Utrecht Law Review
igitur
Utrecht Publishing & Archiving Services

This article is published in a peer-reviewed section of the Utrecht Law Review

Can legal research benefit from evaluation studies?

Frans L. Leeuw*

1. Introduction

The evaluation of (public) programmes, laws and other interventions has been characterized by Scriven[1] as a *transdiscipline*. He used this term to indicate that theories, research designs and methods from disciplines like psychology, economics, sociology, public administration, and statistics and, to some extent, law, have merged into a new field, called evaluation (studies). Evaluators are dissecting the backgrounds, processes and effects (a.k.a. 'impacts') of 'tools' or 'interventions' of governments and other (public) organizations. These 'tools' refer to legislation, regulation, covenants, and contracts, but also to subsidies and grants, contracts and (penal) sanctions like 'naming and shaming', and fines. Dissecting not only includes the description and explanation of processes to implement interventions, but also focuses on analyzing the assumptions ('rationale', 'theory') underlying their application and the effects and side-effects that they have on persons, organizations and society.

In order to do so, evaluators need to be capable of applying different theories, methodologies and (research) designs. Moreover, as evaluation research usually is not *l'art pour l'art*, the usability and the use that is made of results of evaluations by policy makers and other officials is also on their agenda.

To a large extent the (academic) field of the study of law is *also* a transdiscipline, taking into account insights from philosophy, logic, language studies, history, behavioural sciences and economics, to name some important ones. And as with evaluations, legal studies are often directed at the parties involved, like governments, victims or contracting parties.

Legal interventions try to influence the behaviour of persons and organizations or the conditions under which they act. Smits[2] puts it as follows. 'The core of legal science is the

* WODC (Research and Documentation Centre), Ministry of Security and Justice, The Hague & Maastricht University (the Netherlands). This paper is based on a speech which the author delivered during the seminar organized by Utrecht University's Faculty of Law, Economics and Governance on legal methodology, April 21, 2010. Corresponding address: f.leeuw@minjus.nl or frans.leeuw@maastrichtuniversity.nl

1 M. Scriven, 'The concept of a transdiscipline; and Evaluation as a transdiscipline', 2008 *Journal of MultiDisciplinary Evaluation*, no. 10, pp. 65-66.

2 J. Smits, *Omstreden rechtswetenschap*, 2009, p. 49.

URN:NBN:NL:UI:10-1-101160

behaviour of the *homo juridicus* (…) legal scholars look at a question that precedes [this] behaviour (i.e. what people are doing, FL): *what is it that people should do as a matter of law?*'

However, designing and implementing interventions like laws, sanctions, contracts and treaties do *not guarantee* that they have the desired or expected impact (on behaviour). It is an empirical question to what extent they do so. To measure their impact, one has to take into account that the interventions operate at different levels and often involve multiple parties; some regard individuals as end-users (e.g. taxpayers, students or farmers), others operate at a more macro level (state officials, international institutions etc.). Some are directed at only one or a few actors, while others focus on chains of activities and organizations. To dissect how the implementation processes operate and to what extent the expected impact has been realized is food for thought for evaluators. Therefore the following question can be asked: to what extent can legal studies benefit from evaluations?

2. Why this question?[3]

Although empirical legal research is winning ground both in the Netherlands and abroad, as is made clear by the existence of specialized journals[4] and organizations,[5] the discussion as to its relevance and appropriateness for the study of law continues. Limiting ourselves to the Netherlands, Smits[6] has asked the question whether legal research is a 'discipline in crisis' and whether it is suffering from an 'identity crisis'. The question is linked to debates about the roles and relevance of methodology and empirical research. Tijssen[7] presents empirical evidence of how Dutch jurists use methodology in their PhDs; he found that 'when they claim to carry out empirical research, they justify their methods largely implicitly, marginally in an explicit manner, or even not at all'.[8] Carrying out empirical research without being transparent about methodology is not a good sign for a field of study. To show how, in a *specific type of empirical research, i.e. evaluations*, theory, methodology and empirical research are combined and carried out, may, we hope, lead to new insights and experiences which are relevant for the current debate on legal research. An example is the practice of theory-driven evaluations, in which argumentation analysis and visualization software, in-depth interviewing and other types of data collection and research reviews are used to unravel or *unpack* underlying normative and causal assumptions of policy makers and lawyers. Another example refers to the diverse quasi-experimental designs developed to evaluate the impact of policies and programmes. Both topics are discussed in this article. By being more precise about what *evaluation studies* have to offer to legal research, this could help the debate and possibly reduce the gap between empirical approaches and normative questions.

The paper first describes some (historical) characteristics of evaluation practice. Then several examples of evaluations related to legal topics are presented while a more systematic

3 Thanks to an anonymous reviewer who suggested asking this question and who also suggested a few other questions. One concerned the problem of the utilization of evaluation findings by lawyers, while the other addressed the ways in which legal scholars and practitioners deal with evaluations that present contradictory outcomes. Due to space limitations we have not been able to discuss these questions here.

