The Canadian Journal of Program Evaluation Vol. 28 No. 2 Pages 43–63
ISSN 0834-1516

OUTSOURCE VERSUS IN-HOUSE? AN IDENTIFICATION OF ORGANIZATIONAL CONDITIONS INFLUENCING THE CHOICE FOR INTERNAL OR EXTERNAL EVALUATORS

Valérie Pattyn

Marleen Brans
University of Leuven, Public Management Institute, Belgium

Abstract: An evaluation can be conducted in-house or can be outsourced to an external party. Yet organizations do not always have full discretion to decide on the locus for evaluation implementation. Certain attributes often push the organization in one direction or another. Via a systematic pairwise comparison of attributes of 18 organizations in the Flemish (Belgian) public sector, we were able to indicate the conditions that matter most in determining the locus of policy evaluation implementation. Our findings can thus enrich existing guidelines on the advantages and disadvantages of internal and external evaluations.

Résumé : Une évaluation peut être conduite en interne ou être sous-traitée à un partenaire extérieur. Une organisation n'aura pas toujours l'entière latitude pour décider du lieu de l'implémentation de l'évaluation. En effet, certains attributs vont souvent orienter l'organisation dans l'une ou l'autre direction. Grâce à une comparaison systématique deux-à-deux des attributs de dix-huit organisations du service public belge flamand, nous sommes en mesure d'indiquer les conditions les plus déterminantes pour le lieu de l'évaluation des politiques publiques. Nos résultats peuvent enrichir les directives existantes sur les arguments en faveur ou en défaveur des évaluations externes.

INTRODUCTION

In principle, with every decision to evaluate, an organization should make the choice whether it will ask staff members to conduct the evaluation or whether it will hire someone from outside the organization for this task. When listing the criteria that determine the choice for a particular locus, the following considerations usually appear in the literature: cost, knowledge, flexibility, objectiv-

Corresponding author: Valérie Pattyn, KU Leuven Public Management Institute, Parkstraat 45, Box 3609, 3000 Leuven, Belgium; <valerie.pattyn@soc.kuleuven.be>

ity, accountability, willingness to criticize, ethics, and utilization of results (Conley-Tyler, 2005). Typically, the advantages of an in-house evaluation are the symmetrical inverse of the disadvantages of an external evaluation, and vice versa (Sonnichsen, 2000, p. 64). Depending on the weight given to each of the relevant considerations, organizations might choose one or the other approach to what we call the "locus" of evaluation.

Much of the evaluation literature appears to have been written on the assumption that organizations can choose a different strategy for each evaluand. Yet, in practice, many organizations are often biased toward one or the other approach, irrespective of the specific evaluand at stake. The freedom of choice for internal or external evaluation is relative, and is heavily influenced by the framing conditions in which an organization operates. Despite the increased attention to context in the evaluation field (Dahler-Larsen, 2012; Sonnichsen, 1994, 2000), there is still a knowledge gap about the conditions that explain why some organizations mainly choose in-house evaluations or dominantly opt for ex-house evaluations. As not all organizations follow the general trend toward more internal evaluations (Mathison, 2011), we are challenged to answer the following question: Which organizational contextual factors are more decisive than others in explaining variations in the locus of evaluation?

This article seeks to answer this question, based on the analysis of 18 organizations of the Flemish (Belgian) public sector. After outlining the relevance of this study, we offer background on the Flemish cases, and discuss the conditions included in the analysis. A subsequent section presents the Most Similar Different Outcome/Most Different Similar Outcome (MSDO/MDSO) technique: an innovative method that helped us to identify the conditions that can best explain why some organizations have a tendency for internal evaluations and others usually rely on external evaluations. We conclude the article with a discussion of our findings.

RELEVANCE

Internal evaluation has been on a steady increase (Mathison, 2011), to the extent that it has now developed into a genuine subfield of evaluation: Internal evaluation is "a key way for organizations to set their own directions, foster change, and know if they were achieving results" (Volkov, 2011, p. 6). Wildavsky's seminal piece on the "self-evaluating organization" (1972) can be considered a major reference

in this subfield. While a strong case for internal evaluation is often made, most scholars now agree, however, that the choice should be made with respect to the purpose and especially to the primary users of the study (Sonnichsen, 2000; Vedung, 2009; Volkov, 2011). Vedung (2009) makes a distinction between the purposes of accountability, improvement, and basic knowledge. First, when the purpose is accountability to external stakeholders, it seems best that evaluations be externally conducted. External evaluators are usually seen as more objective than internal evaluators (Conley-Tyler, 2005). Internal evaluators might find their objectivity compromised by the policies and politics of the organization and its underlying value system. On the downside, however, external evaluators run the risk of losing the confidence of people in the agency, when they appear to be too critical. Moreover, the fact that they are often seeking new contracts can make them vulnerable to direct conflicts of interest (Vedung, 2009, p. 118; Worthen et al., 1997, p. 18). It also happens that external evaluations are commissioned for mere symbolic or strategic accountability purposes: to justify pre-existing organizational decisions or to provide "appearances more flattering than reality" (Vedung, 2009, p. 111).

