Evaluation and Limitations of Social Interventions: The Case of Spain

Juan Andrés Ligero Lasa
University of Madrid

Background: Although the number and practice of evaluation varies enormously among policy areas, there are very few studies about the unequal evaluation development in the different policy sectors.

Purpose: This article aims to (1) acknowledge the different evaluation development among different policy sectors in Spain (2) identify the factors and causes that provoke this disproportion, and (3) explore potential consequences of this unequal distribution of evaluation studies among policy areas.

Setting: Spain.

Intervention: Public polices in Spain.

Research Design: A sample of evaluations is classified by policy sector and the number of evaluations in each sector is analyzed and compared. Then, other significant variables are identified for explaining differences among sectors.

Data Collection and Analysis: The cases (evaluation studies) are drawn from two samples: (1) a data base of evaluation studies and (2) a survey to Spanish evaluators held in 2009. The comparison was done with difference in proportions, adjusted standardised residuals and crosstabs.

Findings: Analysis of Spanish evaluations shows that program evaluations are much more frequent in the social policies' area than in the areas of security, defense or justice. A variable with a high ability to predict whether or not evaluations will be carried out is identified: the selective versus universal nature of the policies being evaluated. Selective interventions are more frequently evaluated than universal policies. This lack of balance makes selective interventions more prone to severe critical analysis. This evaluation bias, in turn, produces a series of perverse effects such as a greater probability of cutting down programs based on selective application strategies.

Keywords: *Evaluation, social policies, pubic policy, Spain, policy limitation*

There is still much work to be done in Spain in the field of evaluation. This is easily explained looking at its slow evolution since its early steps.

It is not until the 1980s and 1990s that we can find a substantial body of work written in Spanish (either original or translated), even though there was very interesting work on evaluation published before. During these two decades in Spain, there were different public institutions whose functions included evaluation. However, this work was not substantial enough to create a

comprehensive culture that would spread to other institutions. In 2002, The International Atlas of Evaluation (Furubo et al.) placed Spain in the lowest rank of the 21 countries under study.

It was not until the turn of the century that evaluation took off in Spain with a steady pace (Pazos & Zapico-Goñi, 2002; Bustelo 2006; Fernandez-Ramírez & Rebolloso, 2006). Some of the signs of this growth are:

- In the year 2001 the Spanish Evaluation Society is created to foster the development of evaluation culture.
- A supply of specific graduate and post-graduate training appears. Evaluation courses are offered within different undergraduate degrees. There are several postgraduate Master's degrees, the first was offered in 2002.
- Also in this year, the National Agency for Quality Assessment and Accreditation (ANECA) was set up. In 2004 a Commission was also created to build the Agency for Evaluation (National Agency for the Evaluation of Policies and Quality of Services) founded later in 2007. These initiatives evidence the presence of a political will supporting and endorsing the stronger development of Evaluation in Spain.
- A significant growth in the production of papers, theses and books on the subject, although it is still meager.
- The above mentioned and other factors have contributed to the emergence of an incipient professional market (III Seminar on Evaluation Experiences for Programmes and Policies, 2006).[1]

The development of this discipline is no doubt satisfactory. The improvements which evaluation seems to bring to public action will no doubt be beneficial for the general welfare. However, although we are still in the early days, it is worth questioning whether evaluation is developing adequately.

A possible deficiency in the way in which the professional field is taking shape is the fact that certain public policy sectors are subjected to evaluation more often than others. This unequal application of evaluation would have consequences on the development of policies themselves and, therefore, on the target populations, point I make on in this paper. The question that arises is whether the same effort is invested in evaluating the actions of the different sectors of public intervention.

In Spanish literature there are good, well supported articles on the state of the evaluation issue, meta-evaluation projects and case studies[2] but there are no studies

[1] Entitled "Professional Development: Emerging Fields in Evaluation" held in the Centre for Management Studies of the Complutense University of Madrid on June 1st 2006.