4 See: *Journal of Empirical Legal Studies, Law & Social Inquiry* and the *Journal of Legal Studies*. See also: <http://www.elsblog.org/>. Topics studied are, for example, 'Competition in the Courtroom: When Does Expert Testimony Improve Jurors' Decisions?' and 'Empirical Research for Public Policy: With Examples from Family Law'.

5 See, for example, Gordon (The Empiricists: Legal Scholars at the Forefront of Data-Based Research. <http://stanfordlawyer.law.stanford.edu/2010/05/the-empiricists/>) who described the role of empirical/data-based legal research at Stanford University in the USA. Many other examples are available.

6 J. Smits, *Omstreden rechtswetenschap*, 2009, p. 15.

7 H. Tijssen, *De juridische dissertatie onder de loep. De verantwoording van methodologische keuzes in juridische dissertaties,* 2009, p. 207.

8 Tijssen studied 90 legal PhDs. His central question was how research choices in recent Dutch legal doctoral theses have been justified.

overview of types of evaluations (ex ante, process and impact evaluations) is then presented, linking them to legal topics and approaches. Finally, some conclusions are drawn.

3. The history of evaluations

During the first part of the 20the century, in particular in the US, UK, Sweden and to a lesser extent the Netherlands, evaluations started to take place.[9] An early Dutch example stemmed from 1921. In that year de Jongh received his PhD at the University of Amsterdam for a quasi-experimental study on the changes in moral judgments of boys and girls of different age groups. The experimental groups were in custody at that time, and the control groups consisted of 'normal' juveniles. It took a further 40 years before another (criminological) experiment was carried out, but during the last 20 years much more experimental studies have been reported.[10] The results of experimental evaluations can be found in knowledge repositories such as the Campbell Collaboration Crime & Justice Group (<http://www.campbellcollaboration.org/crime_and_justice/index.php>). This site summarizes the results of hundreds of (experimental) evaluations (and there are several more available).[11] The 'what works' tradition in criminology started a little earlier but it also looked into the impact that sanctions and behaviour modification programmes had on preventing and reducing (juvenile) delinquency. Already during the 1950s in the UK, experimental evaluations of prison regimes were being carried out. The 1960s came to be known in the USA as the Golden Years of Evaluation, due to President Johnson's Great Society Policy. Many programmes, serious public money, and ambitious goals, together with the (beginnings of) sunset legislation, resulted in evaluations rapidly becoming an important 'instrument' to understand what policies and laws were 'doing' in reality. This type of work diffused to other countries in the (Western) world, later also including other continents. Nowadays, a large number of (national) professional evaluation societies, academic journals, professorships, master's programmes and – even — an International Atlas of Evaluation are available.[12] The more governments expand their activities by working with sticks, carrots, sermons and pillories, the more attention is paid to evaluation (systems). The accountability and transparency of the public sector is one of the driving forces behind this.

4. Examples of studies linking evaluation and legal questions/research

I will present six examples of studies that are evaluative in nature and address interesting (legal) questions to show links between evaluations and legal research and practice.

The first study is one of the (if not the) oldest evaluation of legislation.[13] Aubert's goal was to 'establish the extent to which behaviour (of housemaids in Norway) conformed to the rules laid down in the Law on Housemaids of 1948'. The purpose of this law was 'to protect the interests of domestic help'. Aubert used a (probability) sample of some 200 housewives and 200

9 The very first impact 'evaluation' ever made and registered was carried out by a British Royal Navy ship's doctor who experimented with different treatments trying to fight scurvy on naval ships. See R. Youngson & I. Schott, *Medical Blunders: Amazing True Stories of Mad, Bad, and Dangerous Doctors*, 1996.

10 G. Bruinsma & D. Weisburd, 'Experimental and quasi-experimental criminological research in the Netherlands', 2007 *Journal of Experimental Criminology*, no. 3, pp. 83-88.

11 H. Hansen & O. Rieper, 'Institutionalization of Second-Order Evidence-Producing Organizations', in O. Rieper et al. (eds), *The Evidence Book: Concepts, Generation and the Use of Evidence,* 2010, pp.27-52.

12 F.L. Leeuw & J.E. Furubo, 'Evaluation systems: what are they and why study them?', 2008 *Evaluation*, no. 14, pp. 157-169; F.L. Leeuw, 'Evaluation Policy in the Netherlands', 2009 *New Directions for Evaluation*, no. 123, pp. 87-103.

13 V. Aubert (ed.), *Sociology of law*, 1969; V. Aubert, 'Some social functions of legislation', in V. Aubert (ed.), *Sociology of law*, 1969, p.116.

housemaids in Oslo and interviewed them about their conduct, information (about the Law itself), attitudes and motives in so far as these had a bearing upon the content of the Law on Housemaids. 'It has to be concluded that the law was, at least for some years, ineffective in the sense that actual conditions of work remained at variance with the norms laid down'.[14]

The second study is a synthesis of the results of over 50 impact evaluations of Dutch laws spanning the period 1995 to 2007. It is largely desk research. Klein Haarhuis & Niemeijer[15] not only show that the impact of laws on addressees is often rather restricted, but they also present an explanation, based on an analysis of the (behavioural, social and institutional) mechanisms that are believed to make laws 'work' (but apparently do not always do so).