Second, when the goal is program amelioration and refinement, evaluations are preferably conducted by in-house staff (Vedung, 2009, p. 119). Internal evaluators can better formulate the evaluation findings in the language of their organization, and thus enhance the chance of utilization of evaluation findings. It can be argued that "[i] t is difficult for an external evaluator to ever learn as much about the program as the insider knows" (Worthen et al., 1997, p. 18). However, external evaluators may also have the ability to communicate relevant information, on condition that they work closely with stakeholders in a participative mode (Conley-Tyler, 2005; Fetterman, 2001).

Third, basic knowledge evaluations refer to the type of "fundamental research that seeks to increase the general understanding of reality" (Vedung, 2009, p. 110). Meta-evaluations are a classic example. The primary audiences of this type of evaluation are not particular agencies, but rather government operators, the evaluation community, or academia. Therefore, basic knowledge evaluations might suitably be conducted by an external party, such as research institutes or universities. But it can be equally useful for the general culture of an agency to conduct a synthesizing meta-evaluation in-house (Vedung, 2009, p. 120).

Although we only touched upon some of the most apparent advantages and disadvantages of internal or external evaluation (for a more extensive list, see, e.g., Sonnichsen, 2000, p. 64), it is apparent that the evaluation literature is not always conclusive about the best locus for each evaluation purpose. Besides, it is clear that the literature primarily focuses on individual evaluation studies. The organizational context, which transcends individual evaluation studies, often remains out of the picture. Yet we can assume that this context strongly determines the organizational inclination toward one locus or the other. As Dahler-Larsen (2012, p. 36) argues, "as evaluation becomes an organizational functional entity, it is subject to the structures, values and rules of the organization." Especially in times where internal evaluation is on the increase, it is relevant to understand (a) which conditions facilitate organizations to comply with this trend, and (b) which conditions explain why an organization is still outsourcing most of its evaluations. Unlike Wildavsky (1972, p. 509), who focused on preconditions "depressing or facilitating evaluation," we systematically compare organizations that are all active in evaluation, but with varying biases in locus. We focus not on the ideal, and do not consider whether evaluations indeed get "high impact" (Sonnichsen, 2000).

The demarcation line between internal and external evaluations can be an issue of discussion. Scriven (1980, quoted in Mathison, 1991, p. 159) has emphasized that "internal/external is really a difference of degree rather than kind." In the present study, we consider the stage of "conclusion formulation" as the demarcation point. When the conclusions are formulated by an external party (e.g., consultant or university), we label this evaluation as "external." When these conclusions are instead formulated by the organization itself, the evaluation is considered "internal" (see also Vedung, 2009). With evaluations, we refer to Scriven's (1980) conceptualization by defining them as "scientific analyses 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." The definition of evaluations studied in this article is more restrictive than the encompassing conceptualizations applied in many of the works on internal evaluation (e.g., Sonnichsen, 2000). Performance measurements, for example, fall outside the scope of our analysis.

THE FLEMISH PUBLIC SECTOR AS OUR DOMAIN OF INVESTIGATION

Evaluation practice has spread rather slowly and unevenly among European countries (Furubo, Rist, & Sandahl, 2002). Belgium is usually

situated in a second wave of countries that have adopted evaluation practice, mainly following external pressures. In Belgium the momentum for increased attention to policy evaluation only came when the second wave was already slowing down. Our article will focus on Flanders, to date a largely undiscovered evaluation area. Unlike in a few pioneer countries or regions, evaluation was not introduced here as an autonomous tool, but as part of a broader package of reform. In 2006, the Flemish administration implemented a government-wide reform package called "Better Administrative Policy" (BAP), which was modelled along the New Public Management (NPM) blueprint. Although policy evaluation was not a core principle of the reforms, it has indirectly been given important attention. Departments have been explicitly assigned the policy evaluation function, and autonomized agencies are supposed to generate input by means of relevant policy and managerial information for policy evaluation (VlaamsParlement, 2003). The fact that the reforms incorporate substantial NPM-inspired elements makes it a very interesting case. Generally, as described by Mathison (2011), NPM has served as an important catalyst for the promotion of internal evaluation. Several years after the implementation of the reforms, policy evaluation practice is now practiced by a large variety of public sector organizations in Flanders, across a large variety of policy fields. The present research is based on a study of 18 of these organizations. They represent the different policy fields in which Flanders is active: education; mobility; environment; energy; housing; agriculture; work and social economics; welfare, public health, and family; culture and youth; and economics. Of the cases we investigated, 11 mainly conduct their evaluations in-house. The other 7 cases are more inclined to outsource their evaluations.

