

Summary

Exploring evaluation capacity in Dutch central policy departments

In the strive for regulatory quality, ex post evaluation research has an important role to play in the Netherlands. To actually fulfil this role, requires that (1) evaluations of laws and regulations (further: evaluations of laws) are planned and carried out; (2) these evaluations are designed in such a way as to generate useful, valid and reliable insights and (3) they are being used by policy officials and designers of laws to inform new or revised laws and policies. These three requirements are not automatically reached in practice. Laws and legal amendments are often complex units of analysis: a mix of norms, rules and manifestations of policy interventions. Apart from social-scientific skills, evaluations of laws also require a legal perspective. What is more, individual policy officials and jurists (designers of laws) indicated that know-how is lacking as to the commissioning and management of these evaluations. These factors may have caused evaluations of laws to vary quite strongly in terms of process, scope and methodological design. In the Netherlands there is no central regulatory evaluation policy. This should not have to be problematic at all. However, significant variation in terms of validity and reliability of evaluations of laws was found earlier.

This widely varying practice of evaluations of legislation has bearing on issues of *evaluation capacity*: the capacity to conduct evaluations, and, to use them in the regulatory cycle. Which conditions are present in policy departments to facilitate the production and utilization of evaluations of laws, in view of the evaluation process as well as content?

The aim of this study is to *gain insight into existing departmental practices as regards the conduct and use of ex post evaluations – particularly, evaluations of laws*. The study was commissioned by the Knowledge Centre for legislation and legal affairs of the Ministry of Security and Justice (Legislative quality division). The Knowledge Centre strives for more comparability between evaluations of laws, in terms of both process and content, to be able to better accumulate knowledge and draw lessons from experiences with these evaluations.

Research questions and approach

- 1 Which general insights about core aspects of evaluation capacity are offered by the international (a) evaluation literature and (b) norms and guides of international institutions and larger countries?
- 2 Which experiences as regards evaluation capacity have been accumulated by policy departments, by policy officials and jurists?
- 3 What research designs are deemed relevant for methodological quality and content of evaluations of laws?
- 4 Which lessons can be drawn from the answers to the above questions for the (further) development of evaluation capacity – particularly in relation to legislation?

This study was structured along with the following aspects of evaluation capacity and their characteristics, based on evaluation literature (research question 1). These constituted the conceptual backbone of the report:

- evaluation structure: goals and organizational (evaluation) parts or units;

- evaluation programming and obligations;
- financial and other means for evaluation;
- evaluation frameworks & programs (process and content);
- safeguarding evaluation knowledge and experience;
- the promotion of evaluation utilization.

We used this conceptual frame to, first, report a non-exhaustive number of approaches and experiences by leading international organisations and larger (western) countries. Second, desk study combined with 35 interviews with key informants were conducted to collect information from departmental practice as regards the six aspects.

Evaluations, including those of laws, do not take place in a vacuum. They are conducted from one or more of a variety of possible motives, e.g.: political accountability, for example, addressing a parliamentary question; financial accountability, to complete the budget cycle; to learn, i.e. to close the policy or regulatory cycle and use evaluation knowledge to feedback into new or adjusted policies. Evaluations may also serve a strategic goal. Two ideal-typical evaluation motives are discerned in the literature: (a) accountability and (b) learning, as in enlightenment and/or policy embetterment. The political context influences which of these motives is dominant. In turn, the evaluation motive affects how evaluations are dealt with and used. What is more, there is a correlation between motive and evaluation type. Summative evaluations, focused on the question 'did it work', more often suit the accountability motive, whereas formative evaluations, focused on understanding (intermediate) results from policies or laws, particularly fit the learning motive.

International experiences with evaluation capacity (research question 1)

A variety of international handbooks, guides and related materials offered by (among others) the UN, European Commission (EC), US, UK and World Bank demonstrate that policy officials – often the commissioners – have been assigned a significantly larger role in the commissioning and management of evaluations over the past decennium.