[2] Amongst the possible references it is worth highlighting the following as an example: the chapter by Pazos and Zapico (2002) for the International Atlas of Evaluation, the Commission report for the Study and creation of the National Agency for the Evaluation of Public Policies and Quality of Services (AEVAL) (2004), the paper by Fernández and Rebolloso referred to above, the article by Bustelo (2006) on the development of an evaluation culture, the issue by Garde on the institutionalisation of evaluation in Spain (2006), the document for the European Commission by Bustelo et al (2006) and the article by Díaz-Puente, Cazorla and Borrego (2007) in which they analyse international evaluation journals and present data for Spain. Metaevaluations or case

examining the incidence of evaluation broken down by political sector. Despite the fact that research on evaluation has increased substantially over the last years still "we lack systematic studies describing the varied and little known Spanish reality" (Fernandez- Ramírez & Rebolloso, 2006, p. 135).

Taking into account local studies only, it is impossible to know what the level of evaluation development by sectors is. It is necessary to use other approaches of investigation in order to reply to the question. The following section describes the methods and the choice selected.

Method

In an article about evaluation activities in Europe, Leeuw, Toulemonde and Brouwers (1999) suggest certain ways of monitoring growth (or decline) of evaluation activities, such as interviews, taking stock of evaluation reports, bibliometrics and indirect evidence. In their case they choose a survey carried out with evaluation service providers, both in the private and the academic sectors.

Other significant studies (although already old) on evaluation by political sectors have obtained information through an evaluation sample. This is the case of the work by Nioche and Poinsard (1984) in France, with 269 evaluations. Another example is the analysis of contracts offered by the United States administration, such as the ¨Request for Proposal¨ study included in Freeman and Solomon (1981).

Amongst the different strategies that have been identified for the present essay, two evaluation samples have been selected for analysis. One of them is made up of the information presented in evaluation congresses and seminars held in Spain and the other is the result of a self-completion survey carried out on evaluators.

The other options were more complex and difficult to put into practice. In Spain there is no obligation or reason to motivate communicating evaluation assignments to any register or data base, which do not exist anyway. If these registers actually existed, as Leeuw et al. (1999) state, given the large number of evaluation instances, they could very well become unmanageable. On the other hand, bibliometric studies on evaluations in the Spanish context are not very reliable, as there are no specialized publications or a shared and extended tradition amongst the different institutions of publishing their reports or passing them on to documentation centers. The two samples used are described below.

Study of the Evaluation Database

The Master on Evaluation of Public Programmes and Policies of the Complutense University of Madrid has compiled an evaluation data base in order to provide students, academics and professionals with references. All the evaluations that had been presented or mentioned in different specialized evaluation forums between 2001 and May 2009[3] were identified, classified and

studies include the work by García Sánchez (2005) on educational reform, Bustelo (2004) on gender equality policies and Díaz-Puente, Yagüe and Alfonso (2008) on European structural funds, which present information on their own sector alone.

[3] Presentations in evaluation congresses or seminars: I, II, III, IV, V Seminar on Evaluation Experiences, UCM (2004, 2005, 2006,2007 and 2008);IV, V Congress of the Spanish Evaluation Society (2005, 2009);VIII Spanish Congress of

included in a data base. Evaluations of several seminars on drug prevention were also included. Finally, Master's theses were also included whenever the authors gave their authorization.