A third example concerns the effectiveness of the complaints procedures of the Dutch National Ombudsman.[16] The authors use the Ombudsman's complaints database which contains information concerning over 140,000 dossiers covering 25 years. They applied a mix of conventional database techniques and a genetic-based mining algorithm[17] to scan and map this database. One of the questions was what the success rates are of different types of intermediaries in *reaching* the Ombudsman and in persuading the Ombudsman to label the complaint as *(partially or fully) founded.* Intermediaries are organizations like legal aid institutions, trade unions, and lawyers that help to 'link' and submit individual complaints to the National Ombudsman. Apparently there was a need for this type of linking arrangement. The authors defined the success of complaints in two ways, i.e. complaints that were labelled as *fully or partially founded* and as *fully founded* by the Ombudsman. Table 1 presents some results. It must be noted that '*reports*' refer to the number of reports that were yielded due to complaints being submitted by an intermediary.

If we look into the success rates (in terms of 'fully founded cases') over the years, the 'best' intermediary has been an association or foundation (for example, the Netherlands Refugee Foundation); 54% of the complaints submitted by this type of intermediary are declared to be founded by the Ombudsman. The second best type of intermediary are private individuals (51%) and the third most successful type of intermediary are the legal aid agencies/Legal Services Counters (50%). The relatively low success rate of lawyers is remarkable, since lawyers may be considered to be professionals with a special interest and expertise in this field.

14 V. Aubert, 'Some social functions of legislation', in V. Aubert (ed.), *Sociology of law*, 1969, p. 121. In his book several other empirical studies dealing with the behaviour of the judiciary, the role of courts and legal decision-making can be found. Nowadays some of these studies would be labelled as 'evaluations'.

15 C. Klein Haarhuis & B. Niemeijer, *Wet en Werkelijkheid*, 2008.

16 J. van Dijk et al. , *Analyzing a complaint database by means of a genetic-based data mining algorithm*, WODC memorandum, 2009.

17 R. Choenni, 'Design and Implementation of a Genetic-Based Algorithm for Data Mining', 2000 *Proceedings of the 26th International Conference on Very Large Data Bases,* pp. 33-42.

Table 1 Success rates of different types of intermediaries.

Type of intermediary	Reports	Success rate (fully or partially founded)	Success rate (fully founded)
Association/foundation	133	86%	54%
Private (individual)	127	82%	51%
Legal aid agency / Legal Service Counter	299	79%	50%
Trade union	81	70%	48%
Citizens Advice Bureaus	62	84%	47%
Lawyer	1,590	82%	43%
Administrative office/tax consultant	352	65%	38%
No intermediary	9,333	73%	38%
General social services	90	74%	37%
Action group/interest group/information lines	73	84%	27%
Legal complaints office	213	74%	18%

The fourth example is an evaluation published by Eric Posner and Miguel de Figueiredo[18] addressing the question whether the International Court of Justice (ICJ) is biased (in terms of the members of the ICJ voting in the interests of the states that appoint them). As prior empirical studies were ambiguous according to the authors, they test the charge of bias using statistical methods. Amongst other things, they find evidence that judges favour the states that appoint them and that judges favour states whose wealth level is close to that of the their own states.

The fifth example is Richard Posner's analysis of the impact of regulations on sexual behaviour and relationships.[19] This study used existing literature from a multitude of disciplines, including demography, biology, law and economics, to sort out what impacts, if any, laws of a different nature have on relationships and sexual behaviour.

Farnsworth's *The Legal Analyst* (2007) is the sixth example; it is a textbook that shows how the study of law is related to social and behavioural sciences and evaluation. Farnsworth acts like a legal evaluator and a good one too. He unravels ex ante and ex post legal decisions in terms of (behavioural) mechanisms that are in part responsible for their 'effectiveness'. The book describes mechanisms such as the slippery slope mechanism, cognitive biases, the role of incentives and many more. Evaluators have for a long time paid attention to detecting and testing theories underlying policies and programmes in which mechanisms, contexts and 'outcomes' are central.[20]

What these examples show is that there are strong links between evaluation and legal questions. Let us therefore go into more detail by presenting four different *types of evaluation studies* and link these approaches to legal questions and studies.

18 E.A. Posner & M.F. P. de Figueiredo, 'Is the International Court of Justice Biased?', 2005 *Journal of Legal Studies,* no. 34, pp. 599-630.

19 R. Posner, *Sex and reason*, 1992.

20 J. Elster, *Nuts and Bolts for the Social Sciences,* 1989; J. Elster, *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences,* 2007*;* B. Astbury & F.L. Leeuw, 'Unpacking black boxes: mechanisms and theory-building in evaluation', 2010 *American Journal of Evaluation* 31, no. 3, pp. 363-381.

