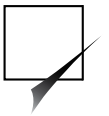

Article

Evaluation
2015, Vol. 21(1) 6–31

DOI: 10.1177/1356389014564248
evi.sagepub.com

The institutionalization of evaluation matters: Updating the *International Atlas of Evaluation* 10 years later

Steve Jacob
Laval University, Canada

Sandra Speer
Independent evaluator, Germany

Jan-Eric Furubo
National Audit Office, Sweden

Abstract
This text provides a comparative cross-country analysis of evaluation culture and the institutionalization of evaluation. The countries included in this research are the 19 OECD countries examined by the authors of the *International Atlas of Evaluation* 10 years ago (Furubo et al., 2002). The analysis is based on the results of an expert survey of four to five evaluation experts from different backgrounds for each country, as well as additional information from the literature. Using the nine indicators from Furubo et al. (2002) with a focus on the institutional characteristics of reforms, trends in evaluation culture over the last decade have been identified.

Keywords
Evaluation culture, institutionalization, Evaluation society, Supreme Audit Institution, expert survey

Introduction

Administrative organizations in modern welfare states have historically been exposed to accountability pressures and management influences. Evaluation plays multiple roles:

Corresponding author:
Steve Jacob, Faculty of Social Sciences, Department of Political Science, Laval University, Charles-de Koninck Pavillon, Office 4443, 1030 Avenue des Sciences Humaines, Quebec, QC, Canada, G1V 0A6.
Email: steve.jacob@pol.ulaval.ca

> Both politically, in terms of being accountable to those who fund the system, and also ethically, in terms of making sure that you make the best use possible of available resources, evaluation is absolutely critical. (Julio Frenk, Mexican Health Minister quoted by Oxman et al., 2010: 427)

However, few normative claims exist regarding how evaluation should be embedded in the architecture of governance. Exceptions include a rationale for mandatory impact evaluations (Oxman et al., 2010) and a call for a centralized evaluation function as recommended by Hallsworth and Rutter (2011: 33):

> The government's Head of Policy Effectiveness should take a significant role in evaluations. She or he would receive a proportion of departments' current evaluation spending to establish an institutional base that had three main functions: to oversee departmental commissioning; to run an open evaluation commissioning process; and to commission its own lessons learned reviews in cases of exceptional policy failure. The Head of Policy Effectiveness should ensure that general lessons emerging from evaluations are incorporated into policy making guidance.

Despite such assertions supporting evaluation practice and promoting its institutionalization, there is little discussion about forms of institutionalization with their advantages and disadvantages. However, differing national evaluation cultures have emerged (Barbier and Hawkins, 2012), as can be seen in the *International Atlas of Evaluation* (Furubo et al., 2002; hereafter the *International Atlas*), which provides an in-depth overview of the differences between evaluation cultures in various national settings. National policy styles can shape patterns of policymaking in systems of public administration, and it can be assumed that some of these national characteristics have an impact on evaluation regardless of the particularities of different policy fields and organizations. The *International Atlas* provided the first systematic comparative overview of evaluation cultures within a framework of selected indicators measuring nine dimensions. In 2001, evaluation cultures in 21 nations were described and analyzed.[1] Since then, an increasing number of evaluations have been undertaken and, due to external and internal pressures, the evaluation cultures in many countries have been strengthened. Some country-based studies have been published in recent years (e.g. Bussmann, 2008; Leeuw, 2009; Feinstein and Zapico-Goñi, 2010; Jacob and Slaïbi, 2014; Smits and Jacob, 2014); however, little systematic comparative research across countries exists (Jacob and Varone, 2004; Jacob, 2005a; Widmer et al., 2009) and this body of research is still at a relatively early stage. Gaarder and Briceño (2010) analyze aspects of evaluation institutionalization across developing countries classified as having low and medium income (Mexico, Columbia, Chile, South Africa, China), which are less comparable to the political and administrative systems of the OECD countries reviewed here.

Evaluation can follow various designs, is embedded in different forms of institutionalization, and has widely varying usages within different sectors and on different levels. The development of evaluation culture does not follow a one-dimensional model. This makes developments empirically difficult to capture and these challenges are further compounded by the varied historical roots for governance. Being conscious of these complexities and challenges, we sought to analyze the diverse forms and functions of evaluation culture in 19 OECD countries around the world.

To create a systematic overview 10 years after the initial study, a comparison of these original findings with current developments in evaluation culture was undertaken. This article investigates trends from the last decade and identifies conditions for strong evaluation cultures as well as paths of development, which can vary significantly. Ten years is an appropriate timeframe for examining

an issue such as institutionalization and its effects as well as the development of the supply side. Reforms require many years to become effective, although in such a timeframe, in some cases, existing institutions might be dismantled and set up again. In this article, we begin by discussing the methodology for measuring institutional and cultural changes in evaluation and also specify the mechanisms for such changes. Is there a converging trend of evaluation institutions and practices across these countries? This is followed by a discussion of the results, indicator by indicator and by clusters of trends. Finally, prospects for further research are discussed.

Methodology

Several reviews of (individual) national as well as sectorial evaluation cultures have been carried out since 2001. It is difficult to fully grasp a national evaluation culture. There is no single way of measuring it and changes in critical dimensions in several institutions need to be tracked. Furubo et al. (2002) refer to the following nine indicators:

I. evaluation takes place in many policy domains;
II. there should be a supply of evaluators specializing in different disciplines;
III. discussions and debates fuel a national discourse regarding evaluation;
IV. a national evaluation society exists;
V. institutional arrangements in the government for conducting evaluations and disseminating their results exist;
VI. institutional arrangements in Parliament for conducting and disseminating evaluations exists;
VII. pluralism exists within each policy domain;
VIII. evaluation activities occur within the supreme audit institution; and
IX. evaluations do not just focus on inputs/outputs, but also on outcomes.

Another study of institutionalization of evaluation in various countries (Varone and Jacob, 2004) made distinctions between the presence of national evaluation bodies and the epistemic evaluation community. The existence of evaluation bodies was analyzed within the executive branch of government (= V), the parliament (= VI) and the supreme audit institution (= VIII). The epistemic community is defined by the existence of a national society (= IV), a scientific journal as well as quality standards. 'Scientific journal' and 'quality standards' are not covered by the framework in the *International Atlas* and could be interpreted as an extension of Indicator IV ('profession with its own societies'). Many elements are very similar despite the fact that both research teams were working on these indexes at the same time without knowing each other. However, both indexes emphasize indicators that are situated more upstream and somewhat indirectly linked to the effective evaluation praxis. Indeed, in order to measure the degree of institutionalization and the effective practices of evaluation, the most obvious approach would be to use results indicators such as the number of conducted evaluations or meta-evaluations, or, alternatively, process indicators such as the number of requests for proposals or the proportion of the public budget dedicated to evaluation. Unfortunately, these statistical elements are, for the most part, unavailable since evaluation is much too polycentric.

Ten years ago, scores for evaluation cultures within the *International Atlas* were exclusively based on subjective views by the authors of that time as well as the individual contributors of individual *Atlas* chapters reflecting their perceptions. The scores were produced in an iterative process between the editors and the chapter authors. Out of the previous 21 countries, 19 were included in

this survey. Zimbabwe and China were excluded because of very different political environments and for a better comparability between the more homogeneous 19 OECD countries included.

Instead of rewriting the individual chapters of the *International Atlas*, a survey of evaluation experts with a broad knowledge of their respective national evaluation landscape was conducted. Although the comparison still relies on subjective expert views, the basis for the comparison has been widened. Given the nature of our research, the expert survey appeared to offer the most appropriate approach. We could not use a classic survey of the members of an evaluation society, or of political or administrative officeholders, since they do not have a suitably detailed view of all the evaluation mechanisms present in a given country. A review of the literature to obtain information would also not have been appropriate, given that detailed national studies are rare and were not available for all the countries in our sample.

For over 30 years, expert surveys have been used in the field of political science to position political parties in a policy space. A corpus of literature has built up in the field, and the use of the method itself has also been scrutinized. To summarize, the main strengths of the method mentioned in the literature are: flexibility about issues and topics and validity since experts use multiple sources of information to base their judgment. The weaknesses are: subjective judgment; informational asymmetry among respondents; conflating preferences and behaviour (do experts evaluate rhetoric or action?) and temporal constraints on retroactive judgment (Huber and Inglehart, 1995; Budge, 2000).