EXPLANATORY CONDITIONS

Our study being the first of its kind, at least to our knowledge, we consider it most reliable to proceed with an open, inductive approach, and not to exclude any potentially interesting conditions beforehand. To this goal, we applied a mixed strategy of evaluation literature screening and semistructured interviews with representatives from our particular administrative area. 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 enabled us to control for a large number of conditions.

We distinguish between five categories of conditions, inspired by actor-centred neo-institutionalism (Scharpf, 1997). Two of them are

actor related. The other three categories are of a more structural nature. In the actor-related categories, we make a distinction between capabilities of the organization and actor orientations vis-à-vis evaluation. From a more structural perspective, we distinguish between conditions relating to the institutional setting, conditions concerning policy issue characteristics, and conditions concerning the path of the organization. Appendix 1 (first column) lists the 26 conditions of the five categories with potential explanatory relevance.

A SYSTEMATIC PAIRWISE COMPARISON OF ORGANIZATIONS

Our search for interesting explanatory conditions resulted in a long list of factors. The challenge was then to identify those conditions that have most explanatory strength. We chose to rely on the Most Similar Different Outcome/Most Different Similar Outcome (MSDO/MDSO) technique. This method was originally developed by G. De Meur as a systematic comparative tool to reduce the complexity of a large data set (see, e.g., De Meur, 1996; De Meur & Berg-Schlosser, 1994). The technique has proven to be especially useful to get out of the usual "small n–many variables dilemma" (De Meur & Berg-Schlosser, 1994). It helped us to keep the overview of our 18 cases: a number too small for the application of most statistical techniques, but too big for in-depth, case-oriented analysis. Despite its potentials, MSDO/MDSO has only been applied in a limited way to date. The technique basically involves a rigorous and systematic application of the assumptions of J. S. Mill (1973), which underpin most comparative research designs in the social sciences. Rather than focusing on similar and different cases that differ or share only one condition, MSDO/MDSO takes a more realistic approach. Its major focus goes to *most similar* and *most different* cases (De Meur, 1996; De Meur & Gottcheiner, 2009). The underlying idea is that the most "extreme" pairs of cases, in terms of degree of (dis)similarity, embody the strongest explanatory potential. Key objectives are to identify the factors that can explain why two very different organizations share the same evaluation locus or vice versa: why organizations that share many attributes differ in evaluation locus. Phrasing it differently: when two organizations share hardly any of the conditions of our list but both are outsourcing their evaluations, we can only understand this similar evaluation behaviour by focusing on their limited similarities. And inversely: when two organizations share almost all conditions but differ in evaluation locus, we can only understand this variety by concentrating on the few conditions on which they differ. To decide which pairs are most similar or most different, the

technique requires the translation of all organizational attributes (conditions) into binary codes (values: 0 or 1). Dichotomization deliberately simplifies social reality, to keep the focus on large trends and general patterns (De Meur & Berg-Schlosser, 1996; De Meur, Rihoux, & Yamasaki, 2009). The main source for the coding of our conditions was a survey with closed questions, sent to the management of our organizations. The survey was only sent to civil servants that had already been interviewed in a prior stage. It was assumed that the interview helped both parties (respondents and interviewer) to understand each other's interpretation of core evaluation concepts. The actual content of the survey mainly overlapped with the interviews' content, but enabled the collection of more structured answers. Appendix 1 (second column) details what the closed answers of the survey questions looked like. The closed answers made comparisons across organizations more reliable. Interviews with advisers of the ministerial cabinet were conducted to double-check the data with what was received from the organizations themselves. In addition, data were supplemented with document analysis, so we could verify whether reported evaluation studies really fell within this study's scope of analysis. We take the perceptions of respondents as "proxies" for the actual reality. Via the triangulation of different kinds of sources and different types of respondents, we nonetheless received a robust and reliable picture of each organization. The binary coding of the data proved helpful to compensate for the slight differences in subjective assessment between and within organizations. We noticed that respondents sometimes use different nuances throughout time to assess the same situation, but are consistent in the general pattern. For instance, no matter whether respondents considered the availability of monitoring information as *rather sufficient* or *fully sufficient*, we coded this condition as "1." Appendix 1 lists the thresholds used for the binary coding. Appendix 2 provides the overview of codes for our organizations.