5. What are evaluators doing and why is it relevant for legal research?

5.1. Ex ante evaluations

A first activity of evaluators is to predict what the consequences of policies, rules, legislation and other 'tools of governments' will be once they are implemented. This is often referred to as 'ex ante' evaluations or 'prospective evaluations'.[21] Evaluators do this by unravelling the behavioural, social and institutional assumptions that underlie the policy, law or programme to be implemented. To put it a little differently: by articulating the 'theory' underlying the intervention, they try to get hold of what is to be expected when it is implemented. The more evidence-based this so-called 'policy or intervention theory' is, the larger the likelihood that the policy will make a difference. Questions that in practice (in the Netherlands) have recently been addressed by evaluators are the following:

– Why is it believed that introducing a new sanction focusing on behaviour modification by young offenders will result in less reoffending?
– Why is it believed that a new and more solid covenant between the government and internet service providers will help to prevent and reduce cybercrime?
– What are the conditions under which a multimedia mass communication campaign focused on a safe(r) use of the internet will be effective in modifying unsafe surf behaviour?

Evaluators also look into the 'operational logic' of prospective interventions and arrangements: who has to do what, when, for how long, under which (legal) constraints and opportunities to make the implementation a success? A similar question is whether the actors involved are able and competent to deliver.

Sometimes this work is done because legislators do not have enough information or are not able to *choose* between two or more options. An interesting case concerns two competing proposals for a law focused on reducing unwanted teenage pregnancies. The US GAO[22] decided to articulate the theories underlying the two *(possible) prospective laws*, confronted them with evidence from behavioural research and presented the findings to the US Congress.

In the Netherlands legislators are asking similar questions to some extent, as is indicated in a concept like 'evidence-based laws'.[23]. The '*Aanwijzingen voor de regelgeving (Art. 6 e.v.)*' (Drafting instructions for legislation (Art. 6 et seq.)) are an important guidance. They specify the actions with which lawmakers and legal policy should become involved when starting to produce regulations and legislation. Some of these actions are:

– That (empirical) evidence has to be found regarding facts and contexts that are relevant for the law which is to come;
– the goals of the prospective law have to be articulated as clearly as possible;

21 US GAO, *Prospective evaluations methods*, 1995; J.M. Verschuuren (ed.) *The impact of legislation: A critical analysis of ex ante evaluation,* 2009, pp. 3-10.

22 US GAO, *Prospective evaluation methods*, 1995.

23 R. van Gestel, 'Evidence-based Lawmaking and the Quality of Legislation. Regulatory Impact Assessments in the European Union and the Netherlands', in H. Schäffer & J. Iliopoulos-Strangas (eds.) *State Modernization in Europe ,* 2007*,* pp. 139-165; J.M. Verschuuren (ed.), *The impact of legislation: A critical analysis of ex ante evaluation*, 2009, pp. 3-10.

– the relationship between the goals of the law, self-regulation and (governmental) regulation has to be articulated with the objective being to clarify that governmental actions (in terms of legislation and regulation) are needed.

The Dutch Cabinet has also started an experiment in which the *Framework for Integral Decision making for Policy and Regulation (IAK)* is used to help realize a reduction in the administrative costs of regulation and legislation but also to realize more transparency and coordination between policies, legislation and implementation.[24]

In his book on law as an argumentative discipline, Smits[25] also addresses an ex ante question. One of his points is that legal research is comparative in nature. By comparing different (legal) systems and framing the comparison in terms of which one is better, Smits believes that it is possible to answer the questions of what the *homo juridicus* should and should not do. The problem, however, is *how* to decide on what is 'better'. Smits' reply is to find the answer 'in the normative *presuppositions* underlying the acceptance of an argument' (emphasis added). However, *how* these presuppositions can be found in a valid and reliable way and *how* they can be 'tested' (against each other) is unclear. Here, the ex ante evaluation approach, in particular the one that follows a theory-driven methodology, can help. Methodological rules, software to visualize argumentations and ways as to how to (empirically) test presuppositions are nowadays available in the evaluation literature.[26]

5.2. Process evaluations

A second type of work which evaluators are doing is to find out to what extent policies, programmes, regulations and other types of measures have been implemented *as agreed*. This type of work is usually called 'process evaluations' and is descriptive in nature. Here evaluators investigate the execution of (court) orders, including sanctions, behaviour modification programmes and financial penalties. The central question is the following: *are the policies and laws and other (legal) arrangements executed as intended, respectively as was formally agreed upon?*

Examples of the relevance of this type of work for legal research are easy to find. Take the problem of sanctions which courts decide to impose. They range from penalties and community service orders to behavioural interventions and incarceration. To what extent their execution is carried out in a way that is compliant with the goals and procedures of the intervention has been found not to be a great and challenging question for the (Dutch) courts. Boone et al.[27] recently evaluated the ways in which (Dutch) 'courts obtain information on the execution and efficiency of sanctions and how this information is incorporated in their rulings. The interviews present a fragmented picture. Courts have different sources with information on the execution and efficiency of sanctions at their disposal, but whether judges actually use them is mainly a matter of personal interest. (…) *Generally speaking, judges* [interviewed in this study, FL] *are not well informed on the developments in the prison system. They have hardly any idea how the imprisonment they have ordered is executed, when the person whom they have sentenced to imprisonment*

24 H. Lokin, & M.J.W. Stokkermans, 'Uitvoeri(n)g getoetst, uitvoerbaar geregeld', 2008 *RegelMaat,* pp. 37-45.

25 J. Smits, *Omstreden rechtswetenschap*, 2009, p. 53.

26 F.L. Leeuw, 'Reconstructing program theories: methods available and problems to be solved', 2003 *American Journal of Evaluation* 24, no. 1, pp. 5-20; S. W. van den Braak, *Sensemaking Software for Crime Analysis*, 2010.

