara (2001: 167) who claimed that analytical eclecticism, in which the causal importance of different types of factors is taken into account (e.g. structure; agency; ...) is highly needed to understand complex social and political phenomena. In absence of available systematic evidence on organisational variety in evaluation quality measures, we consider this the most fruitful strategy to advance the field. The research mainly proceeded in an inductive way. To come to a list of conditions with potential explanatory relevance, we combined two sources. First, we conducted an extensive screening of the evaluation literature. Most explanatory indications were found in the evaluation capacity building (ECB) literature. We continued our search for variables until we reached a point of saturation, in which no new variables were encountered. Complementary to literature screening, other relevant explanatory conditions emerged inductively through explorative interviews with top managers and evaluation staff of the field of education and training. This domain, commonly known as the pioneer in evaluation practice in Flanders, constituted our pilot case. The mixed approach of field work and literature review yielded a large list of potentially interesting conditions.

These conditions were subsequently ordered in five major categories. For this purpose, we were inspired by the principles underlying the actor centered neo-institutionalist framework (ACI), developed by Fritz Scharpf (1997). We deliberately used a ‘framework’, instead of a specific theory. As Scharpf (1997: 30) argued, instead of providing detailed general law-like regularities, the framework provides a descriptive language, which allows for the comparison of uniquely qualified cases by showing how a particular constellation of factors could bring about a specific effect. In other words, we rely on ACI as a heuristic tool to bring more structure in the inductively derived conditions. Although sharing a lot of premises with historical neo-institutionalism, ACI offers us a more suitable vehicle to tackle the challenges of this research. Most of the applications of historical neo-institutionalism generally relate to macro issues of socio-economic history, which are much broader than a phenomenon such as evaluation quality assurance of organizations. Secondly, many historical neo-institutionalist inspired works are characterized by long narratives that do not really provide clear analytical concepts (Witte, 2006). Thirdly, historical neo-institutionalist research ideally requires longitudinal historical analyses. Given the reshuffling of the Flemish public sector in 2006, we do not consider it feasible to get extensive and valid data from the past. Three elements supported the choice for ACI. First, it has been inductively grounded in and particularly developed for policy analysis. Second, ACI stands out for its feasible and relatively easily operationalisable definitions of actors, institutions and the relationships between both (Witte, 2006). Third, ACI fits very well with the assumptions behind the methodological strategy followed in this research. Scharpf shares the belief in conjunctural causation to understand complex social phenomena. In addition, ACI provides the theoretical justification to analyze the structure-related conditions separately from the actor-related conditions (see below, i.e. two step QCA).

ACI proceeds from the conviction that “social phenomena are to be explained as the outcome of interactions among intentional (individual and/or collective) actors, but that these interactions are structured, and the outcomes shaped, by the characteristics of the institutional settings in which they occur” (Scharpf, 1997: 1). In line with ACI, we analytically distinguish between two broad explanatory dimensions, which correspond with the major research traditions in the social sciences: on the one hand those emphasizing agency, and those that stress the role of structure on the other hand. Each of these dimensions can be disentangled into particular explanatory categories, five in total.

We conceptualize ‘structure’ in a broad way. We analyse the relevance of the characteristics of the institutional setting; characteristics of the tasks of the organisation and the history of the organisation.

1. Characteristics of the institutional setting: According to Scharpf (1997), 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. Based on our duo of sources, we deduce six conditions of institutional nature which can frame the actions of actors: the size of the organisation; the extent of formal autonomy that an organisation enjoys; the legal status of the organisation (department or agency); the anchorage of the evaluation function; the engagement of staff in specialized evaluation training, and the existence of regulatory evaluation requirements to which the organisation is subjected.

2. Whether evaluation quality assurance is considered necessary and feasible by a public sector organization, or by its unit in charge of evaluations, will presumably depend on the ‘fit’ with the issue-specific characteristics of the policy organisation itself (Schmidt, 1993). The scrutiny of the evaluation literature combined with our interviews alerted us to four different policy issue related condition variables that can exert important influence. Accordingly, we

will investigate the role of media and parliamentary attention for tasks of the organisation; perceived competition on tasks of the organisation; the perceived measurability of the organisational output and outcomes.

3. A third category of conditions deals with the path of the organisation. With this category, we include more longitudinal considerations in our research. As Scharpf (2000: 768) 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”. Accordingly, whether organisations employ evaluation quality measures, will likely depend on their experience with evaluations in the past. We will examine the influence of the length of evaluation experience; organisational stability; and ministerial stability.

In line with ACI, we treat actors as a theoretically distinct category. This is based on the observation that policies can sometimes change when the cognitive and normative orientations of policy makers change, even when the external and institutional conditions remain constant (Scharpf, 2000).

4. Actor orientations: Especially in the context of public sector evaluations, which obviously concern issues of public stake, various types of actors claim a voice. The majority of the testimonials of ECB sources we screened indeed referred to demand in general terms or by referring to the demand coming from a particular actor. In the present research, we focus on the influence of demand for evaluation from organisational management; minister; parliament; civil society organisations; and other organisations active within the same policy field. Finally, we test the explanatory relevance of organisation-wide support for evaluations.