For each country, five experts from three different backgrounds – public, private and academic – were invited to participate, ensuring the inclusion of broader perspectives. In the end, seventy-eight evaluation experts participated in this survey administered from April to September 2011. For each country, four or five experts from various backgrounds were included. Authors who wrote a chapter in the *International Atlas* were invited to participate in the new expert survey if they could be contacted. Additional experts were identified by literature searches, previous personal knowledge and the help of a snowball-system.

To ensure comparability, the same indicators and the same scale were used as previously in Furubo et al. (2002). Every expert was asked to give a rating according to the explanatory text (see Box 1) and to comment on it. Additionally, five open-ended questions were included on the main changes in evaluation culture: the triggers for the institutionalization of evaluation, the utilization of evaluation, the suggestion of relevant documents or literature, and the possibility for further comments. Answers to these additional questions helped in interpreting the data and explaining changes. The questionnaire was pre-tested with six international experts. Experts were not provided with the ratings from 2001, but, of course, any of them could look this up in the *International Atlas*.

This research relies on 'subjective' assessments because there is a scarcity of 'objective' data on evaluation, and only in this way could country specific contexts and nuances be taken into account. In general, the scores provided by the experts varied, but were globally convergent. The results presented in Table 1 show the average scores provided by the national experts for each dimension. This approach allowed us to reflect the perceptions the experts have of evaluation practices in their own country. The text of the article has not been validated with the experts. Several national experts have provided almost the same rationale to explain their ratings. When available in the literature, information provided by the experts has also been triangulated with the literature. In conclusion, we are aware that the experts selected may have a biased perception of the situation in their own country. For this reason we used several experts for each country in order to obtain a more precise and detailed view. This represents an improvement compared to the *International Atlas of Evaluation*.

Box 1. The Nine Indicators from the *International Atlas of Evaluation*.

I. Evaluation takes place in many policy domains: There are frequent evaluation activities within various policy fields.

0: If evaluation activities take place only in a very limited part of the public sphere, perhaps in only one policy domain or only in relation to one or two programs or in relation to externally funded programs (e.g. EU or World Bank funded programs), we regard evaluation as an isolated activity, and the country will get a score of 0.

1: A score of 1 shall be given to countries where evaluation activities are clearly frequent, but where they are not regarded as an integrated part of the whole public sector.

2: To get a score of 2, evaluation activities must be taking place in most of the public sector.

II. Supply of domestic evaluators from different disciplines: There is a supply of evaluators from different academic disciplines who have mastered different evaluation methods and who conduct and provide advice over evaluations. This criterion is also intended to grasp the diffusion and pluralism of evaluation praxis in a country.

0: Countries where there exist perhaps only a handful of institutions conducting evaluations with a rather monolithic perspective get a score of 0.

1: Countries somewhere in-between these two positions receive a score of 1.

2: Countries with a flourishing supply of evaluators in which evaluative problems are seen from different perspectives, and with evaluators from different disciplines specializing in different methods, will receive a score of 2.

III. National discourse concerning evaluation: There is a national discourse concerning evaluation in which more general discussions are adjusted to the specific national environment.

0: Countries where the discussion is totally based on 'imported goods' get a score of 0.

1: The countries in between get a score of 1.

2: A score of 2 will been given to countries in which it is obvious that discussions about questions such as organizational structures, systems for training evaluators, evaluation utilization as well as potential adverse effects result from the country's own national experience and preconditions.

IV. Professional organizations: Evaluators have their own societies, networks or frequent attendance at meetings of international societies and at least some discussion concerning evaluation standards or ethics.

0: A score of 0 is reserved for countries with only ad hoc meetings.

1: Countries without societies but where meetings are held on a more or less regular basis receive a score of 1.

2: Countries that have networks or societies for evaluators get a score of 2.

V. Degree of institutionalization – Government: Institutional arrangements in the government for conducting evaluations and disseminating their results to decision makers. In several countries, a large number of evaluations are conducted, but their results seem to reach decision makers more by chance than anything else. This criterion attempts to take into consideration permanent arrangements or systems whereby evaluation initiatives are commissioned to different evaluators and, at the same time, arrangements are developed to ensure that the evaluations conducted are put to suitable use. Examples for this kind of institutionalization can be central evaluation units in various national ministries or the existence of an evaluation budget. This is a form of guarantee that utilization – at least in formal terms – will take place.

(Box Continued)

Box 1. (Continued)

0: Countries lacking such arrangements get a score of 0.

1: A score of 1 is an 'in-between-value.'

2: Countries with well-developed structures and processes for conducting and disseminating evaluations get a score of 2.

VI. Degree of institutionalization – Parliament: Institutional arrangements are present in parliament for conducting evaluations and disseminating them to decision makers. This criterion tries to cover the same kind of arrangements as criterion V, but this time at the parliamentary level. The reason for having the same criterion for parliament is that we find it more likely that different political groups will be involved and perhaps other kind of evaluative questions will be raised if the initiative comes from the parliamentary sphere. Members of parliament may be involved in the ad hoc initiation of evaluation, but also evaluation clauses are introduced into laws as a means of finding political compromises, or subject committees conduct respectively tender evaluations. Furthermore, evaluation results may be utilized in the parliamentarian debate.

0: Countries lacking such arrangements get a score of 0.

1: A score of 1 is an 'in-between-value'.

2: Countries with well-developed institutionalization for conducting and disseminating evaluations get a score of 2.

VII. Pluralism of institutions or evaluators performing evaluations within each policy domain: An element of pluralism exists, that is, within each policy domain there are different people or agencies commissioning and performing evaluations. This criterion is obviously intended to capture the degree of pluralism. If we imagine that we have only one very dominant organizational entity in a policy domain, which at once formulates the evaluative problems, decides which evaluators to use, and thus also decides what kind of methods to employ, etc., there is no scope for pluralism. A country with this kind of situation is regarded as less mature than a country in which there are a number of commissioners and conductors of evaluations.

0: A score of 0 is given to countries with a very monolithic structure.

1: A score of 1 is for countries in the middle.

2: A score of 2 is given to countries with a high ranking.

VIII. Evaluation within the Supreme Audit Institution: The existence of evaluation activities within the Supreme Audit Institution (SAI) can be of different kinds. The SAI might conduct evaluation activities themselves (e.g. Performance/Value for Money Audits) or look at conditions for undertaking evaluations within the public sector or even carry out different forms of meta-evaluation.

0: Where evaluation is absent, the score shall be 0.

1: A country which has evaluative activities within the SAI, but not to the same extent, or to countries which have only recently brought evaluation into the activities of their SAI, gets a score of 1.

2: A score of 2 shall characterize countries in which evaluation plays an important part in fulfilling the activities of the SAI.

(Box Continued)

Box 1. (Continued)

IX. Proportion of impact and outcome evaluations in relation to output and process evaluations: The evaluations conducted should not just be focused on the relation between inputs/outputs or technical production. Some public sector evaluations must show program or policy outcomes as their object and raise such questions as whether the public interventions actually had impacts on the problems they were intended to solve.

0: A score of 0 is given to countries that seem to concentrate too heavily on input/output measurements or on the production process itself.

1: A score of 1 is given to countries in between.

2: A score of 2 is given to countries with a very pluralistic set of activities in this respect.

Source: Furubo et al. (2002: 7–9) with some adaptations of the explanatory text.

The lay of the land in 2011

According to other information on national evaluation cultures besides this survey and our own knowledge from many countries, the data does not show a seam effect[2]. On the contrary, in individual cases, the previous scores have been readjusted with decreased results. The choice of repeating the survey with the previous scale was not without constraints; ceiling effects may have occurred for countries which already had the highest scores in 2001 and could not reflect further improvements and variations in the scores. The interpretation of the data attempts to account for this.

Table 1 presents the scores for the 19 countries included in our sample. Based on these results, we divided the sample into three categories according to the respective degree of evaluation culture maturity in each country. A high degree of maturity is defined by a score of 12 or higher, a score between 6 and 11.9 represents a medium degree of maturity, and countries with a score lower than 6 have a low degree of evaluation culture maturity:

- High degree of maturity (n = 15): Australia, Canada, Denmark, Finland, France, Germany, Israel, Japan, the Netherlands, Norway, South Korea, Sweden, Switzerland, the United Kingdom, the United States;
- Medium degree of maturity (n = 4): Ireland, Italy, New Zealand, Spain;
- Low degree of maturity (n = 0).

The vast majority (79%) of the countries included in our sample showed a high degree of evaluation culture maturity. Four countries (21%) were in the middle of the scale while no country (0%) was characterized by a low degree of maturity.