27 M. Boone et al., *De tenuitvoerlegging van sancties: maatwerk door de rechter*?, 2008, p. 76.

will or may first be released and how imprisonment can be used to have a positive effect on the condemned person's behaviour'. (emphasis added)

Another study looked into compliance with verdicts in Dutch civil law.[28] For the more serious cases, the compliance level after three years is 85% and for cases that can be labelled as a 'judicial decision in a defended case', the percentage is 74. When dealing with default judgments the percentage is much lower: 31%. Eshuis presents explanations for these differences.

A third example is of a different nature; it is a synthesis of 20 process evaluations carried out in the Netherlands. All evaluations addressed the same question: *to what extent are (judiciary and behavioural) interventions confronted with implementation problems?*[29] The study describes a number of recurring implementation problems. Table 2 presents the data. Some of these findings can be linked to insights produced by public choice economists studying the behaviour of public-sector agents. Time and again they found that mechanisms like budget maximization, bureau politics and 'what's in it for me' are important determinants of inefficiency and ineffectiveness in public-sector decision making.[30]

Table 2 Incidence of problems when implementing (penal) sanctions/ behaviour (modification) programmes (based on 20 Dutch process evaluations).

IMPLEMENTATION PROBLEM	Total times found in the (N=20) process evaluations
Item: Collaboration in the (Justice) chain	
Partners do not collaborate in an adequate way / competition between policy actors	***7***
Item: Social acceptance of programmes/interventions	
The acceptance of programmes, interventions by participants /stakeholders is insufficient	***10***
Item: Guidance	
Inadequate guidance documents	***10***
Guidance documents not taken seriously/not followed	***15***
'Freies Ermessen' by 'agents'	***5***
Item: Participants	
Not enough participants (clients, inmates etc.) for the programmes	***9***
Inclusion criteria regarding interventions & programmes not complied with	***8***
Item: Human Resources Management	
Not enough personnel to do the job; too many changes in persons doing the job	***10***
Differences in the quality of personnel and training	***9***
Not enough trainers	***4***

28 R. Eshuis, *De daad bij het woord. Het naleven van rechterlijke uitspraken en schikkingsafspraken,* 2009, p. 113.

29 J. Wieman et al. , *Procesevaluaties: wat kunnen we ervan leren? Een onderzoek naar 20 procesevaluaties van justitiële gedragsinterventies*, 2011 (forthcoming).

30 D.C. Mueller, *Public Choice III,* 2003.

The conclusion is that when (legal/behavioural) interventions are implemented, there is *no a priori evidence that what is decided is in fact implemented.* This can not only lead to problems because stakeholders, Parliament, judges and others do not expect this, but it is also causing problems for evaluators aiming to find out to what extent interventions are effective. When the implementation is (in part) a failure, it is difficult to find out about the impact.[31]

5.3. Impact ('effectiveness') evaluations

A *third type* of work which evaluators carry out is to find out to what extent the goals of a policy, intervention, law, subsidy, levy and other 'tools' have been attained and to what extent this has been caused by the intervention under investigation. This type of work is called impact evaluations or effectiveness evaluations. Examples of questions are the following.

To what extent are styles of oversight and inspection different in terms of impact?
Evaluations of the 'Regulatory Compliance Pyramid' articulate the (institutional) conditions under which 'soft' compliance activities such as information campaigns and incentivization and 'tougher' interventions like fines or imprisonment work best.

To what extent are naming and shaming campaigns and policies able to prevent reoffending among released sex offenders?
Pawson[32] has shown that the chances that naming and shaming interventions are serious in their effects are rather small. To realize less recidivism by released sex offenders through this pillory mechanism is difficult to accomplish. One reason is that a complex set of activities by the policy, the municipality, and neighbourhoods has to take place, and have to be coordinated to make the pillorying work. Pawson also found evidence[33] that the chances of (just in time) detection of released pedosexual offenders by such a 'system' in the US were limited. Nevertheless, recently in the UK, Sarah's Law has been rolled out nationally.[34]

To what extent do laws ('statuses') realize the goals they have set?
Government policies in many European countries are primarily regulated by statutes. In most countries, they are adopted by parliament. They set out – in an abstract and generalized form – the rules that apply to citizens and/or to government agencies and in a number of cases have articulated the specific goals that they want to achieve in terms of modifying the behaviour of civilians, corporations or public-sector organizations. Bussmann[35] shows, for Switzerland, that the evaluation of legislation 'has not generated much interest' in the world of policy and programme evaluation, but 'is becoming an issue' (ibid.). Bussmann[36] also discusses strategies for all-encompassing evaluations of legislation, ranging from experimental designs to more descriptive and even narrative ones.