5. Besides orientations, actors are characterized by capabilities. Capabilities are understood as all action resources allowing an actor to influence a certain outcome. Without, these, actors are not able to realize their orientations. Capabilities are of all kinds and for

instance refer to personal properties, physical resources, technological capabilities, etc. In the evaluation literature, the following capabilities are considered relevant for achieving high quality evaluations : skills to conduct evaluations; budgetary resources to evaluate; availability of experienced staff to conduct evaluations; availability of external evaluators with expertise on the themes of the organisation; and the availability of monitoring information. The overview of conditions we finally included in the research, can be consulted in annex 1. Their operationalization will be discussed further in the article.

Two Boolean methods of comparison

This article covers two research questions: (1) which conditions are most relevant to explain organizational variety in evaluation quality assurance; and (2) how do they combine? To answer these questions, we sequentially apply two different Boolean techniques. They imply a systematic pairwise comparison of cases.

(1) The inductive screening yielded a long list of explanatory variables that might account for organisational variation in the application of evaluation quality measures. To identify the conditions with most explanatory power, we apply the *Most Similar Different Outcome/Most Different Similar Outcome (MSDO/MDSO) technique* (De Meur, 1996). The technique is only rarely applied. This is surprising, especially given its potentials for explorative studies with a large number of conditions, which is typical for the present study too. One can recognize Mill's canons of agreement and difference (1973 [1843]) in the name and logic of the technique (De Meur, 1996; De Meur & Berg-Schlosser, 1994; De Meur & Gottcheiner, 2009). MSDO/MDSO seeks to identify the relevant conditions that are capable to explain why a pair of organisations that shares a lot of organisational attributes ('most similar'), differs in the extent to which evaluation quality measures are applied. To understand this different behavior, the (few) conditions on which these organisations differ will likely have most

explanatory potential and are worth exploring more in depth. Inversely, we focus on similarities, to understand why cases that are considered ‘most different’, nonetheless have a similar profile in evaluation quality management (De Meur, Bursens & Gottcheiner, 2006).

(2) The MSDO/MDSO technique assists us to identify the conditions that are most capable to explain organizational variety in evaluation quality assurance. Yet, such focus on individual conditions runs somehow counter to the complex nature of the evaluation reality, and the important role of context in this regard (Cooksy & Mark, 2011; Dahler-Larsen, 2012; Cooksy and Caracelli, 2005). We explicitly assume that it is a combination of factors (a conjunction) that explains the outcome, and that different combinations might produce the same outcome (equifinality) (Berg-Schlosser et al., 2009: 8; Schneider & Wagemann, 2012: 78). In the same vein, we expect that the effect of the presence or absence of a condition may differ depending on the wider context. While the MSDO/MDSO analyses identifies the individual conditions with most explanatory relevance, we are uncertain about which combinations can sufficiently explain why organisations have developed routine in using evaluation quality measures. For evaluation capacity building interests, having this information is beneficial for knowing which possible ‘recipes’ can account for routinized quality assurance. *Qualitative Comparative Analysis* (QCA) is particularly suited to unravel these sufficient combinations of conditions. Also Scharpf (1997) is convinced that this approach to causality best fits the complexity of social reality. QCA’s analysis of sufficiency corresponds with a ‘minimisation process’, which is again engrafted on Mill, and more particularly his method of pairwise difference (Ragin, 2000). Concisely formulated, this method builds upon the assumption that ‘if two configurations differ in only one condition, but show the same outcome, this particular condition can be eliminated’ (Schneider & Wagemann, 2012: 105). Iteratively repeating this premise until minimisations are no longer possible, the result of the process is an overview of ‘prime implicants’ that can imply the

entire set of configurations (Schneider and Wagemann, 2012, p. 109). They no longer comprise any explanatory ‘redundant’ variables. In the analysis section of this article, we explain the two techniques, MSDO/MDSO and QCA, in more detail.

A binary translation of the cases

The two techniques that we apply, rely on Boolean logic. Such logic requires a translation of our cases in terms of a combinations of binary (values: 0 and 1) conditions and outcomes. The codes ‘0’ and ‘1’ express qualitatively different states. Conditions with code ‘1’ stand for ‘present’ in the case, while a score of ‘0’ indicates ‘absence’. The concepts ‘present’ and ‘absent’ should not be considered in absolute terms. They can equally indicate nominal differences as ‘department’ versus ‘agency’, or refer to qualitative states as ‘strong’ versus ‘weak’, etc. Admittedly, coding our cases in dichotomous values inevitably implies a loss of detail. For the present research purposes, we prefer to focus on broad trends that we could measure in a reliable and valid way, rather than to focus on detailed nuances that could not be robustly measured. The latter is a side effect from using perceptions as proxies for most of our conditions. We observed that respondents were not always consistent in their perception about the extent of presence or absence of the same condition, in the interviews and the survey. Moreover, respondents of a single organisation did not always share the same nuances in perception, but agreed on broad lines. Focusing on dichotomous trends was therefore helpful to compensate for these differences in nuances. Furthermore, to our knowledge, the MSDO/MDSO technique is at present only applicable for binary coded settings. Annex 1 indicates how the different conditions were operationalized. The above-mentioned survey constituted the major source for the binary translation of our data. The columns of annex 1 refer to the answer possibilities of the survey questions. As the table

illustrates, for instance, no matter whether organisations assessed the availability of skills as ‘rather sufficient’ or ‘sufficient’, a score ‘1’ was given.