With our data dichotomized, we proceeded to the identification of pairs that are most similar or most different by simply calculating the number of dichotomous conditions for which cases differ. This calculation is done per category of conditions, since cases can be similar in one category (e.g., actor capabilities) but dissimilar for another (e.g., policy issue characteristics). Compare, for instance, organizations INTA2 and DEPT7 in Appendix 2. They respectively differ in 2, 4, 2, 0, and 1 conditions for categories A to E. To come to a comprehensive judgement on (dis)similarity, three analyses were conducted in parallel and compared with each other: (a) a pairwise

comparison of all cases that dominantly conduct in-house evaluations (grey zone in Appendix 2); (b) a pairwise comparison of all cases that mainly outsource their evaluations (white zone); and (c) a pairwise comparison of cases that result in different values on the locus of evaluation (grey zone versus white zone). Within each of the zones, we identified those pair(s) of cases that were most (dis) similar for the highest number of categories. DEPT1 and DEPT2, for instance, can be considered among the most dissimilar pairs of the white zone, as they strongly differ in three categories (conditions relating to the path of the organization, issue characteristics, and actor orientations). Both outsource their evaluations.

Once the most (dis)similar pairs were identified, we looked for the conditions that have most explanatory value. For most different pairs with the same evaluation locus, we were especially interested in the conditions for which they have the same value. The highly dissimilar cases DEPT1 and DEPT2, for instance, share the perception of outcomes that are hard to measure, strong managerial demand for evaluations, and high demand of civil society organizations for evaluations. These conditions can thus be considered to have strong explanatory potential. Inversely, for most similar pairs of cases with different evaluation loci, we were interested in the conditions on which they differ. Organizations EXTA2 and INTALP2 are among the most similar pairs of cases. They share almost all characteristics for the category relating to the institutional setting of the organization. They have a different value only for the anchorage of an evaluation function. This condition thus seems to be more relevant than the others. Following Bursens (1999), we only kept the conditions that were mentioned at least twice across several (dis)similar pairs of cases. For more technical details on the procedure, we refer the reader to more specialized works (De Meur, 1996; De Meur & Berg-Schlosser, 1996; De Meur & Gottcheiner, 2009).[1]

WHICH INDIVIDUAL CONDITIONS HAVE MOST EXPLANATORY POTENTIAL?

Table 1 lists the conditions that were identified as most explanatory powerful.[2]

The various categories of actor-centred neo-institutionalism all turn out to be relevant. Yet, in our Flemish setting, structural conditions appear generally more explanatory than conditions of an actor-related nature. Starting with the actor-related factors, it is evident that

Table 1
Overview of Most Relevant Conditions

Conditions with most explanatory potential for cases with a tendency to outsource evaluations (+ category)	
Availability of external evaluators	A
Evaluation demand from civil society organizations	B
No/limited evaluation demand from sector minister	B
Anchored evaluation unit	C
Outcomes that are hard to measure	D
No evaluation experience before the NPM reforms	E
Organizational stability	E
Conditions with most explanatory potential for cases with a tendency for in-house evaluations (+ category)	
Evaluation demand from sector minister	B
Absent/limited legislative evaluation requirements	C
Absent/limited competition	D
Outcomes that are easy to measure	D
Outputs that are easy to measure	D
Evaluation experience before the NPM reforms	E
Weak organizational stability	E

outsourcing evaluations will be facilitated when an organization can rely on *available external human capital*. In our interviews this element was repeatedly named as the condition that often jeopardizes the successful implementation of planned evaluations. In contrast to that of many other countries, the pool of available evaluators in Flanders is very limited. True, variation exists between policy fields. But the calls for evaluations often need to be re-issued, in the absence of any qualified response. The fact that the Flemish evaluation market is very small not only has repercussions for the extent of outsourcing evaluations but also has important consequences in many other respects. As only a limited number of applicants usually respond to evaluation calls, inevitably the same evaluators are repeatedly evaluating the same policies time and again. Similarly, a handful of private evaluation firms can easily build competitive advantages in a limited market. Methodologically, the pool of techniques used is also restricted. Furthermore, peer reviews are hard to implement, as everybody knows one another. Finally, and importantly, the restriction of a small domestic evaluation market in Belgium can cast doubt on the traditional argument that external evaluations are more independ-

ent and objective than evaluations conducted by organization employees. Such a small market makes commissioners and evaluators inevitably mutually dependent (Brans, Pattyn, & De Peuter, 2011).

Evaluation demand of civil society organizations is another important element. As claimed by several respondents, civil society organizations often insist on having an evaluation conducted by an external party, rather than by the administration itself. Although perhaps not always correct, the general image that external evaluators are more independent and objective compared with in-house evaluators is a major argument in this respect. A civil servant put it this way: "[T]he administrator-general is very formal in this regard. He always says: even if we can conduct the evaluation itself, if externals do it, it will be accepted, and it is objective … and much better … He formulates it in a cynical way. But he realizes that if we conduct the evaluation in-house, people will be much more critical about the objectivity of the results" (interview with DEPT8, our translation). The name on the report is often key to lending credibility to an evaluation.