First, commissioners make crucial decisions about the evaluation process. They co-determine evaluation content and quality – as well as use. For example, they have a say or determine the overall evaluation plan / program and budget as well as the scope, focus and main evaluation questions. A multitude of criteria and guides circulate on formulating Terms of Reference (ToRs), the process of commissioning, supervision, judgement and utilization of evaluations (see Annex 3 to this report). These guides often also include information on suitable evaluation designs and methodologies, particularly for evaluations of effectiveness. Here, two strands can be discerned: *black-box* and *clear-box* evaluation approaches. Black-box approaches address the attribution problem: to what extent did the policy or law under evaluation actually trigger the results aimed for – rather than factors external to it? *How* this happened is outside the scope of study. In this strand an emphasis is put on experimental designs – including randomized controlled trials (RCTs) or quasi-experimental ones, where confounding factors are filtered out as much as possible. Clear-box approaches address the question: how did a policy or law work out, and why or under which circumstances? The two strands are not mutually exclusive, but can be complementary. In this regard, fit-for-purpose is an emerging mantra; 'the right' evaluation design hardly exists. Whether a design is adequate depends on, for example, type of policy, policy stage, topic, main questions to be answered and target group type(s). For example, the US General Audit Office (GAO) published a

guide for evaluating highly complex programs: how to make the most out of an effectiveness evaluation when a program can hardly be disentangled from its context?

By making policy officials more aware of the relative strengths and limitations of different evaluation designs, they are enabled to make better informed decisions in relation to evaluation – so it is thought.

Second, apart from key players in the evaluation process, policy officials also help create evaluation preconditions by shaping their policies in the process of policy development. It is they, not evaluators who determine the extent to which policies are clear and based on an articulated policy theory. Explicit assumptions about the intended chain of effects helps to focus an evaluation study at a later stage, adding to evaluation quality. A sharper focus could also enhance the cost-efficiency of the evaluation process, according to an EC-guide.

Apart from the policy theory, policy officials can help forward evaluation quality at an early stage by setting up a monitoring system from the inception of a new or revised policy or law. It is viewed as a missed opportunity if baseline and follow-up measurements of process, intermediate and end result indicators are not made from the start. Target groups like communities, citizens and companies can contribute to these monitoring processes.

Third and related, evaluation *use* is increasingly viewed as part of the policy-making process, rather than as the trickling down of an end report. The EC 'Smart Regulation' program prescribes a feedback loop from ex post evaluation findings to the first stage of policy formation or adjustment, through the obligatory ex ante Regulatory Impact Assessments (RIA). In these broad assessments ex post evaluation findings need to be included. At the UN Evaluation Group (UNEG), use is also prescribed and also meta research and knowledge repositories are promoted to keep the oversight of evaluation knowledge. At the country level, however, it has been hard to stimulate the systematic use of evaluations, according to a recent international comparative study. Nevertheless, evaluation societies and parliaments do (or can) fulfil a stimulating role in drawing lessons from evaluation.

Evaluators remain a key factor in evaluation quality as well as their use, judging, for example, from guides to write and present findings in a transparent manner. Their importance is also demonstrated by overviews of evaluator competencies, to be used by, among others, commissioners in the selection process. Not only methodological skills are judged important, also interpersonal and communication skills, flexibility and an antenna for the political and policy context.

Although internationally, much has been published regarding the promotion of evaluation capacity, the overall effectiveness of such initiatives is unknown as of yet. For example, adverse effects, for which some have warned, could be at work, like opportunistic behavior or evaluation ritualization.

Experiences by Dutch policy departments (research questions 2 and 3)

Policy departments very differently shaped evaluation capacity, as regards the involved organizational structure(s) as well as evaluation processes and methodological frameworks/content. Accountability in terms of budget spending turned out to be an important driving force behind evaluating. The key instrument of the central evaluation system is 'policy appraisals', syntheses of existing evaluation work covering specific parts ('articles') of the policy budget. Policy appraisals ought to take place every four to seven years at most and are to be sent to Parliament. The policy divisions within departments are responsible for setting up 'their' own policy

appraisals. This may have an effect on day-to-day evaluation practice, for example on the programming of studies.

Evaluation structure

In the Netherlands, the vast majority of evaluation studies are commissioned to external parties like bureaus, research institutes or university departments, in view of, according to our respondents, internal capacity restrictions and objectivity. Within policy departments, a variety of units assist/are involved in managing evaluations. The initiative to conduct evaluations is most often taken by the research coordinator within a policy division or directorate. None of the policy departments we studied has a separate evaluation division or unit, but there are a number of semi-independent research institutes performing also evaluations, among which the Research and documentation centre (WODC) at the Ministry of Security and Justice. All policy departments have a financial-economic division, responsible for accountability and control. To varying extents, these financial-economic divisions help stimulate the planning, quality and use of policy evaluation research. For example, at the Ministry of Social affairs, this role is quite pronounced as opposed to other departments. Within the Economic Affairs department, since 2012-2013 a network of divisions and units, including the financial-economic division, has addressed the quality of monitoring and evaluation (m&e) activities, mostly in relation to *effectiveness* evaluations.