As for our outcome, i.e. the extent of evaluation quality measures applied, we chose to focus on those quality instruments that are most prominently present in the evaluation literature. The quality instruments selected suit all types of evaluations (in-house/outsourced, etc.). With a set of five indicators, we attempted to come to an overall judgment of the evaluation quality assurance at an organisational level. These indicators represent a mixture of structural, formative and summative approaches to evaluation quality assurance. Structural approaches refer to efforts in using guidelines and standards in evaluation. Formative approaches are quality assurance measures that aim to check quality during the course of conducting the evaluation work. And summative approaches refer to the assessment of evaluation quality once the evaluation report is finished (Schwartz & Mayne, 2005: 4-5). We specifically measured the extent of presence of the following indicators per organisation: (1) appointment of a steering committee or working group for the evaluations; (2) involvement of second readers for evaluation reports and interim versions; (3) involvement of stakeholders in the evaluation process; (4) conduct of meta-evaluations, which assess previous evaluation studies; and (5) use of evaluation manuals. Besides these five quality instruments, we asked organisations whether they apply any other quality measures. Each case was scored on the presence of the various quality assurance instruments, by asking respondents about the extent of the application of these quality tools. Scores per instrument ranged from 0 (never applied), 1 (when occasionally applied) and 2 (when systematically applied for all evaluations). The sum of the scores constitutes our composite indicator for this outcome variable.

-INSERT TABLE 1-

Comparing the use of quality assurance instruments across the cases, two major groups can be identified. On top of table 1 (indicated in bold and italics), we list the few

organisations for which a variety of quality assurance mechanisms are common practice. They obtain a minimum score of 7. They all share the habit of composing steering or working groups for each evaluation and to involve second readers. In addition, they rely on manuals for their evaluations, at least on an occasional basis. One of the organisations already conducted meta-evaluations for a specific subtype of its evaluations. The use of this type of quality instrument does not seem to be habitual practice yet in the Flemish context. Besides the quality instruments directly assessed, one organisation also reported having developed an ‘evaluability assessment’ procedure, as compulsory practice prior to every evaluation considered. All but one organisation of this group also engaged stakeholders in their evaluation processes. Involving stakeholders is probably the most contested element of the quality mechanisms studied. Whether we include it or not, the distribution of organisations across the two groups would remain unchanged.

Lower down in the table are the bulk of organisations for which quality assurance instruments are not yet mainstreamed. Their overall scores range from 1 to 5. Whereas most indeed have the tradition of working with steering or working groups or with second readers, they only ‘occasionally’ work with evaluation manuals. Apparently, meta-evaluations are, for this group, also a step ‘too far’. The identification of these two groups paves the way to code our outcome in binary terms. The ‘top group’ receives a score of ‘1’. The group with ‘laggards’ is assigned score ‘0’. The table in annex 2 gives the overview of binary scores for all cases.

Which individual conditions have most explanatory power?

To answer this first research question, we applied the MSDO/MDSO technique. Restricted by the size of this article, we focus on the main characteristics of the technique, and refer to more extensive descriptions elsewhere (see e.g. De Meur, 1996; De Meur & Berg-Schlosser, 1996;

De Meur, Bursens & Gottcheiner, 2006; De Meur & Gottcheiner, 2009). MSDO/MDSO requires three separate analyses: a comparison of (1) most different organisations for which a variety of quality assurance mechanisms are common practice (grey zone of the table in annex 2); (2) most different organisations for which these instruments are not yet mainstreamed (white zone in the table); and (3) most similar organisations but with a different evaluation quality profile (grey zone versus white zone).

To determine which pairs of cases are ‘most similar’ or ‘most different’, we relied on the binary data table. To measure the distance (proximity) between two cases, we simply counted the number of Boolean conditions in which they differ (are similar). To get a nuanced assessment of (dis)similarity, MSDO/MDSO requires the identification of categories that cluster related conditions. Our theoretical approach, ACI, provides the framework for the identification of the relevant categories for the present work. Compare for instance DEPT3 and DEPT11. They respectively differ in two; one; two; one and one conditions for the five explanatory categories. In MSDO/MDSO, (dis)similarities are first calculated per category and then aggregated (De Meur, 1996). As such, we get a comprehensive image of the breadth and depth of (dis)similarity between cases. Having the Boolean distances computed for each pair of cases made it possible to identify the minimum distance for pairs of cases with a different value on the outcome variable (MSDO) and the maximum distance for pairs with the same value for the routinization of evaluation (MDSO) (Berg-Schlosser and De Meur, 2009). Those pairs that differ most from each other can be said to differ at ‘level 0’ $D(0)$. ‘Level 1’ $D(1)$ is 1 away from ‘level 0’, whether there is a pair with this value of distance or not. The inverse reasoning was followed to calculate the similarity levels $S(0)$ to $S(k)$ between cases for MSDO (De Meur, 1994). As an assisting tool in the pairwise comparison of cases, we relied on the MSDO/MDSO software (beta-version 8/7/2006), developed by G. De Meur

(available via <http://www.jchr.be/01/beta.htm>). Only those pairs of cases that reach levels S(0)/D(0) and/or S(1)/D(1) for the highest number of ‘categories’ (h) were given this status.