In order to determine which factor influenced scores, we compared and contrasted data for the countries at the top and the bottom of the ranking (see Table 2). The ‘Top 3’ is composed of Finland (16.6), Switzerland (16.4), and Canada (16), while the ‘Bottom 3’ includes Ireland (9), Italy (10.7), and Spain (11.3).

In 2011, the differences between countries are less contrasted than in the past (see next section). The average score of countries with the highest degree of maturity is 16.3 while

Table 1. Evaluation Culture in 2011.

	I. Domains	II. Disciplines	III. Discourse	IV. Profession	V. Inst. – Government	VI. Inst – Parliament	VII. Pluralism	VIII. SAI	IX. Impact	SUM
Australia	1,3	1,7	1,7	2,0	0,7	1,0	1,7	2,0	1,7	**13,7**
Canada	2,0	2,0	2,0	2,0	1,8	0,8	2,0	1,8	1,8	**16,0**
Denmark	1,8	1,8	1,8	2,0	1,3	1,0	2,0	1,5	1,3	**14,3**
Finland	2,0	2,0	1,8	2,0	1,8	1,2	2,0	2,0	1,8	**16,6**
France	1,6	1,4	1,8	2,0	1,4	1,2	1,2	1,0	1,4	**13,0**
Germany	1,3	2,0	1,3	1,8	1,0	1,0	2,0	1,3	1,5	**13,3**
Ireland	1,0	1,3	1,5	1,0	1,0	0,3	1,3	1,0	0,8	**9,0**
Israel	1,3	1,8	1,0	1,8	1,3	1,0	1,8	1,3	1,3	**12,3**
Italy	1,7	1,7	1,3	2,0	1,3	0,7	1,0	0,3	0,7	**10,7**
Japan	2,0	1,8	1,5	1,3	2,0	0,3	1,5	1,3	1,3	**12,9**
Netherlands	2,0	1,9	1,5	1,8	1,8	1,5	1,8	1,8	1,4	**15,3**
New Zealand	1,4	1,0	1,4	2,0	1,2	0,6	1,4	1,4	1,2	**11,6**
Norway	1,9	1,5	1,1	1,8	1,4	0,9	1,8	1,8	1,3	**13,5**
South Korea	2,0	2,0	1,7	1,7	2,0	1,7	1,7	1,3	1,3	**15,3**
Spain	1,3	1,8	1,5	2,0	1,3	0,5	1,3	0,3	1,5	**11,3**
Sweden	1,8	1,6	1,6	1,8	1,8	1,4	1,6	1,7	1,6	**14,8**
Switzerland	1,8	2,0	1,6	2,0	1,3	2,0	1,8	2,0	2,0	**16,4**
United Kingdom	2,0	2,0	1,5	2,0	1,5	1,3	2,0	1,8	1,3	**15,3**
United States	1,6	2,0	1,8	2,0	1,8	1,4	1,6	1,8	1,8	**15,8**
Mean	**1,7**	**1,8**	**1,5**	**1,8**	**1,5**	**1,0**	**1,7**	**1,4**	**1,4**	**13,7**
Top 3	1,9	2,0	1,8	2,0	1,6	1,3	1,9	1,9	1,9	16,3
Bottom 3	1,3	1,6	1,4	1,7	1,2	0,5	1,2	0,5	1,0	10,3

the average score is 10.3 among the countries with the least mature evaluation culture. The difference between these two groups is only 6 points. Table 3 shows almost no distinction (deviation is equal to or less than 0.5) in term of disciplines, national discourse, professionalization, and institutionalization within the government. The strongest deviation (more than 1) results from the involvement of the SAI (supreme audit institution) in the field of evaluation.

The countries with the highest degree of evaluation maturity ('Top 3') scored high on every indicator while the countries with the lowest degree ('Bottom 3') showed more volatility on the indicators composing the index. Institutionalization within the parliament is the indicator where 'Top 3' countries and 'Bottom 3' countries received the lowest score. On the other hand, the existence of professional organizations is the indicator where both groups received their highest scores. 'Top 3 countries' obtained perfect scores (i.e. 2) or almost perfect scores (i.e. 1.8 or 1.9) on 7 indicators out of 9. 'Bottom 3' countries showed room for improvement on many indicators, especially on the institutionalization of evaluation within the parliament, the involvement of SAIs in evaluation activities and the orientation of evaluation toward impact assessment where their scores were 1 or lower.

Evaluation in different policy domains

Every country scored 1 or higher on this indicator and six countries obtained the maximum score of 2. The mean for this indicator was 1.7. These results show that policy domains are widely covered in most of the countries being examined. Nonetheless, regulations such as 'all public programs ought to be evaluated regularly' are the exception and can be found in Canada, France, Japan, and Switzerland. These regulations might encourage evaluation activities to be commissioned within all policy sectors in a given country. However, more information on the implementation of these regulations is needed to fully understand their impact on national evaluation culture. In the other countries, practices were unevenly distributed. In most of the countries in our sample, we found evaluation activities in the fields of education, health, labor markets, social policy, aid, industry policy, environmental policy, and research and development. Education, development aid, and research are often said to be the most intensively evaluated – even 'over-evaluated' as said about the field of education in Israel – and are, at the same time, the ones with the longest history in evaluation from where practices spread to other policy sectors. On the other hand, fields such as law, finance policy, defense, police, transport, and foreign policy are not evaluated with the same intensity or are not evaluated at all. Differences in policy sectors result from differing governmental structures. The rating is mainly linked to the national level and reflects the fact that most evaluation activities seem to take place at the national level. However, in federally-structured countries, responsibilities for certain policy fields lay on subordinated levels where evaluation culture might be at a developmental stage.

Supply from different disciplines

Historically, evaluators stem from different disciplinary backgrounds (Alkin, 2004; Jacob, 2008; Vaessen and Leeuw, 2010). The number of books available to introduce specialists in

Table 2. 'Top 3' and 'Bottom 3' groups in 2011.

	Ranking	Country	Score
Top 3	1	Finland	16.6
	2	Switzerland	16.4
	3	Canada	16.0
	…		
Bottom 3	17	Spain	11.3
	18	Italy	10.7
	19	Ireland	9.0

Table 3. Indicators of 'Top 3' and 'Bottom 3' groups in 2011.

	Top 3	Bottom 3	Deviation
I. Domains	1.9	1.3	0.6
II. Disciplines	2.0	1.6	0.4
III. Discourse	1.8	1.4	0.4
IV. Profession	2.0	1.7	0.3
V. Inst. − Government	1.6	1.2	0.4
VI. Inst − Parliament	1.3	0.5	0.8
VII. Pluralism	1.9	1.2	0.7
VIII. SAI	1.9	0.5	1.4
IX. Impact	1.9	1.0	0.9
SUM	16.3	10.3	6.0

various fields (social psychology, economics, social work, etc.) to evaluation practices continues to increase (Drummond and McGuire, 2001; Grinnell et al., 2011; Mark et al., 2011). As a result, every country scored 1 or higher on this indicator and seven countries even reached the maximum score of 2. The mean score for this indicator was 1.8. As a result, experts did not see the attraction of personnel from different disciplinary backgrounds as a problem, and saw a large variety in the educational opportunities offered to evaluators. In the last decade, an increasing number of economists have entered the evaluation community, which is partly perceived with criticism as a new dominance in the Netherlands and Sweden. However, the broad spectrum of social scientists including sociologists, psychologists, political scientists, public administrators, and educational scientists is represented, with specific mixtures in the various policy sectors. New varieties of evaluators are emerging from the disciplines of IT and law. Statisticians and mathematicians are noticeably less involved and in the United States, 'evaluation is less and less a preoccupation for schools of public administration and other disciplines that were involved in evaluation in the last decades. Now, evaluation is more and more driven by education. As a result, the face of evaluation evolves' (US expert).

Quantitative data informing the disciplinary background of evaluators is, for the most part, lacking. A study by the AEA showed that members of the American evaluation community come from very diverse fields (AEA, 2008); a Swiss evaluators' database also

showed disciplinary heterogeneity. This situation might be explained by the fact that evaluation is becoming increasingly integrated into higher education and, in many countries, postgraduate interdisciplinary masters programs exist. Moreover, many national societies offer interdisciplinary training, such as the Japanese Evaluation Society, which also awards a 'Certificate of Professional Evaluator'. However, some experts from Canada and France reported a certain shortage of interdisciplinary evaluators. A study in Canada (Breen and Associates, 2005) showed a lack of capacity within the government and noted that most evaluators were quite inexperienced, which has led to a system of professional designations by the Canadian Evaluation Society and a Consortium of Universities for Evaluation Education developed in response to a federal government initiative (Cousins et al., 2009). In France, a certain scarcity of interdisciplinary evaluators was also reported despite the existence of masters programs and the recent initiation of a French summer school in evaluation (2010). However, many experts saw room for further cross-fertilization between the disciplines which could translate into an adaptation of current curricula.