As opposed to financial-economic divisions, 'knowledge divisions' are more concerned with strategic, future-oriented research than with ex post evaluations. The latter account for just a fraction of their research activities.

Several of the organizational structures and units are quite fresh and require time to take shape in everyday practice. An exception is the Commission for Regulatory Evaluation (CER) at ZonMw, which has been operating in the domain of evaluations of health law since almost two decades.

Evaluation programming and obligations

In all annual departmental budgets, an evaluation program is included and published. Besides this, almost no parallel programs were retrieved. These evaluation programs partly also result from obligations in about 10-20 per cent of concrete laws, to (periodically) evaluate the law in question, most often in terms of effectiveness and efficiency. As indicated, these evaluation studies form the base material for the periodic policy appraisals, in which findings are to be synthesized to account for budget spending. The process of evaluation programming seems to be dominated by accountability motives and political incentives. Most departments first determine an overall research budget, within which individual evaluations ought to fit. Research institutions like the WODC or the Netherlands Institute for Social research (SCP) administer their own budget.

As concerns non-financial means, respondents often reported that policy divisions spend a limited percentage of their time on research and evaluation, judging from (among other things) limited capacity for the co-ordination of research activities. This is not without consequence: it puts pressure on the programming of evaluations. What is more, time pressure makes policy officials to postpone the start of an evaluation, so that measurement opportunities are missed, like forming a control group, performing a baseline measurement and/or setting up a monitoring system.

Over the past few years, financial-economic divisions and sometimes other types of organizational divisions have tried to make evaluation programs more pro-active, particularly in view of the obligatory periodic policy appraisals and the strive for

establishing policy effectiveness. It is too early to establish the results of these initiatives.

Evaluation frameworks & programs (process and content)

In chapter 3, we described a number of programs that serve to manage or enhance both the process and content of evaluation studies. Three of these programs particularly addressed legislation – in the domains of the environment, health and administrative law, respectively. The evaluation programs are examples of (temporary) evaluation capacity and offer lessons as regards the programming and demarcation of evaluation questions, design and evaluation criteria. One overarching lesson resulting from these experiences is the importance of a multidisciplinary, empirical-legal approach when evaluating legislation. Effectiveness in a legal sense does not imply that a law will achieve the intended societal effects in practice – and vice versa. In the Program for Regulatory Evaluation (PER, managed by the aforementioned CER) in the health domain, evaluations of laws need to fulfil seven main requirements.

Besides these programs, we described a number of general programs applicable to policy making in general (potentially including laws). At the department of Education, the knowledge division stimulates combinations of black-box (experimental) and clear-box (explanatory) evaluation approaches. At the department of Economic affairs, the financial-economic division and the new evaluation network developed a guide for enhancing *policy* quality, of which an integrated part is how to perform evaluations. So here, evaluation is regarded as one of a set of indispensable means to achieve the overarching aim of (more) effective policy making. The guide encompasses all stages of the policy cycle, from problem analysis, choice of instruments and ex ante appraisal to political decision-making, implementation and finally, ex post evaluation. It also contains tips and tricks for doing 'early' (small scale) policy experiments and for setting up a monitoring system.

The dominant approach in evaluation research differs between the policy departments studied, depending in part on the dominant departmental divisions or units involved. For example, if the financial-economic division is very active, summative approaches often prevail over explanatory ones. For example, at the Economic Affairs department, commissioners can make use of guides and checklists for the commissioning and content of (experimental) *effectiveness* evaluations, including checklists for assessing tenders, TORs and final reports. If knowledge divisions or scientific research institutes are involved in evaluation, more often a mix of evaluation approaches is encountered (e.g. at the Education department).

In a recent ex post evaluation program at the Ministry of Infrastructure and the Environment, the ambition is to hark back to earlier findings from ex ante regulatory impact assessments of the policy or law.

Safeguarding evaluation knowledge and experience

Like internationally, the commissioning party is increasingly regarded as a key player in the process and content of evaluation research by the Dutch policy departments. To what extent and how do policy departments make sure that commissioners have the know-how and experience to adequately manage evaluations? We addressed: which parts of the departmental organization fulfill a role in this regard, the presence of evaluation training courses, how external evaluators are selected and then supervised, and how evaluation experience and knowledge is retained and shared (collective memory).

In most policy departments, research coordinators – often senior policy officials – fulfil a key function in this regard; they are relatively experienced and help forward less experienced colleagues. Besides this, support functions vary significantly be-

tween departments. For example, the Ministry of Social Affairs has a *Chief Science Officer* who supports the co-ordination and methodology of research including evaluations. At the Education Ministry, the Knowledge division fulfils a similar role as well as a Policy Analyst Team at the Ministry of Economic Affairs that forms part of the aforementioned evaluation *network*. At the Ministry of Security and Justice, project managers of (evaluation) research more or less 'take over' the supervision of evaluations after the TOR-stage. Nevertheless, the commissioner from the Ministry (most often a policy official) is member of a supervisory committee, which is always installed.