The MSDO/MDSO analysis further hints at the explanatory power of *ministerial demand for evaluations*. The organizations confronted with evaluation requests from the minister tend to conduct more in-house evaluations than those where these requests are not present. Explaining this tendency is not obvious. We can speculate that a minister prefers to keep evaluations within the circle of the administration to better control the findings, although we admit that the personality and background of the minister in charge can also have some influence in this regard. Although this requires further research, it is revealing to observe that the department (DEPT8) subjected to the minister with the strongest scientific background dominantly relies on external expertise.

As to the conditions relating to the institutional setting, the *anchorage of the evaluation function* seems powerful. On one hand, many of the organizations with a tendency to outsource evaluations have staff at their disposal who are charged with the follow-up of the evaluations and/or who arrange steering committee or working group meetings. On the other hand, some organizations also indicate that their evaluation unit enables them to conduct internal evaluations. Yet, while potentially functional for in- and ex-house evaluations, the anchorage of an evaluation function has more explanatory power for the cases with a bias to external evaluations. The follow-up of an external evaluation, and also the assessment of the offers received

from external evaluators, requires the appropriate steering of the organization. This is facilitated when an organization has staff at its disposal for whom evaluation is one of the core tasks. With the exception of the organizations that did not conduct evaluations before the NPM reforms, all cases that dominantly outsource their evaluations have an institutionalized evaluation unit.

Focusing on policy issue characteristics, the organizations that usually outsource their evaluations are overall more characterized by *outcomes that are hard to measure*. Importantly, all but one of the organizations to which this applies are departments. These indeed have been formally entrusted with the evaluation task since the NPM reforms took place. Tasks of policy preparation and evaluation are admittedly more difficult to measure than policy interventions. In such a situation, input from an external party is desirable to successfully complete the evaluation.

As for the "historical track" of the organization, the length of evaluation experience matters. Not surprisingly, *organizations will need some time before taking up a large share of in-house evaluations*. This finding is relevant in light of the importance attributed to the introduction of the NPM philosophy in the public sector for the increase in internal evaluations (Mathison, 2011). In Flanders, the organizations that became engaged in evaluation only after the NPM-oriented reforms predominantly outsource their evaluation.

Organizational stability, in turn, is also a key distinguishing factor. In times of severe reshufflings, organizations seem to be more inclined to conduct their evaluations in-house. This is an interesting observation, especially when we take into account that evaluation requires change (Wildavsky, 1972). Kiesling (2000, p. 130) argued that evaluations involving double-loop learning are usually conducted by external evaluators. Double-loop learning can concern a painful questioning of an organization's policies and value. We can speculate that stable organizations are more ready to consider the kind of double-loop changes suggested by an external party. Another reason for our observation is presumably related to the scale of the evaluations that are usually outsourced. Because outsourcing is mainly reserved for large-scale evaluands, "solid" organizations will be better positioned to implement these. The involvement of an external evaluator is usually more expensive than conducting an internal evaluation.

Notice that the availability of an adequate evaluation budget does not have the most explanatory strength. Also, the organizations with a bias toward external evaluations often mention struggling to get the funds necessary for an evaluation. Organizational size, measured in terms of staff and budget, will be comparably more important. As we can derive from our data table, the conduct of dominantly external evaluations remains a privilege of the largest organizations. Organizations in charge of large budgets usually have more spare resources at their disposal, which allow them to afford more expensive evaluations (Carpenter, 2001). But while shared by the cases that dominantly outsource evaluations, organizational size is not a key explanatory factor.

The MSDO/MDSO analysis yields a largely, but not entirely, inverse picture for the cases with a tendency for internal evaluations. A factor characterizing the cases with a tendency for in-house evaluation practice is the *absence of legislative evaluation requirements*. Indeed, where evaluation clauses exist, the lawmaker often stipulates that the study should be conducted by an external party.

Next comes the *absence of competition* as a powerful discriminatory condition. Although our empirical evidence is scarce, we can speculate that in case of competition, the need for perceived objectivity will be more pertinent than in the case of noncompetition. An external party can provide the necessary ammunition to justify the organization's existence. Conley-Tyler (2005) also seems to suggest this in her checklist for deciding between internal and external evaluations. For "sensitive evaluations" (p. 9), she advises proceeding with external evaluators.