National discourses

Concerning the national discourse surrounding evaluation, results showed a widespread proliferation of discourses in every country in our sample. Canada received the highest score while Israel received the lowest score for this indicator (1); the mean for all scores was 1.5. This result means that evaluation and performance management issues are on the political agenda. Topics often mentioned were 'evidence-based policy', 'performance measurement' and 'credentialing evaluators'. In some countries, the financial crisis and the general macroeconomic context affected political debates. For instance, in recent years, there has been a great deal of discussion in Ireland about 'value for money', 'accountability', the need to 'control spending', and the need to 'escape the wasteful habits of the past'.

The national discourse was also fed by evaluation findings reported in the media or by criticism against the work of inspectorates (in the Netherlands) or complaints about the meaningfulness of evaluation (in Denmark). The national discourse was also centered on internal topics from the evaluation community such as the best way to institutionalize evaluation, quality improvement and evaluator credentialing. In this case, evaluation societies were, for the most part, the locus of specific discourse regarding evaluation. This discourse remains mainly within the circle of evaluators and is rarely visible to the general public. For instance, in the United States, experts mentioned the existence of a vivid discourse in public agencies, foundations and advocacy organizations and that 'the discourse has moved from whether there should be evaluation to what kind should be performed' (US expert). In Japan, the national discourse takes place mainly within the executive branch of government and among ministries and agencies.

Apart from the United States (Alkin, 2004), specific evaluation approaches and schools of thought were less existent. Exceptions could be found in the United Kingdom with the realist approach to evaluation (e.g. Pawson and Tilley, 1997). In addition, diversity and theoretical eclecticism are starting to be recognized in Europe, Australia and New Zealand (Rogers and Davidson, 2013; Stame, 2013). Alongside these theoretical developments, other initiatives reflect the concerns of national evaluation communities. Examples include France, where emphasis is placed on the 'plurality principle'

which is anchored in the French '*Charte de l'Évaluation*', and New Zealand, where discourse and developments around indigenous evaluation approaches and cross-cultural evaluation are burgeoning, going back to the history of the Maori striving for self-determination.

Professional organizations

This indicator had the least impact for determining the maturity of evaluation culture in our study. Indeed, eleven countries received the highest possible score (2), the mean was 1.8 and no blind spots existed for the countries under review. Most of the professional evaluation networks and societies already existed prior to 2001. Only five were newly founded: the Dutch Evaluation Society in 2002, the Irish Evaluation Network in 2002, the Swedish Evaluation Society in 2003, the New Zealand Evaluation Association in 2007 and the Norwegian Evaluation Society in 2009. Before the existence of the New Zealand Evaluation Society, New Zealanders were members of the Australasian Evaluation Society (AES) and have maintained their membership. Countries such as the Netherlands, Sweden and Norway already had a long history of evaluation in their country before finally organizing evaluation professionals into an association. Other countries set up evaluation societies with the emergence of a new professional field, which was, in some cases, also fostered from outside influences such as the European Union.

In addition to their respective national evaluation societies, Norway has an evaluation network operated by the Norwegian Government Agency for Financial Management, Switzerland has an Evaluation Network of the Federal Administration which exists alongside the Swiss Evaluation Society, and the United States has a National Legislative Program Evaluation Society (NLPES) which was created within the National Conference of State Legislatures.

However, different professional organizations conduct different activities; they are not comparable in terms of membership, and the intensity of their involvement varies significantly from one country to another. In reality, the existence of a professional organization within a given country does not guarantee a strong national discourse as discussed above. Many experts highlighted a lack of communication, collective vision and initiative transcending the traditional boundaries of evaluation societies.

Institutionalization of evaluation within governments

The degree of institutionalization within governments varied considerably among the countries studied. Even though the mean of this indicator was high (1.5), results showed contrasted situations between high and low scorers. Japan and South Korea received a '2' while Australia scored 0.7; Ireland and Germany both received a 1, the median score. According to the OECD (1997), a certain level of institutionalization must be reached for evaluation to exercise its full role in public governance. Institutionalization provides the conditions for sustained and systematic data collection on policy implementation and the effects and outcomes of programs. Institutionalization also renders the presence of highly qualified evaluators within the public administration, universities or consultants' firms more likely. Moreover, institutionalization facilitates cooperation among concerned authorities in situations of multi-level governance (for instance, in federal systems) and increases the

acceptability of an evaluation culture. Finally, institutionalization encourages learning processes within policy networks and promotes an efficient implementation of evaluation activities (Leeuw and Rozendal, 1994).

Commitments to evaluation lead to the creation of institutional mechanisms. In this context, the term 'institution' refers to a formal organization or a procedural rule which contributes to the development and the continuity of evaluation in a given jurisdiction. Patterns of institutionalization vary significantly across countries (Jacob, 2005a, 2005b). There is no 'one size fits all' solution in this area. In many countries, the institutionalization of evaluation leads to the enactment and implementation of regulations (e.g. constitutional articles, laws on public administration and performance), evaluation policies and the creation of specific institutional arrangements within the government such as evaluation units.

The use of regulations, which can be more or less permissive or restrictive, is a way to anchor an evaluation reflex in a given bureaucracy and to overcome the inhibitions of certain actors who might fear openness, transparency and accountability procedures. Among the countries studied, France and Switzerland have rooted the principle of evaluation into their constitutions. The first constitutional provision was adopted in Switzerland in 2001. Article 170 of the Swiss constitution stipulates that 'the Federal Assembly shall ensure that the effectiveness of measures taken by the Confederation is evaluated'. The same constitutional principle was adopted in France in the Law of July 23, 2008, which 'introduced the principle that Parliament 'evaluates policies'' (Barbier, 2010: 45, translation). In other countries, evaluation activities are encouraged by the adoption of regulations related to accountability (the Netherlands) or public management reforms (Japan). In the United States, legislation such as the Government Performance and Results Act and management reform initiatives coming from the Office of Management and Budget in the Bush Jr and Obama administrations have encouraged more extensive production and use of evaluations. In Ireland, the government's Value for Money and Policy Review initiative ensures that evaluations are conducted in a wide range of policy sectors. In Japan, the government is obliged to report each year how policy evaluation has been conducted and how the evaluation results have been reflected in policy planning and development. The Korean Integrated Public Service Evaluation System is the result of the implementation of the Basic Law of Government Affairs Evaluation enacted in 2006. According to this legislation, every public program supported by the government or ministries must be regularly evaluated. A few other countries have adopted this type of policy specifically devoted to evaluation. This is the case in Japan with the Public Policies Evaluation Law stipulating that all ministries are to submit an annual policy evaluation plan and conduct evaluation within their jurisdiction. A very similar policy exists in Canada, the new 2009 policy on evaluation of the Government of Canada, which requires all departments to evaluate 'all departmental direct program spending excluding ongoing programs of grants and contributions over five years' (Canadian expert).

Evaluation units exist in many countries and reflect the variety of practices within a government or across policy sectors. For instance, in Finland, the evaluation function is embedded in strategic support units (e.g. development aid, science, innovation, education, social, and health affairs) and in New Zealand, most of the larger government agencies have evaluation capacities, particularly in areas such as health, education, and social services. In Sweden, a number of autonomous agencies are wholly or partially dedicated to evaluation.

These agencies assist ministries in analyzing the effects of public policy in their respective subsectors. In Canada and Israel, all government departments have an evaluation unit and have had so for decades. In many countries, evaluations are frequent but, in spite of policy documents, there is no 'champion' for ensuring that evaluation is addressed systematically across the public service. In some countries, the involvement of the ministry of finance (the Netherlands, Japan) or budget reforms (France, United States, United Kingdom) result in developing evaluation systems committed to planning, budgeting and performance measurement. On the other hand, evaluations are sometimes funded by program managers on an ad hoc basis (Australia).