---

31 If one interprets policy implementation as a 'natural experiment', the existence of implementation problems is less of a problem, because evaluators can monitor the differences in implementation and subsequently use that information to understand what is going on during this process.

32 R. Pawson, 'Evidence and Policy and Naming and Shaming', 2002 *Policy Studies*, no. 23, pp. 211-230.

33 A. Petrosino & C. Petrosino, 'The public safety potential of Megan's Law in Massachusetts: an assessment from a sample of criminal sexual psychopaths', 1999 *Crime & Delinquency* 45, pp.140-158

34 See *The Independent*, 3 March 2010; 2 August 2010.

35 W. Bussmann, 'Evaluation of Legislation: Skating on Thin Ice', 2010 *Evaluation 16*, no. 3, p. 280.

36 Ibid., p. 287.

Impact evaluations not only describe to what extent changes in behaviour, or in other dependent variables, have taken place, but they also address the *attribution question*: *are these changes causally linked to the policy, programme, intervention, law, levy, inspection or any type and combination of interventions and arrangements under review*? This is the core question in any impact evaluation.

Six basic criteria have to be fulfilled in order to be able to speak about an impact evaluation.[37]

The first criterion is to identify the type and scope of the intervention that has to be evaluated. An example is given by Tremper, Thomas and Wagenaar.[38] If the goal is to evaluate the impact of legislation (in the USA), it is important not to restrict the evaluation to the 'statutes passed by a legislative body', but to also take into account 'case law from the court system and regulations adopted by administrative agencies'. Otherwise, the "evidence" that results from such an evaluation is probably incorrect. Another example they give is that the legislator and the evaluator have to be clear on what type of effects are studied: '(a) anticipation effects; (b) lagged effects; and (c) changes in the effect over time (cumulative effects, trajectories of decaying effects)'.[39]

The second criterion is to agree on what is valued. When conducting impact evaluations, evaluators need to ask, next to the question on the impact *of what* and *on what*, also this question: impact *for whom?* The principles to follow here are to agree on the most important, and most valued, objectives of the intervention, and as far as possible to try to translate these objectives into measurable indicators.

The third criterion is to carefully articulate the theories linking interventions to outcomes. The theories refer to the (set of) (behavioural) assumptions underlying the activity, programme, law or other legal arrangement. Why is it believed that the programmes and legal interventions will realize what they intend to realize? How robust are these theories, if one is familiar with results from studies in which these or similar ones have been tested? An example is given by Ehren et al.[40] It deals with the Dutch Educational Supervision Act, implemented in 2003, that guides, to a large extent, the work of the Netherlands Inspectorate for Education. The act specifies the working methods and describes the framework for inspection. The act also specifies certain expectations as to how schools should be inspected, the effects such inspections are expected to have, and how these effects should be realized. Together, these assumptions form the programme theory, underlying the Educational Supervision Act. The authors have identified and articulated this programme theory and have also evaluated the accuracy of the assumptions.

The fourth element to address is the attribution problem. This is the problem of how to be sure that the policy programme is indeed resulting in or contributing to changes in behaviour. To address this problem, the design of the evaluation needs to be able to compare a situation with and a situation without the policy or legal arrangement. There are different approaches and designs available to make this doable. In the world of penal law and criminology, many examples ranging from evaluations of behaviour modification programmes to reduce reoffending and boot camps to restorative justice and victimization programmes are easy to find in the repositories of, among others, the Campbell Collaboration Crime and Justice Group (<http://www.campbellcollaboration.org/reviews_crime_justice/index.php>). They also describe the protocols that are

37 F.L. Leeuw & J. Vaessen, *Impact evaluation and development,* 2009.
38 C. Tremper et al., 'Measuring Law for Evaluation Research', 2010 *Evaluation Review* 34, no. 3, p. 243.
39 Ibid., p. 249.
40 M. Ehren et al., 'On the Impact of the Dutch Educational Supervision Act', 2005 *American Journal of Evaluation* 26, pp. 60-76.

used to deal with the attribution problem. Evaluations of interventions focused on preventing or reducing (persistent) antisocial behaviour that combine social, psychological and neuro-biological knowledge have recently become available.[41]

The fifth criterion is also a methodological one: use a mixed-methods approach during data collection and analysis and apply the logic of the comparative advantages of methods and designs. A lens by which to examine these comparative advantages is the four different types of validity:

– *Internal validity*: Establishing the causal relationship between intervention outputs and processes of change leading to outcomes and impacts.
– *Construct validity*: Ensuring that the variables measured adequately represent the underlying realities of development interventions linked to processes of change.
– *External validity*: Establishing the generalizability of findings to other settings.
– *Statistical conclusion validity*: For quantitative techniques, ensuring the degree of confidence about the existence of a relationship between intervention and impact variable and the magnitude of change.