Note, finally, that the presence of the skills to conduct internal evaluations is an obvious requirement for the conduct of in-house evaluations. An organization may be keen on conducting internal evaluations, but if it is lacking the skills to do so, it will be more likely to turn to an external party. All cases with a tendency for in-house evaluations share this characteristic. But this condition does not belong to the most critical explanatory factors to distinguish between the different outcomes.

DISCUSSION AND CONCLUSION

An evaluation can be conducted in-house or can be outsourced to an external party. To date, the evaluation literature is not conclusive as to the best strategy to follow. It should be clear, however, that an

organization will not always have full discretion to decide on the locus for evaluation implementation. No matter the extent of belief of organizations in self-evaluation (cf. Wildavsky, 1972), and irrespective of the purpose and ultimate users of the study (Sonnichsen, 2000; Vedung, 2009; Volkov, 2011), certain attributes often "push" the organization in one or another direction. Via a systematic pairwise comparison of attributes of 18 organizations of the Flemish public sector, we were able to indicate the conditions that seem to matter most in determining the locus of policy evaluation implementation. We investigated conditions that are actor-related, but also conditions that are of a more structural nature. Within the scope of this article, we were not able to unravel their respective weight, but overall, the structural conditions were more frequently indicated as relevant than the actor-related ones.

These findings are enlightening not only for theoretical purposes. They can also inspire practice and complement existing checklists for choosing between internal and external evaluations (see, e.g., Conley-Tyler, 2005). We should like to emphasize the role of five conditions that appear decisive to explain organizations' bias to one or the other locus.

1. *Organizational stability matters*. An unstable setting discourages organizations from engaging in large-scale external evaluation, which can involve double-loop learning.
2. We noticed that the organizations without a pre-reform *evaluation track* all rely mainly on external evaluators. We can speculate that these organizations might shift to in-house evaluations with the gradual development of internal knowledge.
3. *Competition*, or at least the perception of competition with other organizations, turns out to be important. Organizations that operate in a competitive environment seem to feel more comfortable with external evaluators. Although not always objectively true, external evaluators are considered to have a more objective view than in-house staff members. The results can back up the commissioning organization's legitimacy.
4. The *measurability of organizational outputs and outcomes* matters. Where outputs and outcomes are difficult to observe, expertise of outsiders is called in more readily.
5. The *presence of an evaluation unit* is another valuable resource. Although an evaluation unit can be helpful for

> the execution of both internal and external evaluations, it seems a vital advantage especially for the latter. After all, the outsourcing of evaluations does not imply a complete abandonment of responsibilities in the evaluation. Should the external evaluation meet the expectations, it requires appropriate steering and follow-up. An evaluation unit can in this regard carry out a useful brokerage role.

We do not defend a particular evaluation locus in favour of another. As Mathison (2011, p. 19) has stated: "[A]sking whether one or the other is better is not a question that can be answered generically." This is not to say that both stances are completely neutral. In the introduction, we explicitly underlined the possible impact of the evaluation locus on the quality of evaluation findings and their use. Particularly in an evaluation setting where the pool of evaluators is very limited, the question can be raised whether it is desirable that an organization fully relies on external evaluators. From a sustainability perspective, we can argue that organizations should ideally be able to conduct a minimum of the evaluations themselves. This would be in line with Wildavsky's dream of the self-evaluating organization (Wildavsky, 1972). As expressed by one of our respondents, "we should not become too dependent on them [i.e., external evaluators]" (DEPT2). Time should reveal whether we will indeed see a development toward more internal evaluations in the long run in Flanders, in line with the global trend (Sonnichsen, 2000; Mathison, 2011).

This article has been restricted to an identification of individual conditions that appear most decisive in distinguishing most-similar organizations with different outcomes and most-different organizations with similar outcomes. The MSDO/MDSO technique is often used before moving to a combinatory analysis of conditions. A qualitative comparative analysis (Ragin, 1987, 2000) with these conditions as input would be a logical avenue for further research. Future studies should also test the extent of external validity of our findings. Are our results unique to a region that has only recently started to develop an evaluation culture, or can we claim wider generalizability?

NOTES

1 As an assisting tool, we relied on the MSDO/MDSO software (beta-version 8/7/2006), developed by G. De Meur (available at http://www.jchr.be/01/beta.htm).

2 Within the scope of this article, we list only the results of the MSDO/MDSO analyses. The full analyses can be requested via the authors.

REFERENCES

Brans, M., Pattyn, V., & De Peuter, B. (2011, September). *The evaluation of labour market policies in Belgium: A meta-analysis*. Paper submitted in the framework of the European Commission's Mutual Learning Programme: Peer review of Evaluation of labour market policies and programmes: Methodology and practice. United Kingdom.