Some countries have, in addition to evaluation units in various ministries, central evaluation units. In Spain, there is the National Evaluation Agency. Within Japan's Ministry of Internal Affairs and Communication, the central unit has more than 100 employees dedicated to evaluation. The Korean Office of the Prime Minister, the Canadian Center for Excellence in Evaluation of the Treasury Board, and the Norwegian Government Agency for Financial Management in the Ministry of Finance play coordinating roles.

There is a certain logic supporting central units. They can oversee the quality of evaluations conducted or commissioned by other evaluation units; they can also provide technical support to government bodies by developing, testing, and disseminating methods for the ex ante, mid-term and ex post evaluation of public investment projects and programs (e.g. the Italian Public Investment Evaluation Unit − UVAL). As in the European Union, the central evaluation unit − formerly within the DG Budget, now in the Secretariat-General − coordinates the evaluation function, keeps an overview of evaluation findings, and supports evaluation activities in sectorial units.

Outcomes of institutionalization mechanisms at an international level are not easy to gauge. Authors argue that institutionalization contributes to results utilization and quality improvements. According to our panel of international experts, evidence supporting these claims is often missing. The situation in many countries varies dramatically (from one country to another and across policy sectors within a country) and relies on external factors such as evidence-based policy initiatives or management practices. At this time, no institutional mechanism seems to ensure (on its own) the systematic uptake of evaluation results in the policy cycle. In the United Kingdom, many government departments have found that integrating evaluators (usually researchers, economists, and statisticians) into policy and strategy units has been a powerful way of ensuring that evaluation findings are incorporated into ongoing policy development from initial inception through to implementation.

Institutionalization of evaluation within parliaments

Parliaments have the weakest institutionalization of evaluation across all countries. The mean of this indicator is 1 and seven countries received a score below this median score. The lowest score (0.3) was attributed to Ireland and Japan. Only one country (Switzerland) received the maximum score in this category. While parliaments are not themselves among the big producers of evaluations, a higher general interest in evaluation by parliaments was reported by the international experts. This means that parliaments could be very active users of evaluation while lacking their own 'institutional arrangements for conducting evaluations and dissemi-

nating them to decision makers' (Korean expert). Evaluation could play a double role in parliamentarian activities.

First of all, parliaments are sometimes evaluation producers. Some parliaments devoted evaluation to specific internal units. However, none of them were newly created; they already existed a decade ago. The Swiss parliament has an evaluation unit (Parliamentary Control of the Administration), which was inaugurated in 1991 and is the only one directly under the responsibility of the legislature. This might be due to the fact that the Swiss members of parliament are not professional politicians and their need for support might be higher. In 2009, the French National Assembly created a bipartisan (majority and opposition) *Comité d'évaluation et de contrôle des politiques publiques* (public policy evaluation and monitoring committee) to produce ten evaluation reports a year. In Australia, the parliament conducts its own reviews of major and minor policy issues including the use of meta-analysis. The Swedish parliamentary office also conducts evaluations. Additionally, many parliaments have standing organizations for Technology Assessments for new technologies, which will help parliaments in decision-making.

Members of parliament can also commission evaluations by adopting provisions, laws, or constitutional amendments requiring evaluation to be conducted by an autonomous agency or on a specific issue or agenda. An example of the creation of an evaluation agency can be seen in Germany with their parliament's decision to have an independent evaluation agency for development aid. More common is the introduction of evaluation clauses into laws (e.g. in Denmark, Germany, the Netherlands, New Zealand, and Switzerland), which is nothing new as it was already practiced in the United States during the 1960s. In more recent years, this mechanism has been adopted in Germany while significant policy reforms were under way (e.g. within the labour market reform) as well as for experimental laws. In some cases, evaluation clauses form part of political negotiations and help in the adoption of controversial legislation or coalition agreements such as in Germany and Spain. Modern law-making usually includes background information on expected impacts; this has often been routinized and strengthened in recent years in Denmark, Finland, New Zealand, and Germany. However, rigorous ex ante evaluations seem to be the exception; for example, within environmental impact assessments, more often estimations of financial implications are conducted. Another aspect often included in impact assessments is the 'Administrative Burden Reduction Assessments', which has been practiced in the United States since the 1980s and been introduced more widely across European countries in the early 2000s, with the Netherlands being the front runner since the 1990s.

Few parliaments pay attention to evaluation quality and the processes they commission. The Netherlands has a small review center to review and 'verify' evaluations done by the executive branch of government. In Germany, within major or minor interpellations, information on evaluations commissioned by the executive is increasingly sought. They often do not ask for the results, but more for process information, such as when the evaluation started, who was selected as evaluator/evaluation institute and when the evaluation results can be expected. Questions on this level make evaluation a new discursive element in parliamentary discussions rather than one leading to more use of evaluative information.

Second, MPs can use evaluation knowledge produced by others. Normally, parliaments conduct extensive hearings during the budget process. The Korean National Assembly Budget Office (NABO) has an integrated function of audit, research and budget office (see section below about Evaluation within SAIs). In the United States, the Congressional Budget Office's (CBO) mission is also to provide evaluations. In Australia, 'the government's annual reports

to the parliament on program spending and performance (Portfolio Budget Statements – PBSs) are meant to include whatever evaluation findings are available; however, many fail to do this, or only report in a perfunctory manner. Parliament conducts extensive hearings during the budget process, focusing on the PBSs' (Australian expert). In Spain, a Parliamentary Budget Office has recently been created and could provide a permanent arrangement for conducting and diffusing evaluation.

Most parliaments have research and information services which help individual MPs or parties by answering requests and ordering smaller studies. However, they usually do not conduct evaluations but may refer to existing evaluations or write syntheses, taking on more the function of knowledge brokers.

The influence of evaluations within parliament is very difficult to trace as is the use of evaluation in general. Sometimes MPs refer explicitly to evaluations conducted elsewhere. In the Netherlands, the use of evaluation results in parliamentary discussions has risen significantly, especially concerning development aid. MPs often use evaluation reports to give informed answers to parliamentary questions. In Norway, it is common practice to present evaluation results in White Papers, which are then discussed in parliament. The use of evaluation results occurs more frequently within parliamentary committees, where people with more expert knowledge work and where evidence in particular policy areas is sought, partly 'away from the public gaze'. For instance, evaluation findings are often scrutinized by the various parliamentary committees in the United Kingdom. However, for some countries such as Ireland or Spain, discussions based on evaluations' findings rarely take place in Parliament.

To summarize, parliaments have access to evaluation results produced outside, such as from SAIs or commissioned by the executive branch of government. In part, processes are routinized, such as for ex ante impact assessments. However, parliaments can also trigger evaluations by laws and create and change the institutional setting for evaluation, include evaluation clauses into laws and include evaluation results in the budget process. These activities mainly aim at holding the government accountable. The government can - of course, selectively – rely on evaluations produced by the executive to justify its actions, but can also 'tie its own hands' by publicly commissioning an evaluation. In sum, many comments from the experts in the survey were about triggering evaluation in the government and other branches as well as using evaluative information. The use of evaluation in political processes is quite a different matter than its institutionalization within parliament, but can, of course, overlap.

Pluralism of institutions and evaluators

Every country scores 1 or higher on this indicator and five countries obtained the maximum score of 2. This means that a pluralism of institutions and evaluators is commonly observed across the countries being studied. The mean score of this indicator is 1.7. Degrees of pluralism vary generally across the different policy sectors. For instance, it is low in agriculture but high in social affairs. These seem to be general tendencies across nations and policy sectors. In some countries, which do not generally organize evaluation centrally, public agencies exist for chosen sectors; such is the case for education in Denmark. These agencies then limit pluralism on the demand side when they are carrying out internally the majority of evaluations in a given policy sector. It is interesting here how the more centrally-steered evaluation systems are described as more pluralistic. The supply side is often more pluralistic, with many players with different backgrounds (mix of internal and external evaluation). The demand side is partly

centralized, but some countries have evaluation committees consisting of outsiders as well. The very slight increase of the score for all countries reflects the involvement of new actors in evaluation, such as NGOs and other agencies. Within the answers of this survey, a description of decreased pluralism has been the exception, e.g. social policy/social assistance in France.