Once these pairs were identified, we looked for the conditions that matter most in the categories for which the cases concerned were most (dis)similar. This exercise was conducted for each of the above mentioned zones. We only kept the conditions that were at least mentioned twice across several (dis)similar pairs of cases. Table 2 lists the conditions that met this criterion for the presence and the absence of evaluation quality assurance measures.

-INSERT TABLE 2-

First of all, the cases with a substantial number of quality measures can rely on skilled staff who know how to conduct evaluations in-house. Not surprisingly, knowledge of evaluation makes staff members more sensitive for the importance of quality assurance measures. Interestingly, budgetary impediments to evaluation are not a major obstacle to the introduction of systematic quality measures within the organisation. On the contrary, they can serve as a possible incentive not to waste precious financial sources and to conduct the evaluations with utmost care. Providing high quality evaluations is not necessarily a costly operation. The absence of external evaluation experts (consultants, researchers, etc.) neither constitutes a major hindrance to institutionalize quality measures. Furthermore, in our case setting, ministerial demand for evaluations turns out to be essential to proceed to high quality arrangements. Explaining this observation is not straightforward. Presumably a lot will depend on the character and the specific background of the minister in charge. As a respondent testifies: the profile of the sector minister will not only influence the “frequency of evaluation questions, but also the level of detail, the granularity of the evaluation questions, and the degree of demandingness”. We return to this issue, in the analysis of sufficiency. In addition, the length of evaluation experience of the organisation is of key relevance. We

herewith refer to whether the organisation already practiced evaluation prior to the NPM reforms. The development of high quality arrangements is a process that requires maturing with the practice of evaluation. Those organisations that only recently engaged in policy evaluation did not yet move to the systematic introduction of evaluation quality checks.

As for the other more structural conditions, the anchorage of the evaluation function has a key explanatory value. In our case area, anchorage ranges from having a large evaluation unit to the organisation having just one single staff member in charge of the coordination of evaluations. No matter the modalities of anchorage, we observe that an anchored evaluation function creates professional capacity for evaluation and makes it easier to systematically introduce quality measures for evaluations. The staff members working in these units operate as important ‘mediators’, or ‘evaluation quality guides’. In many instances in Flanders, the installation of an evaluation unit provided the momentum to structurally embed quality arrangements in the organisation. The position of international evaluation requirements is equally remarkable. Although typically considered critical for the diffusion of evaluation practices around the globe (Furubo & Sandahl, 2002), evaluation requirements by international institutions (i.e. European Union) do not always seem to be of key importance for explaining why organisations have substantially developed quality measures. As we will reveal later in the article, the effect of these requirements will depend on a favorable actor constellation.

Turning to the conditions of most explanatory relevance for cases with low quality assurance, the analysis again highlights the importance of an anchored evaluation function; though this time it focuses on its absence. While the presence of the evaluation unit was conducive to institutionalize evaluation quality measures within several Flemish organisations, the absence of such unit proved to hamper the introduction of quality measures. Additionally, organisational instability seems to impede the implementation of quality

measures. Several organisations that had been heavily reshuffled after NPM reforms reported having other priorities than introducing evaluation quality measures. An interviewee described it as follows: “At this moment, we are in a starting phase. In a first stage we focus on data collection and monitoring. We conduct some evaluations on an ad hoc basis, but these are not very well developed yet”.

The MSDO/MDSO analysis finally reveals some other conditions that can explain why ‘most similar’ pairs of cases differ in evaluation quality profile, and/or that can explain why ‘most different’ pairs of cases have a similar profile, but these conditions not always match with the same outcome value across different pairs of cases. Three conditions have this ambiguous status: evaluation demand from civil society organisations; evaluation clauses embedded in Flemish legislation, and measurability of outcomes. They have, at first sight, a weaker explanatory weight compared to the conditions that consistently match with the same outcome values. But, as these conditions can possibly play a decisive explanatory role in combination with other conditions, we will nonetheless include them in our ‘analysis of sufficiency’ below.

Which combination(s) of conditions can explain the high usage of evaluation quality measures?

A QCA analysis of sufficiency can help to unravel how the conditions with the most explanatory potential, interact. Stated differently, QCA can assist in finding the combinations of conditions that lead to a high usage of evaluation quality measures. We proceeded in two steps. Among practitioners of configurational comparative methods, this two-step approach has increasingly gained ground (Mannewitz, 2011; Schneider and Wagemann, 2006). In a first step, our analysis just includes the structural conditions. Only if these can not account for a fully sufficient explanation (consistency percentage of 100, see below), the conditions of

more actor-oriented nature are included (on top of the remote conditions which prove relevant). For our research purposes, this approach is extremely apt. As Scharpf (1997: 764) explains: “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”. The explicit distinction between causal remoteness and proximity offers a powerful analytical tool to come to more subtle explanations. As remote conditions constitute the setting in which the actor interactions take place, for many purposes, knowledge of these factors will be sufficient to understand a particular situation. If that is the case, there is in theory no analytical need to descend to a more proximate level of analysis (Scharpf, 1997; 2000). Besides its theoretical value, the two-level strategy is a useful vehicle to cope with the large number of conditions of our explorative research (Schneider and Wagemann, 2003, Mannewitz, 2011).