Bursens, P. (1999). *Impact van instituties op besluitvorming. Een institutioneel perspectief op besluitvorming in de communautaire pijler van de Europese Unie* [The impact of institutions on decision-making. An institutional perspective on decision-making within the first pillar of the European Union] (Doctoral thesis, University of Antwerp, Antwerp, Belgium). Antwerpen, Belgium: UA.

Carpenter, D. P. (2001). *The forging of bureaucratic autonomy: Reputations, networks, and policy innovation in executive agencies, 1862–1928*. Princeton, NJ: Princeton University Press.

Conley-Tyler, M. (2005). A fundamental choice: Internal or external evaluation? *Evaluation Journal of Australasia*, *4*(1 & 2), 3–11.

Dahler-Larsen, P. (2012). *The evaluation society*. Stanford, CA: Stanford University Press.

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*(2), 193–219. http://dx.doi.org/10.1111/j.1475-6765.1994.tb00440.x

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. http://dx.doi.org/10.1177/0010414096029004003

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, UK: Sage. http://dx.doi.org/10.4135/9781446249413.n12

De Meur, G., Rihoux, B., & Yamasaki, S. (2009). Addressing the critiques of QCA. In B. Rihoux & C. C. Ragin (Eds.), *Configurational comparative methods: Qualititative comparative analysis (QCA) and related techniques* (pp. 147–178). Thousand Oaks. CA: Sage. http://dx.doi.org/10.4135/9781452226569.n7

Fetterman, D. (2001). *Foundations of empowerment evaluation*. Thousand Oaks, CA: Sage.

Furubo, J. E., Rist, R. C., & Sandahl, R. (Eds.). (2002). *International atlas of evaluation*. Piscataway, NJ: Transaction.

IAVA. (2007, 2008, 2009). *Jaarverslag van het Auditcomité en het Agentschap Interne Audit van de Vlaamse Administratie* [Annual Report of the Audit Committee and the Agency Internal Audit of the Flemish Administration]. Brussels, Belgium: IAVA.

Kiesling, H. J. (2000). *Collected goods, Neglected goods: Dealing with methodological failure in the social sciences.* River Edge, NJ: World Scientific. http://dx.doi.org/10.1142/4077

Mathison, S. (1991). What do we know about internal evaluation? *Evaluation and Program Planning*, *14*(3), 159–165. http://dx.doi.org/10.1016/0149-7189(91)90051-H

Mathison, S. (2011). Internal evaluation, historically speaking. In B. B. Volkov & M. E. Baron (Eds.), *Internal evaluation in the 21st century, New Directions for Evaluation, No. 132*, 13–23.

Mill, J. S. (1973 [1843]). Of the four methods of experimental inquiry, chap. 8. In *The Collected Works of John Stuart Mill* (Vol. 7, *A system of logic ratiocinative and inductive*). London, UK: Routledge and Kegan Paul.

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

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

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

Scriven, M. (1980). *Evaluation Thesaurus* (4th ed.). Newbury Park, CA: Sage.

Sonnichsen, R. C. (1994). Effective internal evaluation: An approach to organizational learning. In F. L. Leeuw, R. C. Rist, & R. C. Sonnichsen (Eds.), *Can governments learn?* (pp. 125–141). New Brunswick, NJ: Transaction.

Sonnichsen, R. C. (2000). *High impact internal evaluation: A practitioner's guide to evaluating and consulting inside organizations*. London, UK: Sage.

Vedung, E. (2009). *Public policy and program evaluation* (4th ed.). New Brunswick, NJ: Transaction.

Vlaams Parlement. (2003). *Kaderdecreet Bestuurlijk Beleid* [Framework Decree 'Better Administrative Policy']. Brussels, Belgium: Author.

Volkov, B. B. (2011). Internal evaluation a quarter-century later: A conversation with Arnold J. Love. In B. B. Volkov & M. E. Baron (Eds.), *Internal evaluation in the 21st century. New Directions for Evaluation, No. 132*, 5–12.

Wildavsky, A. (1972). The self-evaluating organization. *Public Administration Review*, *32*(5), 509–520. http://dx.doi.org/10.2307/975158

Worthen, B. R., Sanders, J., & Fitzpatrick, J. (1997). *Program evaluation: Alternative approaches and practical guidelines* (2nd ed.). New York, NY: Longman.

Valérie Pattyn is affiliated with Leuven University Public Management Institute, Belgium. The article presents part of her PhD research that focuses on explaining organizational variety in evaluation praxis. Other research interests include comparative methodologies, policy advisors and policy advisory systems, and evidence-based policymaking.

Marleen Brans is a Professor in Public Administration and Policy at Leuven University Public Management Institute, Belgium. Her research interests include politico-administrative relations, policy analytical capacity of civil service systems, and interactions between government and civil society.