Evaluation within SAIs

Results regarding the existence of evaluation activities within a country's supreme audit institution (SAI) vary considerably across countries. The mean of this indicator is 1.4. Australia, Finland and Switzerland obtained the maximum score (2) and were followed very closely by Canada, the Netherlands, Norway, the United Kingdom and the United States, each with a score of 1.8. On the other hand, supreme audit institutions in Italy and Spain played a minor role in evaluation and received a score of 0.3. This situation is probably attributable to the historical development of performance auditing. When the *International Atlas* was published, performance auditing was already established in many countries. A few years earlier, Pollitt et al. published a comparative study of performance audits in five countries (1999). The *International Atlas* emphasized that national audit institutions have played an important role in the more general evaluation discussions in the countries which had developed a more mature evaluation culture. In those countries which developed performance audit praxis in the 1970s and 1980s, performance auditing became an important element in the field of evaluation. In the United States, the Netherlands, Canada, and Sweden, audit institutions play an important role as producer of evaluations and an important position in discussions about evaluation (Mayne et al., 1992; Gray et al., 1993). In other countries, such as France, experts are unanimous to say that the *Cour des comptes* is somewhat active in evaluation but is still in transition to precisely define how it will position itself in the new constitutional and legal environment, focusing more on evaluation and performance management (Jacob, 2005c). Since the constitutional reform of 2008, 'the role of the *Cour des comptes* has been modified and more tightly defined: "the *Cour des comptes* helps Parliament monitor the Government's actions. It helps Parliament and the Government monitor the implementation of financial laws and the application of the laws governing the funding of social security, and also in the evaluation of public policies"' (Barbier, 2010: 45, translation)

Proportion of impact and outcome evaluations to output and process evaluations

Given the diverse institutional, organizational, and methodological settings in which evaluations are performed, it is difficult to summarize the methodological landscape within a single indicator. Internationally, the focus on impact evaluations has been emphasized within the last decade and there is certainly a trend towards more impact and outcome evaluations, which were, at first, influenced from the outside. The mean of this indicator is 1.4 and Switzerland obtained the highest mark closely followed by Canada, Finland and the United States (1.8). Low scorers on this indicator were Italy (0.7) and Ireland (0.8).

Discussions on randomized controlled trials (RCT) have spread from the development aid sector and have also tainted other policy sectors (Hansen and Rieper, 2009). An 'RCT lobby' exists in the United States as well as in other countries such as France where there seems to be a certain push from the supply side, mainly from academics specialized in RCTs. This emphasis towards impacts reflects the rhetoric and/or practice of evidence-based policy. In some countries, the use of impact evaluations is institutionalized through initiatives for evidence-based

policy and meta-analyses. The most prominent examples of this are the American What Works Clearinghouse, the Nordic Cochrane Collaboration and the SFI-Campbell Collaboration as well as the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI − Centre) in the United Kingdom. In many countries, such a commitment and investment towards the evidence-based movement is missing.

However, more governments commissioning evaluations have shifted in this direction, most prominently under the Bush Jr and Obama administrations. In the United Kingdom, there seems to be an attempt to measure economic impact for major policy initiatives. Public management initiatives are pushing the focus of evaluations on results and outcomes (e.g. Canada, Denmark, Ireland, Japan, South Korea, United Kingdom, United States), which is especially evident where the use of outcome indicators has been popularized by national laws on evaluation including the obligation of outcome and impact evaluation.

The number of impact evaluations has risen significantly in the areas of criminal justice, education, labour market policy and social policy, whereas it has traditionally been predominant in health policy. This trend is somewhat influenced by funding to academics carrying out impact evaluations, such as in Denmark. In these fields, quasi-experimental designs have certain traditions; however, RCTs are the newer fashion in some countries such as France where they were absent a decade ago. In other countries, a long tradition of process evaluations continues to co-exist with and is not crowded out by impact evaluations. Impact evaluations have been added to existing evaluation practices in Australia, Canada, Germany, Sweden, New Zealand, and Israel. Process and impact evaluations are also being combined elsewhere such as observed in Switzerland by Balthasar (2007) in a study which offers quantitative information on the relationship of the various designs. For 278 evaluations carried out, 19 concerned implementation only, 48 impact only, and 211 both implementation and impact.

Discussion

Evolution and changes over the last decade

When we compare 2001 (see Table 4) and 2011 ratings (see Table 1), we see a slight increase in the average overall score (from 11.2 to 13.7). This increase means that evaluation culture has matured over the last decade. When we take a closer look at the specific indicators, we see that every indicator shows an increase: the variety of policy domains (from 1.6 to 1.7), the supply from different disciplines (from 1.3 to 1.8), the national discourse surrounding evaluation (from 1.4 to 1.5), the existence of professional organizations (from 1.6 to 1.8), institutionalization in the government (from 1.2 to 1.5) and in parliament (from 0.6 to 1), the pluralism of institutions and evaluators (from 1.4 to 1.7), the involvement of the SAI in evaluation activities (from 1.2 to 1.4) and the focus on impact or outcome evaluations (from 1 to 1.4). No indicator has decreased or remained constant over the last decade.

Countries where evaluation culture has improved between 2001 and 2011 include Denmark, Finland, France, Germany (a slight improvement from 13 to 13.3), Ireland, Israel, Italy, Japan, the Netherlands (a slight improvement from 15 to 15.3), New Zealand, Norway, South Korea, Spain, Switzerland, and the United Kingdom. The most noticeable changes took place in Finland (from 10 to 16.6), Japan (from 3 to 12.9), Spain (from 5 to 11.3), and Switzerland (from 8 to 16.4) where ratings spiked over the last decade.

Table 4. Evaluation Culture in 2001.

	I. Domains	II. Disciplines	III. Discourse	IV. Profession	V. Inst. – Government	VI. Inst – Parliament	VII. Pluralism	VIII. SAI	IX. Impact	SUM
Australia	2	2	2	2	1	1	2	2	2	16
Canada	2	2	2	2	2	1	2	2	2	17
Denmark	2	2	2	1	1	0	2	1	1	12
Finland	2	1	1	1	1	1	1	1	1	10
France	2	1	1	2	2	1	1	1	0	11
Germany	2	2	1	2	1	1	2	1	1	13
Ireland	1	1	1	0	1	0	1	1	1	7
Israel	1	1	1	2	1	0	1	1	1	9
Italy	1	1	1	2	0	0	1	1	0	7
Japan	1	0	0	1	1	0	0	0	0	3
Netherlands	2	2	2	1	2	1	2	2	1	15
New Zealand	1	0	1	2	0	0	1	1	1	7
Norway	2	1	1	1	2	1	1	2	1	12
South Korea	1	1	2	2	2	0	2	1	1	12
Spain	1	0	1	2	1	0	0	0	0	5
Sweden	2	2	2	1	2	1	2	2	2	16
Switzerland	1	1	2	2	0	0	2	0	0	8
United Kingdom	2	2	2	2	1	1	2	1	2	15
United States	2	2	2	2	2	2	2	2	2	18
Mean	1,6	1,3	1,4	1,6	1,2	0,6	1,4	1,2	1,0	11,2
Top 3	2,0	2,0	2,0	2,0	1,7	1,3	2,0	2,0	2,0	17,0
Bottom 3	1,0	0,4	0,8	1,4	0,8	0,0	0,6	0,6	0,6	6,2

Table 5. 'Top 3' and 'Bottom 3' groups in 2001.

	Ranking	Country	Score
Top 3	1	United States	18
	2	Canada	17
	3	Australia	16
	…		
Bottom 3	17	New Zealand, Ireland and Italy	7
	18	Spain	5
	19	Japan	3

Table 6. Scores of 'Top 3' and 'Bottom 3' groups in 2001.

	Top 3	Bottom 3	Deviation
I. Domains	2.0	1.0	1.0
II. Disciplines	2.0	0.4	1.6
III. Discourse	2.0	0.8	1.2
IV. Profession	2.0	1.4	0.6
V. Inst. − Government	1.7	0.8	0.9
VI. Inst − Parliament	1.3	0.0	1.3
VII. Pluralism	2.0	0.6	1.4
VIII. SAI	2.0	0.6	1.4
IX. Impact	2.0	0.6	1.4
SUM	17.0	6.2	10.8

We did not observe noticeable changes in the other direction (from a high level of evaluation maturity to a lower one). This relative stability can be explained by the fact that once implemented, evaluation is vigorously rooted in the political and administrative environment. Moreover, several initiatives contributed to the renewal of the performance paradigm brought by the advocates of new public management theories. However, evaluation culture has declined between 2001 and 2011 in some countries such as Australia, Canada, Sweden, and the United States. In these cases, score deterioration is probably a result of the change in the data collection method. In 2001, it was easier to produce a concordant rating between authors and editors of the *International Atlas* than it was with a group of experts having no contact among themselves. However, several indicators showed a decrease that might not be fully attributable to changes in the research design. For instance, the variety of policy domains covered by evaluation has plummeted in Australia (from 2 to 1.3) and a sharp decline in the institutionalization of evaluation in parliament was reported in the United States (from 2 to 1.4).