The sixth and final element is to build on existing knowledge relevant to the impact of interventions. What is known as a review and synthesis approach can play a pivotal role in marshalling existing evidence to deepen the power and validity of an impact evaluation, to contribute to future knowledge building, and to meet the information needs of stakeholders. Specifically, these methods can serve two major purposes:

– They strengthen external validity by evaluating comparable interventions across different countries and regions – thus assessing the relative effectiveness of alternative interventions in different contexts.
– Because many interventions rely on similar mechanisms of change, they help to refine the hypotheses or expected results chain to assist in greater selectivity for the impact evaluation.

There are several methods that fall into this category.

– *Systematic reviews* are syntheses of primary studies that, from an initial explicit statement of objectives, follow a transparent, systematic, and replicable methodology of literature search, the inclusion and exclusion of studies according to clear criteria, and extracting and synthesizing information from the resulting body of knowledge. Examples can be found in the Campbell Collaboration repository.
– *Meta-analyses***,** a common type of systematic review, quantitatively synthesizing 'scores' for the impact of a similar set of interventions from a number of individual studies across different environments. They follow a strict procedure to search for and select appropriate

41 C.H. de Kogel, *Hersenen in Beeld*, 2009. The (Dutch NSF) program on Brain & Cognition coordinated by WODC (<www.wodc.nl; http://www.nwo.nl/nwohome.nsf/pages/NWOP_88HEHW_Eng>) has as one of its three focal points safety and security issues. Experimental evaluations look into the effectiveness of innovative approaches to prevent and reduce persistent antisocial behaviour, where bio-markers, socio and psycho markers are central. Another related field of impact evaluations deals with the role that food (supplements) play in reducing aggressive behaviour in, for example, prisons. See: A. Zaalberg et al., 'Effects of Nutritional Supplements on Aggression, Rule-Breaking, and Psychopathology Among Young Adult Prisoners', 2009 *Aggressive Behavior* 35, pp. 1-10.

evidence, typically using a hierarchy of methods, with more quantitatively rigorous (experimental) studies being ranked higher as sources of evidence.

– *Realist syntheses* are theory based and do not use a hierarchy of methods. They collect earlier research findings by placing the policy instrument or intervention that is evaluated in the context of other similar instruments and describe the intervention in terms of its context, social and behavioural mechanisms (what makes the intervention work), and outcomes (the deliverables).

Examples of these reviews and synthesis studies in the field of law can be found at the *UK Centre for Evidence and Policy*, the *Campbell Collaboration Crime and Justice Group*, and in specialized journals such as the Journal of Experimental Criminology. An overview of a number of these *repositories* is given by Hansen & Rieper.[42]

Although the attention paid to this type of studies in the world of criminology and penal law is probably greater than in other fields of law, studies on dispute regulation, trust and networks, compliance with contracts, and the role of the embeddedness of contracts are rapidly becoming important. An interesting example can be found in *E-trust*[43] that explores territory where trust, reputation, rules and online relationships intersect. In this book experimental studies and field research are used to examine how trust in anonymous online exchanges can create or diminish cooperation between people. In the world of civilology (= combining civil law and behavioural studies) experiments are also being carried out.

5.4. Organizational evaluations

The fourth *type* of work which evaluators are doing is to evaluate *organizations, how they operate, how they monitor themselves, and how they interact with professional fields and the surrounding environment.* Evaluations of the functioning of hospitals, higher education organizations, prisons, schools, and companies are carried out by using benchmarks that utilize the results from performance monitoring and auditing and that make it possible to compare the performance of (similar) organizations. Sometimes, case studies are preferred. Courts, prisons, offices of the public prosecutor, organizations of notaries, lawyers and bailiffs and many others are being evaluated in terms of their (decision-making) activities, time management, efficiency and effectiveness.

An interesting example that was recently published focused on a (temporarily) new legal organization and the way it functioned: the Dutch CEAS (the *Closed Criminal Cases Evaluation Commission* (*Dutch acronym*: CEAS). The CEAS is a provisory evaluation organ meant to investigate possible mistakes in the stage of criminal investigation/prosecution, which possibly resulted in a wrongful conviction.[44] De Ridder, Klein Haarhuis & de Jongste[45] carried out the evaluation that was 'promised to the Netherlands House of Representatives by the Minister of Justice at the time of the foundation of the CEAS in 2006. The Commission was established following unrest in political circles and in the media that arose after the judicial errors in the

42 H. Hansen & O. Rieper, 'Institutionalization of Second-Order Evidence-Producing Organizations', in O. Rieper et al. (eds.), *The Evidence Book: Concepts, Generation and the Use of Evidence,* 2010, pp. 27-52.

43 K.S. Cook et al. (eds.), *E-trust*, 2009.

44 Based on Carolien Klein Haarhuis & Willemien De Jongste (2010).'Reconstructing and Assessing the Evaluation Logic of the Dutch Closed Criminal Cases Evaluation Commission: Report of a Meta-evaluation', 2010, *Evidence & Policy*,6: pp. 483-503.