Appendix 1
Overview of Conditions, Their Indicators, and Dichotomization Thresholds

Conditions		Code 0 indicators	Code 1 indicators
Category A: Capabilities of the organization			
(1)	Skills to conduct evaluations	Totally insufficient, rather insufficient	Rather sufficient, fully sufficient
(2)	Financial means to evaluate		
(3)	Availability of capable staff		
(4)	Availability of external evaluators		
(5)	Availability of monitoring information		
Category B: Orientations			
Extent of evaluation demand from:		No demand, hardly any demand	Sometimes demand, frequent demand
(6)	Organizational management		
(7)	Sector minister		
(8)	Parliament		
(9)	Civil society organizations		
(10)	Other organizations		
(11)	Extent of organization-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)	Organizational size	Very low, low material weight (*)	At least average material weight
(13)	Organizational autonomy	No legal personality	Legal personality
(14)	Organizational status	Agency	Department
(15)	Anchorage of evaluation function	No evaluation unit	Formal or de facto evaluation unit
(16)	Participation in evaluation community	No engagement in evaluation trainings or networks	Minimally "sometimes" participating in evaluation trainings or networks
Evaluation requirements stipulated in:		No evaluation requirements	Evaluation requirements
(17)	regulation or decrees at Flemish level		
(18)	legislation/regulation at EU level		
(19)	management agreement of organization		

Conditions		Code 0 indicators	Code 1 indicators
Category D: Conditions concerning policy issue characteristics			
(20)	Attention by media or parliament for the tasks of the organization	Not at all, limited, rather limited	Rather much, much, very much
		Highest score of the assessments of attention by each of these 'institutions"	
(21)	Perceived competition on tasks of the organization	Not at all, limited, rather limited	Rather much, much, very much
(22)	Perceived measurability of outputs and	Average score ≤ 3 and/or qualification: very difficult, difficult, rather difficult	Average score ≥ 3 and/ or qualification very easy, rather easy, easy
(23)	outcomes		
		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 organization	
Category E: Conditions characterizing the path of the organization			
(24)	Pre-NPM evaluation experience	No/seldom evaluation practice before the BAP reforms	Sometimes/frequent evaluation practice before the BAP reforms
(25)	Organizational stability	Organizations which underwent medium or large changes (**)	Organizations which underwent no or small changes
(26)	Ministerial stability	≥ 1 minister changes since BAP	No ministerial turnover

(*) The indicator concerns both financial material weight (50%) and material weight with regard to personnel (50%). For financial material weight, the following scales are used [in 10,000 EUR]: (1) very low material weight: 0–50,000; (2) low material weight: 10,000–50,000; (3) average material weight: 50,000–100,000; (4) high material weight: 100,000–500,000; (5) very high material weight: < 500,000. 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 for the years 2007–2008–2009 (IAVA, 2007, 2008, 2009).

(**) Four subcriteria constitute this indicator. Three of them relate to the impact of the NPM-oriented reforms (which account for 60% of the total indicator): (a) changes in the form of management/steering of the organization; (b) changes with regard to the composition of the public entity; (c) 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 those organizations that underwent a large number of changes; 0.3 for those that underwent a medium number of changes; and 0.1. for those organizations that can be characterized by large stability. We calculated the average for the years 2007–2008–2009 (IAVA, 2007, 2008, 2009).

Appendix 2
Binary Data Table

		Category A: Actor capabilities					Category B: Actor orientations							Category C: Institutional setting									Category D: Policy issue character-istics			Category E: Path of organization
	Case (*)	*Skills to conduct evaluations*	*Financial means to evaluate*	*Availability of capable staff*	*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 organizations*	*Evaluation demand from other organizations*	*Organization-wide support for evaluations*	*Organizational size*	*Organizational autonomy*	*Organizational status*	*Anchorage of evaluation function*	*Participation in evaluation community*	*Evaluation requirements in management agreement*	*Evaluation requirement in Flemish legislation*	*EU international evaluation requirements*	*Media/parliamentary attention for organization*	*Competition*	*Measurability of outputs*	*Measurability of outcomes*	*pre-NPM evaluation experience*	*Organizational stability*	*Ministerial stability*
Cases with tendency to conduct internal evaluations	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	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	0	1
	INTALP3	1	1	1	0	1	1	0	0	1	1	1	1	1	0	1	0	1	1	0	1	0	1	0	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	0	0
	ILTALP1	1	1	1	0	1	1	1	0	0	0	1	0	1	0	1	0	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	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	1

	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
	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
	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
Cases that outsource most evaluations	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
	EXTA1	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
	DEPT1	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
	DEPT5	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
	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
	DEPT6	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 of 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.