The distribution of the 19 countries among the three categories of evaluation culture maturity, as rated in 2001, is different from the situation we depict in 2011. In 2001, the majority (53%) of the countries included in our sample presented a high degree of maturity. Seven countries (37%) were in the middle of the scale while only two countries (10%) were qualified as having a low degree of maturity:

- High degree of maturity (equal or higher than 12) ($n = 11$): Australia, Canada, Denmark, Germany, Netherlands, Norway, South Korea, Sweden, United Kingdom, the United States;
- Medium degree of maturity (from 6 to 11,9) ($n = 7$): Finland, France, Ireland, Israel, Italy, New Zealand, and Switzerland;
- Low degree of maturity (below 6) ($n = 2$): Japan and Spain.

Ten years ago, the 'Top 3' was composed of the United States (18), Canada (17), and Australia (16) while the 'Bottom 3' actually grouped five countries due to a tie: Japan (3), Spain (5), New Zealand, Ireland and Italy (7). As mentioned earlier, the differences between countries were clearer in 2001. The overall score of the countries with the highest index of evaluation culture was 17 while it was 6.2 for the lowest one; the difference between these two groups was almost 11 points. The gap is almost twice as wide as the situation we observe in 2011. In 2001, the indicators for professionalization and institutionalization within the government were already among those showing the least contrast between high and low scorers. Conversely, the situation has radically changed for three other indicators (the supply from various disciplines, the pluralism of institutions and evaluators, and the proportion of impact-oriented evaluations). The difference between these three indicators has eroded over the last decade. Most countries have a better supply of evaluators from various disciplines than they did a decade ago. Finland and Switzerland are new to the group of countries with the highest scores for this indicator. Regarding the proportion of evaluation centered on outcomes, more than half of the countries have a higher score today than they did a decade ago. Finally, the institutionalization within the parliament and the involvement of the SAI in evaluation are still the indicators showing the greatest variation among countries belonging to the 'Top 3' and the 'Bottom 3' groups.

Diffusion without convergence

Concerning the diffusion of an innovation (Rogers, 1995), several paths lead to a mature evaluation culture and the institutionalization of evaluation. Institutionalization may occur in particular sectors or at the whole-of-government level. All 19 countries already had intensive evaluation experience a decade ago. Many countries included in this survey already have a long history of evaluation. It means that they can look back on decades of experience in implementing evaluation. To some extent, evaluation has become 'business as usual', which also has its downsides such as evaluation fatigue, which is described as 'hyokazukare' in the Japanese language or as 'evaluitis' as coined by Swiss academic Frey (2007). We can see results from the reforms drafted and initiated in the previous decade. For the Japanese reforms, the turning point can be traced back to 1996; the reforms were rendered even more manifest with the establishment of the central evaluation unit in 2001. In Switzerland, evaluation institutionalization had already started back in the 1990s. We can report on more recent reforms, such as those in Spain from 2004 onwards and those in France in 2008; however, it is still too early to take stock of their effects.

Although evaluation is spreading throughout OECD countries, cross-national institutional differences are not disappearing. A central difference across countries is decentralized versus centralized institutionalization structures. Within the decentralized mode, evaluation entities can be either part of ministries' respective implementing agencies or be separately created as

sectorial evaluation agencies. Coordination between the sectors is fostered in some cases, such as in Canada with the Center of Excellence for Evaluation within the Treasury Board of Canada Secretariat. Within the centralized model, entities are responsible for evaluation over all sectors, like the Spanish evaluation agency, or the *Comité d'évaluation et de contrôle des politiques publiques* in the French parliament, or the Bureau of Evaluation of the Ministry of General Affairs in Japan, or the Korean Division of Evaluation of Programs established under the Budget Office of the National Assembly. In sum, in many countries, new evaluation institutions have been created, which seems to be an overall tendency.

The aforementioned developments are not only about the rise of evaluation but also about the rise and decline of evaluation. France has gone through various phases; after 2001, the inter-ministerial *Conseil National d'Évaluation* was discontinued and evaluation practice was relaunched in 2007. Also, in the Netherlands, evaluation has been more institutionalized over the last decade; however, budget cuts which became effective in 2009 (Leeuw, 2009) may affect the future of this country's evaluation culture.

Limits and future research

Qualitative indicators applied across many countries and based on subjective expert views have their limitations. This survey may be considered biased because of the selection of experts, who can be said to belong to a sort of subculture within evaluation. First, they are all, more or less, involved in governmental evaluation and are often a sort of entrepreneur for evaluation. Second, they belong to certain organizations in which a small minority of evaluators is involved. Additionally, the scale within the survey does not allow for much variation, was chosen for comparability with the former results, and may be changed for future surveys since it does not capture enough variation in a context of convergence and political environment surrounded by a performance paradigm. Ratings are always normative in so far as they make judgments on one country being better than another. However, we did show that many forms of relatively mature evaluation culture exist in our sample.

We can observe very different trajectories of evaluation capacity building across OECD countries. The development is very much embedded in the political culture and determined by other existing institutions. The triggers for evaluation capacity building vary considerably. Institutional designs and organizational attributes have been analyzed here, but no conclusions can be drawn regarding any kind of superiority of these models. Some strong institutionalizations may be weakened within the political process and, conversely, weak institutionalizations may nevertheless contribute to important evaluations and their use. Additionally, it is not only formally institutionalized entities which play important roles in supporting evaluation. However, when describing formal institutional differences, the knowledge about the institutional quality is also lacking; for example, many rules and laws leave room for interpretation. Some evaluation institutions may be imported from outside the country, but the national dynamics and administrative culture persist. For this reason, it is important to reflect on the particular governance settings for evaluation in different countries; there will not be a 'best' evaluation institutionalization for all countries.

The indicators developed in the *International Atlas of Evaluation* offer a broad overview of evaluation capacity building on the country level, including aspects of the supply side as well as the demand side. However, the difficulty of generating results across several countries and many sectors, without generalizing too much, remains. More research on sectorial differences will be

needed and further disaggregated measures could be developed for obtaining a more finely-grained picture. Moreover, some of the indicators (e.g. national discourse, impact/outcome) capture the perceptions of the experts rather than the formal aspects of the institutional setting. In some cases, the results still do not tell us enough about specific rules, mechanisms, and systems. This research was focused on the national level and could not account for variations in the regional and local level. Several experts mention that evaluation is also institutionalized at regional, municipal, and local levels (e.g. Denmark, France, Switzerland, and the Netherlands). We have not investigated the diffusion and transfer of evaluation practices within a country even if it might be interesting to know more about the ripple effect of institutionalization across layers of government. The role of civil society, non-governmental organizations (NGOs), and civil society organizations (CSOs) has not been captured by this expert survey but should be analyzed in the future. It was also out of the scope of this article to analyze the embeddedness of evaluation culture in the different political and administrative systems, such as governance structures, governance processes, welfare systems, political systems, and traditions in political discourses. Comparing the results with broader developments within the countries of this survey would be necessary.

Further research will be needed to explain international institutional differences in evaluation.

Future research might enlarge the scope and include more countries in which evaluation has emerged and has been consolidated in recent years (e.g. Eastern Europe, developing countries in Asia and Africa). Moreover, measuring evaluation culture and institutionalization across international organizations and supranational entities is an avenue to consider for future research. For instance, we are aware that the European Union has played a role in the development of evaluation in Europe but, in this study, we have not collected specific measures about evaluation culture within the European institutions for two reasons. First, we would have to face the problem of comparability between state-based and supra-national structures. Second, we would give undue emphasis to the European Union compared to other regional groupings emerging around the world. Future research could focus on supra-national or international organizations to compare and contrast our results on a state-national level.

Future research could also go beyond description and contribute to theoretical developments. By taking stock of the growing literature in the field it will be possible to elaborate theoretical models about evaluation culture and institutionalization processes. This research avenue will be paved with existing theories about innovation diffusion and knowledge transfer. To do so, reflections and operationalization of concepts such as ‘evaluation culture’, ‘institutionalization process’ and ‘evaluation capacity building’ will be a fundamental initial step.

Acknowledgement

The authors wish to thank all the respondents for their generosity and interest in this study. The first author gratefully acknowledges a research grant from the Canadian Social Sciences and Humanities Research Council (SSHRC) to carry out this study.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for profit sectors.