45 J.A. de Ridder et al., *De CEAS aan het werk; Bevindingen over het functioneren van de Commissie Evaluatie Afgesloten Strafzaken 2006-2008*, 2008, pp. 168-169.

Schiedam park murder case and the publication of the evaluation report by Advocate-General (A-G) Posthumus'.

The task of the CEAS is to examine, by means of an investigation, whether severe shortcomings arose in the investigation, prosecution and/or presentation of evidence at a court session in a specific criminal case, which prevented the court's impartial evaluation of the case. The methodology combined a process evaluation with an audit approach. The criteria looked into were the independence of the CEAS, its impartiality, its comprehensibility, its timeliness and its conclusiveness.

6. Conclusions and discussion

'It is sometimes striking how much the vocabulary of sociology resembles the language of law. The emphasis upon rights, obligations and expectations, upon sanctions and predictability within sociology has its counterpart in the sophisticated analyses of these concepts in the tradition of legal scholarship'. With these words Aubert[46] opened his book on the sociology of law. Now, 40-plus years later, one could say something similar with regard to the relationship between legal scholarship and evaluations.

This paper has brought together a number evaluation approaches that are of direct relevance for questions and empirical research that lawyers are carrying out. Legal research often concerns topics that resemble central evaluation questions: *will it work? Is it going to be implemented (or: executed) in a way which is compliant with what was agreed upon? And: has it worked?*

This brings us to two final questions.

The first is to what extent the different legal fields like penal law, constitutional and administrative law and civil law are confronted with enough substantive and empirical challenges to make the application of methods and theories from evaluation studies, relevant and acceptable in the academic training of lawyers and their scholarly practice.

We believe this is the case.

For penal law, the relationship with evaluation has existed for quite some time. Criminologists, sociologists and to a lesser extent economists are involved in evaluating penal laws and sanctions, programmes and interventions.[47] This tradition goes back many decades.

For the field of constitutional and administrative law, the relationship is more recent and probably less intense. However, over the years lawyers have been confronted with questions of an empirical nature. Examples are why laws do not always realize their goals, how the existence of unintended side-effects of regulations and inspections is to be explained, why the trust in and the acceptance of state interventions varies over years and countries and how civil society 'uses' state organizations such as the National Ombudsman and others for their goals.

Civil law and evaluation is probably the most recent combination. What has come to be known as civilology has increased the relevance of this combination.[48] Civilologists study the determinants, dynamics and delusions of judicial decision making, pay attention to the embeddedness of the legal procedures and the impact of social and psychological factors thereon. If judicial decision making is the culminating point of the legal procedure, this is not a process

46 V. Aubert (ed.) *Sociology of law,* 1969.

47 H. Nelen, *Evidence maze; het doolhof van het evaluatieonderzoek*, 2008.

48 W.H. van Boom et al. (eds.), *Gedrag en privaatrecht. Over gedragspresumpties en gedragseffecten bij privaatrechtelijke leerstukken,* 2008.

which is done in isolation but the final phase of a multistep process: how is this judgment influenced by these consecutive steps? Topics that are studied deal, among other things, with interactions between claimant-plaintiff, client-lawyer and between judges in group decision making but also the impact of ASBOs (anti-social behaviour orders) in the UK and to a limited extent in the Netherlands.[49]

The second question is how to make sure that empirical research and in particular evaluation studies will be able to play a systematic role in legal studies and training. At least three conditions have to be met. The first is to be able to *unpack* legal questions and distinguish between normative and empirical aspects. Argumentation analysis and the visualization of arguments (software),[50] together with content analysis will help. The second condition is to be able to *interpret legal solutions to problems as testable 'theories'* (or a set of assumptions). Evaluators have developed several approaches known as TDE (theory-driven evaluations) that can be used to analyze these 'theories'. The third condition is to become acquainted with the idea of the 'empirical cycle'; in such a cycle questions are formulated, theories are seen as (preliminary) answers to these questions and research designs, methods and techniques of data collection and analysis are seen as mechanisms to test or validate the 'theories'. The final loop of the 'empirical cycle' is to provide feedback to lawyers, making it possible for them to develop their ('evidence-based') answers to the legal-normative questions. Often, if not always, a new 'cycle' then starts again with new, sometimes higher-order (research) questions. To train lawyers and law students in this line of thinking and practice will help to make the combination of evaluation studies and legal research a promising one.

49 Based on information regarding the Seminar on 'Judicial Decision Making in Civil law: Determinants, Dynamics and Delusions' (9 November 2010) organized by Erasmus University Rotterdam. See M.T. Croes for a study on the impact of ASBOs: 'Gedragseffecten van Anti Social Behavior Orders', in W.H. van Boom et al. (eds.), *Gedrag en privaatrecht. Over gedragspresumpties en gedragseffecten bij privaatrechtelijke leerstukken,* 2008, pp. 561-601.

50 See S. W. van den Braak, *Sensemaking Software for Crime Analysis*, 2010 for an empirical analysis of argumentation-visualization programmes (some of them open source, some commercial) and their application in crime analysis.