With QCA, we intend to unravel the trajectories to evaluation quality assurance, that no longer comprise any redundant conditions. To achieve more parsimony, QCA provides the possibility to make simplifying assumptions about non-observed combinations of conditions (logical remainders), for which we can hypothesize the likely outcome (Schneider and Wagemann, 2012: 151). These non-observed combinations differ in plausibility (Ragin & Sonnett, 2004). The most parsimonious solution can be obtained by including all remainders, no matter their theoretical or empirical plausibility. Given the lack of available theoretical and empirical evidence to make plausible assumptions, we predominantly rely, however, on the type of complex solution (i.e. the one that bars all remainders). Yet, the difference between the most complex solution and the most parsimonious solution offers the interesting possibility to make a causal distinction between core causal conditions and peripheral conditions (Fiss, 2011; Ragin, 2008). With Fiss (2011), we conceive causal coreness in terms of the strength of the evidence in relation to the outcome. Core conditions are those that are

shared by the two solutions, i.e. the complex and the parsimonious one. They can be surrounded by peripheral conditions, which correspond with the conditions that are eliminated in the parsimonious solution, but which feature in the most complex solution. Removing the peripheral conditions would imply the inclusion of counterfactuals, for which we cannot judge the plausibility of occurrence.

Within the scope of the present paper, we only present the findingsⁱ of the minimization for those cases in which evaluation quality assurance is routine.

-INSERT TABLES 3a & 3b-

Minimizing the remote conditions reduces the evaluation reality of the five cases using evaluation quality measures to two different paths (table 3a). Both paths share the presence of an evaluation unit, and the acquisition of pre-reform evaluation experience. All of our cases that systematically apply evaluation quality measures already practiced evaluation prior to the NPM oriented reforms.

A first path, capable of explaining 60% of the raw coverage, highlights the beneficial importance of organisational stability, as a core requirement to mainstream evaluation quality checks. This path is especially applicable for organisations with outcomes that are difficult to measure. This refers to these organisations that do not easily provide tangible services or products. Typically, although not exclusively, departments, especially since the NPM reforms, consider their outcomes as more difficult to measure than do agencies. The difficulty of measurability constitutes no hindrance, however, to deploy quality arrangements. Instead, it makes evaluation units consider the deployment of quality measures more necessary so as to check how evaluators cope with the outcomes that are difficult to measure. The lack of

international evaluation requirements is not an obstacle to make these organisations implement systematic quality measures.

The remaining 40% of the empirical coverage can be explained by a second path, in which international evaluation requirements nonetheless play a core role. The three empirical cases to which this path is applicable (DEPT3; DEPT11 and INTALP3) are active in policy sectors that receive European structural funds. They are not subjected to evaluation clauses in Flemish legislation and operate in a setting that recently underwent major reshufflings. However, this constellation of remote conditions can, on its own, not account for a fully sufficient explanation (consistency: 100%), as INTALP3 did not develop substantial quality arrangements. Supplementary conditions, of more actor oriented nature, thus needed to be inserted.

The minimization of the proximate conditions for these cases in scenario 2 can be summarized into one single core constellation, with the ministerial demand for evaluations as a core condition (table 3b). As we alluded to above, the type of actor demanding evaluation will influence the perceived need of organisations to implement quality measures. In our case setting, ministerial demand for evaluations turns out to be essential for proceeding to high quality arrangements. This is particularly what the two cases, DEPT3 and DEPT11, with institutionalized quality arrangements distinguishes from case INTALP3. INTALP 3 indeed lacks this quality reflex. The fact that the sector minister of DEPT11 was an academic clarifies a lot. Obviously, also the special ministerial advisors will here have a great stake. For some respondents, the influence of these advisors is more decisive than the influence of the sector minister him/herself. Other factors, of peripheral importance, are the condition referring to the evaluation demand from civil society and the skills to conduct-in house evaluations. The availability of a pool of external evaluators is substitutable with the availability of the budgetary means to evaluate.

Conclusion

Overall, the extent of evaluation quality assurance in the Flemish NPM-modeled administration is rather limited. The analyses of the distinguishing traits of the five organisations where evaluation quality measures are common practice versus the thirteen organisations where this is not (yet) the case directs theorists and practitioners to several decisive factors. In this conclusion, we would like to highlight the major explanatory factors, with the broader configurational picture in mind.

A first, and empirically most relevant element, is the beneficial role of an evaluation unit for the implementation of high quality measures. Regardless of the modalities of anchorage, the civil servants occupying the evaluation function importantly contribute to the systematic incorporation of quality measures within the organisation. They act as ‘evaluation quality mediators’ or ‘evaluation quality agents’.

Second, but much less empirically and theoretically strong, the application of evaluation quality measures is restricted to those organisations that practiced evaluation prior to the NPM reforms. The development of an evaluation attitude is a time consuming process, and so is the structural embedding of evaluation quality arrangements.

Thirdly, legislative evaluation requirements matter. The organisations that need to comply with international evaluation clauses comparably deploy more evaluation quality assurance mechanisms. The underlying mechanism is presumably to be sought in the evaluation training that are imposed together with the clauses. Such training usually involves instructions on good quality evaluations. An awareness of good practices makes organisations tempted to apply the quality instructions also to other evaluations, which are not initiated by legislation. This line of reasoning is conditional, however, on the presence of a beneficial actor constellation.