A limited number of departments and 'government academies' offer special courses for policy officials, fully or partly about evaluation.

The selection of evaluation research teams is most often determined in the competitive tender procedure, in which each ministry uses its own (more or less explicit) set of criteria. Some ministries have a more elaborate selection procedure than others. For example, the Education Ministry that uses long-term research contracts. At the Ministry of Security and Justice, the WODC keeps a database containing characteristics and capacities of external research teams. Particularly in case of evaluations of laws, a multidisciplinary team is crucial.

Once the evaluation process has started, supervisory committees are regularly used to monitor progress, quality and content. In some policy departments, this is standard practice and there are procedures as regards the role and composition of these committees. For example, the WODC requires the chair of the supervisory committee to be a senior academic.

The exchange of departmental experiences with evaluation projects is often done by informal means, for example, in bilateral or division meetings. Respondents doubt whether this will safeguard 'collective evaluation memory'. Between departments, periodic exchange of evaluation experiences takes place between at least the financial-economic divisions and the Ministry of Finance (often in the light of policy appraisals). This was valued positively by the respondents involved.

The aforementioned 'fresh' evaluation structures and programs appear to be vulnerable when taking into account: departmental cultures as regards research and evaluation, reorganizations and large turnover of policy staff. This is in line with foreign experiences that enhancing and consolidating evaluation capacity is a long-winded task that requires perseverance.

Promoting evaluation utilization

Evaluation use is an indispensable aspect of evaluation capacity. Without processing the fruits of evaluation in the development of laws and policies, each of the capacity aspects described above will lose meaning. Contact and consultation between commissioner and evaluator enhances the probability of evaluation use according to the literature as well as this study. Timeliness and timing in the policy cycle are also important factors. In general, departments publish evaluation reports and often the reports are sent to Parliament. In many cases, an official cabinet response is attached. This response often includes statements as to how a Minister intends to deal with the findings and (if available) recommendations. Besides this, evaluators themselves often valorize evaluation findings by means of presentations for policy officials or academics and articles in professional or academic journals.

In the end, it is up to policy officials and the political context to what extent and how findings are used. There are no fixed or prescribed routes for this, apart from horizon clauses in financial regulations. These regulations expire after five years and can only be prolonged if sufficiently sustained by, among other things, ex post evaluation findings.

To increase their learning value, users argue that ex post evaluations should not only look back, but be oriented towards future action to a larger degree. Evaluations are also deemed instructive if policy assumptions (policy theory) are given a reality check. Finally, more knowledge repositories could be established in which findings from previous evaluation studies are systematized.

Conclusions and discussion (research question 4)

Many of the aforementioned evaluation structures and programs are still quite young, making it hard to determine their impact on learning and accountability and ultimately, on the effectiveness of laws and policies. Hence, it would be interesting to closely follow future developments in the context of evaluation capacity.

One key condition for evaluation capacity is that initiatives to enhance policy relevant knowledge and learning experiences are welcomed rather than punished, in the political context as well as in the central evaluation system. In relation to this, some respondents argued that to learn from evaluations requires formative (process and intermediate) evaluations next to summative work about end effects.

What do the above findings imply for law making divisions and the commissioner of this study, the Knowledge Centre for legislation and legal affairs (Legislative quality division)? Quite rarely are law making divisions involved in the process of ex post evaluations – as opposed to ex ante activity. Only in a limited number of cases, an ex post evaluation of legislation was coordinated by a legal division – for example, when several policy divisions were involved (e.g. regarding the railway system).

Overall, policy divisions rather than jurists are responsible for ex post evaluations. Nevertheless, involving jurists could enhance the content and quality of evaluations of laws. After all, combined legal-empirical approaches proved to add value. What is more, in the very process of law-making, jurists could design evaluation clauses (obligations to evaluate within X years) in such a way that the resulting evaluations become more meaningful. For example, by choosing realistic evaluation periods to allow effects sufficient time to materialize. Finally, legal divisions could be made more aware of the relevance of the *explanatory memorandum* in the clarification of the main policy aims and assumptions as regards effectiveness. A clear policy theory will make it more easy to focus evaluations of legislation at a later stage. The Knowledge Centre for legislation and legal affairs has developed a guide for writing explanatory memoranda.