Notes

1. The *International Atlas* was published in 2002. It depicts the situation in 2001 (Furubo et al., 2002: 11).

2. 'Seam effect' refers to distortion in measurement results using longitudinal or panel data as a result of uneven rates of change in the intervening years between two measurement moments.

References

American Evaluation Association (AEA) (2008) *American Evaluation Association Internal Scan Report to the Membership*, by Goodman Research Group. URL (consulted 28 November 2011): http://www.eval.org

Alkin MC (ed.) (2004) *Evaluation Roots: Tracing Theorists' Views and Influences*. London and New Delhi: SAGE.

Balthasar A (2007) *Institutionelle Verankerung und Verwendung von Evaluationen*. Zürich/Chur: Rüegger Verlag.

Barbier J-C (2010) Éléments pour une sociologie de l'évaluation des politiques publiques en France [Elements for a sociology of public policy evaluation in France]. *Revue française des affaires sociales* 1−2: 25–49.

Barbier J-C and Hawkins P (eds) (2012) *Evaluation Cultures Sense-making in Complex Times*. New Brunswick, NJ and London: Transaction Publishers.

Breen G and Associates (2005) *Interviews with Deputy Ministers Regarding the Evaluation Function*. Ottawa: Centre of Excellence in Evaluation, Treasury Board of Canada, Secretariat.

Budge I (2000) Expert judgments of party policy positions: Uses and limitations in political research. *European Journal of Political Research* 37(1): 103–13.

Bussmann W (2008) The emergence of evaluation in Switzerland. *Evaluation* 14(4): 499–506.

Cousins BJ, Cullen J, Malik S and Maicher B (2009) Debating professional designations for evaluators: Reflections on the Canadian process. *Journal of MultiDisciplinary Evaluation* 6(11): 71–82.

Drummond M and McGuire A (eds) (2001) *Economic Evaluation in Health Care: Merging Theory with Practise*. New York: Oxford University Press.

Feinstein O and Zapico-Goñi E (2010) Evaluation of government performance and public policies in Spain. *ECD Working Paper Series, 22*. Washington, DC: IEG World Bank.

Frey BS (2007) Evaluierungen, evaluierungen … evaluitis. *Perspektiven der Wirtschaftspolitik* 8(3): 207–20.

Furubo J-E, Rist RC and Sandahl R (eds) (2002) *International Atlas of Evaluation*. New Brunswick, NJ and London: Transaction Publishers.

Gaarder MM and Briceño B (2010) Institutionalisation of government evaluation: Balancing trade-offs, Working Paper 3ie, New Delhi.

Gray A, Jenkins B and Segsworth B (eds) (1993) *Budgeting Auditing and Evaluation: Functions and Integration in Seven Governments*. New Brunswick, NJ and London: Transaction Publishers.

Grinnell R, Gabor P and Unrau Y (2011) *Program Evaluation for Social Workers: Foundations of Evidence-Based Programs*. New York: Oxford University Press.

Hallsworth M and Rutter J (2011) *Making Policy Better: Improving Whitehall's Core Business*. London: Institute for Government.

Hansen HF and Rieper O (2009) The evidence movement. The development and consequences of methodologies in review practices. *Evaluation* 15(2): 141–63.

Huber J and Inglehart R (1995) Expert interpretations of party space and party locations in 42 societies. *Party Politics* 1(1): 73–111.

Jacob S (2005a) *Institutionnaliser l'évaluation des politiques publiques. Étude comparée des dispositifs en Belgique, en France, en Suisse et aux Pays-Bas* [*Institutionalizing Public Policy Evaluation. A Comparative Study of the Mechanisms in Belgium, Switzerland and the Netherlands*]. Bruxelles: PIE-Peter Lang.

Jacob S (2005b) Réflexions autour d'une typologie des dispositifs institutionnels d'évaluation [Reflections on a typology of the institutional mechanisms of evaluation]. *Revue canadienne d'évaluation de programme* 20(2): 49–68.

Jacob S (2005c) La volonté des acteurs et le poids des structures dans l'institutionnalisation de l'évaluation des politiques publiques (France, Belgique, Suisse et Pays-Bas) [The will of actors and the weight of structures in the institutionalization of public policy evaluation (France, Belgium, Switzerland, and the Netherlands)]. *Revue française de science politique* 55(5–6): 835–64.

Jacob S (2008) Cross-disciplinarization: A new talisman for evaluation? *American Journal of Evaluation* 29(2) : 175–94.

Jacob S and Slaïbi R (2014) L'évaluation au sein du gouvernement fédéral, un outil de surveillance ? [Evaluation within the federal government: A tool for surveillance?] In : Crête J (ed.) *Les surveillants de l'État démocratique*. Québec, QC: PUL, 135–52.

Jacob S and Varone F (2004) Cheminement institutionnel de l'évaluation des politiques publiques en France, en Suisse et aux Pays-Bas (1970-2003) [Institutional development of public policy evaluation in France, Switzerland and the Netherlands]. *Politiques et Management Public* 22(2): 135–52.

Leeuw FL (2009) Evaluation policy in the Netherlands. *New Directions for Evaluation* 123: 87–102.

Leeuw FL and Rozendal P (1994) Policy evaluation and the Netherland's government: Scope, utilization and organizational learning. In: Leeuw FL, Rist R and Sonnichsen R (eds) *Can Governments Learn? Comparative Perspectives on Evaluation and Organizational Learning*. New Brunswick, NJ and London: Transaction Publishers, 67–89.

Mark M, Donaldson S and Campbell B (2011) *Social Psychology and Evaluation*. New York: Guilford Press.

Mayne J, Bemelmans-Videc M-L, Hudson J and Conner R (1992) *Advancing Public Policy Evaluation: Learning from International Experiences*. Amsterdam: North-Holland.

OECD (1997) *Promouvoir l'utilisation de l'évaluation de programme*. [*Promoting the Use of Program Evaluation*]. Paris: OECD.

Oxman AD, Bjørndal A, Becerra-Posada F, et al. (2010) A framework for mandatory impact evaluation to ensure well informed public policy decisions. *The Lancet* 375(9712): 427–31.

Pawson R and Tilley N (1997) *Realistic Evaluation*. London and New Delhi: SAGE.

Pollitt C, Girre X, Lonsdale J, et al. (1999) *Performance or Compliance? Performance Audit and Public Management in Five Countries*. Oxford: Oxford University Press.

Rogers E (1995) *Diffusion of Innovations*, 4th edn. New York: Free Press.

Rogers P and Davidson J (2013) Australian and New Zealand evaluation theorists. In: Alkin M (ed.) *Evaluation Roots: A Wider Perspective of Theorists' Views and Influences*. Thousand Oaks, CA: SAGE, 371–85.

Stame N (2013) A European evaluation theory tree. In: Alkin M (ed.) *Evaluation Roots: A Wider Perspective of Theorists' Views and Influences*. Thousand Oaks, CA: SAGE, 355–70.

Smits P and Jacob S (2014) La fonction d'évaluation dans l'administration publique québécoise: analyse de la cohérence du système d'actions [Evaluation in Québec's public administration: An analysis of the coherency of the action system]. *Administration publique du Canada* 57(1): 71–96.

Vaessen J and Leeuw FL (2010) (eds) *Mind the Gap. Perspectives on Policy Evaluation and the Social Sciences*. New Brunswick, NJ and London: Transaction Publishers.

Varone F and Jacob S (2004) Institutionnalisation de l'évaluation et Nouvelle Gestion publique: un état des lieux comparatif [The institutionalization of evaluation and new public management: A comparative inventory]. *Revue internationale de politique comparée* 11(2): 271–92.

Widmer T, Beywl W and Fabian C (eds) (2009) *Evaluation. Ein systematisches Handbuch*. VS Verlag für Sozialwissenschaften: Wiesbaden.

Steve Jacob is a full professor in the Department of Political Science at Laval University. He is the founder and director of the research laboratory on public policy performance and evaluation (Per*f*Eval). He conducts research dealing with the mechanisms of performance management and evaluation:

professionalization, institutionalization, and capacity building in Europe and Canada, ethics in evaluation, and participatory approaches.

Sandra Speer is an independent evaluator, based in Wiesbaden, Germany; the focus of her work is on national and international evaluation research for public agencies as well as evaluations in various fields.

Jan-Eric Furubo has held many different positions within the National Audit Office in Sweden. Furubo has published widely on evaluation methodology, the role of evaluation in democratic decision-making processes and its relation to budgeting and auditing.