In this respect, we want to emphasize the quintessential role of evaluation demand from the sector minister and his/her advisors as a fourth condition with high discriminatory potential. An important intermediary element in this regard, as we can speculate, is the background and character of the minister in charge, and especially of his or her personal advisors. The pace of evaluating will often not match the pace of political decision making. The extent to which quality measures are systematically employed, can possibly also influence the length of the evaluation process. Feedback of steering groups, for instance, might lead to an extension of the fieldwork, to the conduct of additional analyses, etc. The attitude, and voluntarism of the minister vis-à-vis evaluation can then be decisive for creating or subtracting space to apply evaluation quality measures.

Organisational stability completes our list of most remarkable influencing conditions. Out of the eighteen organisations studied, only five did not undergo substantial structural changes. Three of these five are among the group of cases where evaluation quality measures are common practice. A stable environment seems conducive to the systematic deployment of quality measures.

Taking stock of the conditions with core explanatory power, we can derive some practical lessons for evaluation capacity building. On a negative note, the promotion of quality measures will not be evident for cases without a beneficial historical legacy, as all of our cases that apply evaluation quality measures have some years of experience with evaluating. Our analyses nonetheless give some hints for cases where this experience is lacking. The installation of an evaluation unit, for instance, can be an important vehicle to structurally embed quality arrangements in the organisation. Alternatively, the introduction of evaluation clauses in legislation may help. Not just because of the clause itself, but supposing that these are associated with the provision of relevant training in good quality evaluation

practice. Where there is a favorable actor constellation, these measures can help to trigger more organisations incorporating substantial quality arrangements.

Flanders is a case within the second wave of countries or regions to develop evaluation. In this group, external (EU) and NPM pressures played a major role in the diffusion of evaluation (Furubo & Sandahl, 2002). We believe that the findings of the present study can be useful for other countries in the same group. From a methodological perspective, the article has demonstrated the strength of MSDO/MDSO and two step Qualitative Comparative Analysis to bring clarity in the contingent evaluation praxis. The combination of the techniques was particularly helpful to deal with the large number of conditions that were included in this explorative study. Future research could elaborate on the precise weight of the conditions that we identified with the MSDO/MDSO technique. Also, the causal process and mechanisms behind the sufficient combinations of condition, merit further analysis.

References

- Benson, A.P., Hinn, M.D., & Lloyd, C. (2001). Preface. In A.P. Benson, M.D. Hinn, & C. Lloyd (Eds.), *Visions of Quality: How evaluators define, understand and represent program quality* (pp. ix-xii). Oxford: Elsevier Science.
- 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* (pp. 19-32). London: Sage.
- Borrmann, A., & Stockmann, R. (2009). *Evaluation in German development cooperation. A system analysis*. Münster: Waxmann Verlag.
- Cooksy, L. & Caracelli, V. (2005). Quality, context and use. Issues in achieving the goals of metaevaluation. *American Journal of Evaluation*, 26(1), 31-42.
- Cooksy, L. & Mark, M.M. (2011). Influence on evaluation quality. *American Journal of Evaluation*, 00(0), 1-9.
- Dahler-Larsen, P. (2012). *The evaluation society*. Stanford: Stanford University Press
- De Meur, G. (1996). La comparaison des systèmes politiques: recherche des similarités et des différences. *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* (pp. 67-94). 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* (pp. 208-221). London, Sage, pp. 208-221
- Delreux, T. & Hesters, D. (2010). Solving contradictory simplifying assumptions in QCA: presentation of a new best practice. *COMPASS Working Paper*, 58. Retrieved from www.compass.org
- European Commission (2012). *The use of evaluations and knowledge management. Evalsed*. Retrieved from http://ec.europa.eu/regional_policy/sources/docgener/evaluation/evalsed/guide/designing_implementing/use_en.htm.
- Fiss, P. C. (2011). Building better causal theories: A fuzzy-set approach to typologies in organisation research. *The Academy of Management Journal*, 54(2), 393-420
- Forss, K., Vedung E., Kruse S.E., Mwaiselage A., & Nilsson, A. (2008). *Are Sida evaluations good enough?* Sida Studies in Evaluation: Swedish International Development Cooperation Agency
- Furubo, J.E. and Sandahl, R. (2002). Introduction. A diffusion perspective on global developments in evaluation. In J.E. Furubo, R.C. Rist & R. Sandahl (Eds.), *International Atlas of Evaluation* (pp.1-23). New Jersey: Transaction publishers.
- IAVA (2007; 2008; 2009). *Jaarverslag van het Auditcomité en het Agentschap Interne Audit van de Vlaamse Administratie*. IAVA: Brussel.
- Johnsen, A., Mecklin, P., Oulasvirta, L. & Vakkuri, J. (2001). Performance auditing in local government: An explorative study of perceived efficiency of municipal value for money auditing. *The European Accounting Review*, 10(3), 583-599.
- Katzenstein, P.J. & Okawara, N. (2001). Japan, Asian-Pacific Security, and the Case for Analytical Eclecticism. *International Security*, 26(3), 153-185.

- Leviton, L. C., & Hughes, E. F. X. (1981). Research on the utilization of evaluations: A review and synthesis. *Evaluation Review*, 5, 525-547.
- Mande (2011). On evaluation quality standards. A list. Retrieved from <http://mande.co.uk/2011/lists/evaluation-quality-standards/on-evaluation-quality-standards-a-list/>
- 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.compassss.org
- 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. (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. Minckenberg (Eds.), *Vergleichen in der Politikwissenschaft* (pp. 180-197). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Sanders, J.R. (2002). Presidential address: On mainstreaming evaluation. *American Journal of Evaluation*, 23(3), 253-259.
- 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 für 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* (pp. 318-350). 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.
- Schneider, C.Q. & Wagemann, C. (2012). *Set-theoretic Methods for the Social Sciences. A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press.

- Schwartz, R., & Mayne, J. (2005). Assuring the quality of evaluative information: Theory and practice. *Evaluation and Program Planning*, 28, 1-14.
- Scriven, M. (1991) *Evaluation Thesaurus. 4th edition*. California: Sage Publications.
- Van der Meer, F.B. (2007). New Public Management and Evaluation. In C. Pollitt, S. Van Thiel & V. M. F. Homburg (Eds.), *The New Public Management in Europe. Adaptation and Alternatives* (pp. 165-180). London: Palgrave Macmillan.
- Varone, F., Jacob, S., & De Winter, L. (2005). Polity, Politics and policy evaluation in Belgium. *Evaluation*, 11(3), 253-273.
- Vlaams Parlement (2003). *Kaderdecreet Bestuurlijk Beleid*. Brussels: Flemish Parliament, 18 July 2003.
- 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* (pp. 41-68). New Brunswick: 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.

Annex 1. Overview of conditions, their indicators, and dichotomization thresholds

<i>Conditions</i>	<i>Code 0 indicators</i>	<i>Code 1 indicators</i>
Category A: Capabilities of the organisation		
(1) Skills to conduct evaluations, (2) Financial means to evaluate, (3) Availability of staff capable to evaluate, (4) Availability of external evaluators, (5) Availability of monitoring information	Totally insufficient, rather insufficient	Rather sufficient, fully sufficient
Category B: Orientations		
Evaluation demand from: (6) Organisational management, (7) Sector minister, (8) Parliament, (9) Civil society organisations, (10) Other organisations	No demand, hardly any demand	Sometimes demand, frequent demand
(11) Extent of organisation-wide support for evaluations	Not at all, to limited extent	To major extent, to large extent
Category C: Conditions with regard to the institutional setting		
(12) Organisational size	Very low, low material	At least average

	weight (*)	material weight
(13) Organisational autonomy	No legal personality	Legal personality
(14) Organisational status	Agency	Department
(15) Anchorage of evaluation function	No evaluation unit	Formal or de facto evaluation unit
(16) Participation of staff in evaluation trainings	No participation in evaluation trainings	Minimally 'sometimes' participating in evaluation trainings
Evaluation requirements stipulated in: (17) regulation or decrees at Flemish level, (18) legislation/regulation at EU level, (19) management agreement of organisation	No evaluation requirements	Evaluation requirements
Category D: Conditions concerning policy issue characteristics		
(20) Attention by media or parliament for the tasks of the organisation	Not at all, limited, rather limited	Rather much, much, very much
	<i>Highest score of the assessments of attention by each of these 'institutions'</i>	
(21) Perceived competition on tasks of the organisation	Not at all, limited, rather limited	Rather much, much, very much
(22) Perceived measurability of outputs, and (23) outcomes	Average score ≤ 3 and/or qualification: very difficult, difficult, rather difficult	Average score ≥ 3 and/or qualification very easy, rather easy,

		easy
	<i>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 and outcomes of the organisation.</i>	
Category E: Conditions characterizing the path of the organisation		
(24) Pre-NPM evaluation experience	No/seldom evaluation practice prior to the NPM oriented reforms	Sometimes/frequent evaluation practice prior to the NPM oriented reforms
(25) Organisational stability	Organisations which underwent medium or large changes (**)	Organisations which underwent no or small changes
(26) Ministerial stability	≥ 1 minister changes since BAP	No ministerial turnover

(*)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 organisation: (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 for the years 2007-2008-2009 (IAVA).

(**) 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 organisation; (2) changes with regard to the composition of the public entity; (3) changes with regard to the

organisation 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 organisations which underwent a large number of changes; 0.3 for these which underwent a medium number of changes and 0.1. for these organisations which can be characterized by large stability. We calculated the average for the years 2007-2008-2009 (IAVA).

Annex 2. Binary data table

	CASE(*)	Category A: Actor capabilities					Category B: Actor orientations					Category C: Institutional setting					Category D: Policy issue characteristics					Category E: Path of the organisation					
		Skills to conduct evaluations	Financial means to evaluate	Availability of staff capable to evaluate	Availability of external evaluators	Availability of monitoring info	Evaluation demand from management	Evaluation demand from minister	Evaluation demand from parliament	Evaluation demand from civil society organisations	Evaluation demand from other organisations	Organisation –wide support for evaluations	Organisational size	Organisational autonomy	Organisational status	Anchorage of the evaluation function	Participation of staff in evaluation training	Evaluation requirements in management agreement	Evaluation requirements in Flemish legislation	Evaluation requirements coming from European Union	Media/parliamentary attention for organisation	Competition	Measurability of outputs	Measurability of outcomes	Pre-NPM evaluation experience	Organisational stability	Ministerial stability
Organisations with high evaluation quality assurance	DEPT3	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
	DEPT11	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
	ILTALP2	0	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
	DEPT2	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
	DEPT8	0	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
Organisations with low evaluation quality assurance	INTA2	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
	DEPT7	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
	INTALP3	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
	INTA9	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
	ILTALP1	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
	EXTA6	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
	INTA5	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
	EXTA2	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
	EXTA3	0	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

	<i>EXTA1</i>	0	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
	<i>DEPT1</i>	0	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
	<i>DEPT5</i>	0	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
	<i>DEPT6</i>	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

(*) Case codes are composed by a letter combination and a number. DEPT stands for Department; EXTA for external agencies of public nature; INTA for internal agencies without legal personality; INTALP for internal agencies with legal personality. The numbers are chosen at random.

Table 1: Overview of evaluation quality measures applied (scores range from 0 to 2)

Code of the organisation	Steering						TOTAL
	Committee/ Working group	Second readers	Stakeholders	Meta- evaluations	Manuals ⁱⁱ	Other	
<i>DEPT2</i>	<i>2</i>	<i>2</i>	<i>2</i>	<i>1</i>	<i>2</i>	<i>0</i>	<i>9</i>
<i>DEPT3</i>	<i>2</i>	<i>2</i>	<i>2</i>	<i>0</i>	<i>2</i>	<i>0</i>	<i>8</i>
<i>DEPT8</i>	<i>2</i>	<i>2</i>	<i>0</i>	<i>0</i>	<i>2</i>	<i>2</i>	<i>8</i>
<i>DEPT11</i>	<i>2</i>	<i>2</i>	<i>2</i>	<i>0</i>	<i>1</i>	<i>0</i>	<i>7</i>
<i>ILTALP2</i>	<i>2</i>	<i>2</i>	<i>1</i>	<i>0</i>	<i>2</i>	<i>0</i>	<i>7</i>
ILTALP1	2	2	1	0	0	0	5
DEPT1	2	2	1	0	0	0	5
DEPT6	2	2	1	0	0	0	5
DEPT7	2	2	0	0	1	0	5
INTA9	2	2	1	0	1	0	5
INTALP3	0	2	2	0	0	0	4
INTA5	1	1	2	0	0	0	4
EXTA1	2	2	0	0	0	0	4
EXTA6	1	1	1	0	0	0	3
EXTA3	1	1	1	0	0	0	3
DEPT5	2	0	0	0	1	0	3
INTA2	0	2	0	0	0	0	2
EXTA2	0	0	0	0	1	0	1

Table 2: Most relevant conditions, identified in the MSDO/MDSO analyses

Identification of the most relevant conditions, for the organisations with high evaluation quality assurance	
	Category
BUDG(0)	A
EVCOMM(0)	A
SKINT(1)	A
DEMMIN(1)	B
ANCH(1)	C
REGINT(0)	C
LEG(1)	E
Identification of the most relevant conditions, for the organisations with low evaluation quality assurance	
ANCH(0)	C
ORGSTAB(0)	E
Conditions leading to ambiguous configurations (C)	
DEMCSO	B
REGFL	C
MEASOC	D

Table 3a: Configurations leading to high evaluation quality assurance.

Results of minimisation of remote conditions

ANALYSIS OF REMOTE CONDITIONS		
	1	2
<i>Category C. Conditions related to the institutional setting of the organisation</i>		
Anchorage of the evaluation function	●	●
Evaluation requirements in Flemish legislation		θ
Evaluation requirements coming from the European Union	θ	●
<i>Category D. Conditions related to the tasks of the organisation</i>		
Measurability of outcomes	○	
<i>Category E. Conditions related to the path of the organisation</i>		
Pre-NPM evaluation experience	●	●
Organisational stability	●	θ
Measures of fit		
Consistency	1.00	0.67
Raw coverage	0.60	0.40
Unique coverage	0.60	0.40
Solution coverage: 1.00 / Solution consistency: 0.83		

Note: The notation used is based on Fiss (2011) and Ragin (2008). Full circles refer to presence of a condition; Crossed-out circles to 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 different core constellations are indicated with a different number above the columns.

Table 3b: Configurations leading to high evaluation quality assurance.

Results of minimisation of the proximate conditions

ANALYSIS OF PROXIMATE CONDITIONS (*)		
	2A	2B
<i>Category A. Actor capabilities</i>		
Skills to conduct evaluations	•	•
Financial means to evaluate	θ	•
Availability of external evaluators	•	θ
<i>Category B. Actor orientations</i>		
Evaluation demand coming from civil society organisations	•	•
Evaluation demand from sector minister	●	●
Measures of fit		
Consistency	1.00	1.00
Raw coverage	$0.50 \cdot 0.40 = 0.20$	$0.50 \cdot 0.40 = 0.20$
Unique coverage	$0.50 \cdot 0.40 = 0.20$	$0.50 \cdot 0.40 = 0.20$
Solution coverage: 1.00 / Solution consistency: 1.00		

(*) ONLY for these cases with an evaluation unit, with pre-NPM evaluation experience, but with weak organisational stability, with evaluation requirements coming from the European Union, and without evaluation requirements stipulated in Flemish legislation.

ⁱ The minimization process of the parsimonious solutions yielded thirteen contradictory simplifying assumptions. These were solved on the basis of the knowledge about the necessary conditions in the analysis. We followed the approach explained in Delreux and Hesters (2010).

ⁱⁱ Two organisations reported to follow fixed procedures for their evaluations, but have not anchored these in a written document. They were given a score of 1 on this quality criterion.