The evaluation base has been used to learn about the distribution of evaluations according to sectors. Only the evaluations carried out in Spain or financed with Spanish money have been selected. As this is not a census or a random sample and given that the base has been compiled with educational purposes it may be biased in different ways. At least three possible biases have been identified:

1. A large number of drug prevention evaluations in comparison to other intervention sectors were found in the database. The question is if this increased volume of drug program evaluations is a true reflection of an increase in demand for this particular kind of evaluation or a result of the master's emphasis on prevention programs. Database analyses were carried out both with and without evaluations of prevention programs in order to compare results. The outcome was more flattering when the evaluations of prevention programs were included. Consequently, the decision was made to exclude them in order to conduct our study in the most unfavourable context. It has to be noted, however, that its exclusion from the final tally might also be a form of bias. The final data base includes 159 evaluations[4].
2. As only specific evaluation forums are taken into account, and given that in Spain they are usually related to the fields of sociology, psychology, economy, political science and administration, the most evaluated sectors may be those closest to these disciplines.
3. Some sectors may be less inclined to presenting their results in public for different reasons, such as the need to preserve the information (home security, defense...), threats to program continuity or lack of incentives to communicate the information. Therefore, sectors more influenced by these possible factors can have less presence in the data base.

To summarize, first of all, the results obtained from this data base will not be representative of the drug prevention sector. Secondly, evaluations in the fields of social sciences, psychology and politics may be overrepresented. Thirdly, in as far as confidentiality may act in detriment of evaluations, sectors with a greater involvement in security issues or working with sensitive or strategic information for the State may be underrepresented.

Self-Completion Survey of Evaluators

Bustelo and Fitzpatrick carried out throughout 2009 the fieldwork for the research project titled "Evaluation in

Sociology. Evaluation of Social Intervention Programmes. Methodology Working Group (2004);Dissertations for the Master on Evaluation of Public Programmes and Policies, UCM (2003, 2004, 2005,2006, 2007,2008).

[4] The "Evaluaciones" data base is available at www.magisterevaluacion.es.

Spain: Practice and Institutionalization"[5]. The results were presented in the Annual AEA Conference: Evaluation 2009: Context and Evaluation, Orlando, Florida held between November 11th-14th, 2009. The authors designed a self-completion questionnaire to be filled in online on key aspects of Spanish evaluation. At one point, the questionnaire requested that the participant name the last evaluation carried out or any other that they considered important and then went on to ask a battery of questions about that evaluation.

The questionnaire was sent to 356 people who carry out evaluations and which are included in a data base administered by the Master on Evaluation of Programmes. Members of the Spanish Evaluation Society were also invited to reply to the questionnaire. These contacts generated 150 filled in questionnaires. This sample has been used in order to find out the distribution of evaluation by sectors.

Several biases have been identified in this sample:

1. The total sample of evaluators in Spain is unknown, there are bound to be people working in evaluation who are not included in the aforementioned data base or who do not belong to the Spanish Evaluation Society and who will not have received the questionnaire.
2. Self-completion surveys are biased by the fact that those who respond represent the more motivated subjects.
3. The respondent chooses according to their own criterion the evaluation that they will describe.

The same analysis has been carried out on both samples, the one obtained from the evaluation base and the one obtained from the self-completion survey. A team[6] classified the evaluated programs according to the sector that the evaluation referred to (the program, policy or service under evaluation). It was decided to classify them according to areas or ministry sectors, as this is a well-known, convenient and intuitive scheme to organize public affairs, and we used the current Spanish model with some corrections[7]: Culture; Defense; Economy; Education and Sport; Employment and Immigration; Environment, Agriculture and Fisheries; Equality; Foreign Affairs and Development Cooperation; Health; Housing; Industry, Tourism and Commerce; Infrastructures; Justice; Home Affairs; Presidency; Public Administration; Science and Technology and Social Services.

Results

The frequency distribution of the two samples is shown in Figure 1.

[5] I am extremely grateful to María Bustelo and Jody Fitzpatrick for allowing me the use of one of the items' survey results before its publishing.

[6] The classification of the evaluated programs was carried out by the coordinators of the Master on Evaluation Belén Rodríguez, Irene Rosales, Isabel Morandeira and Maruxa Fernández and by Juan Andrés Ligero.

[7] Social Services and Health have been separated. Public Administration has been introduced and territorial policy has not been included. "Fomento" (public works) has been translated as Infrastructures, given that this is one of the main activities of the department.

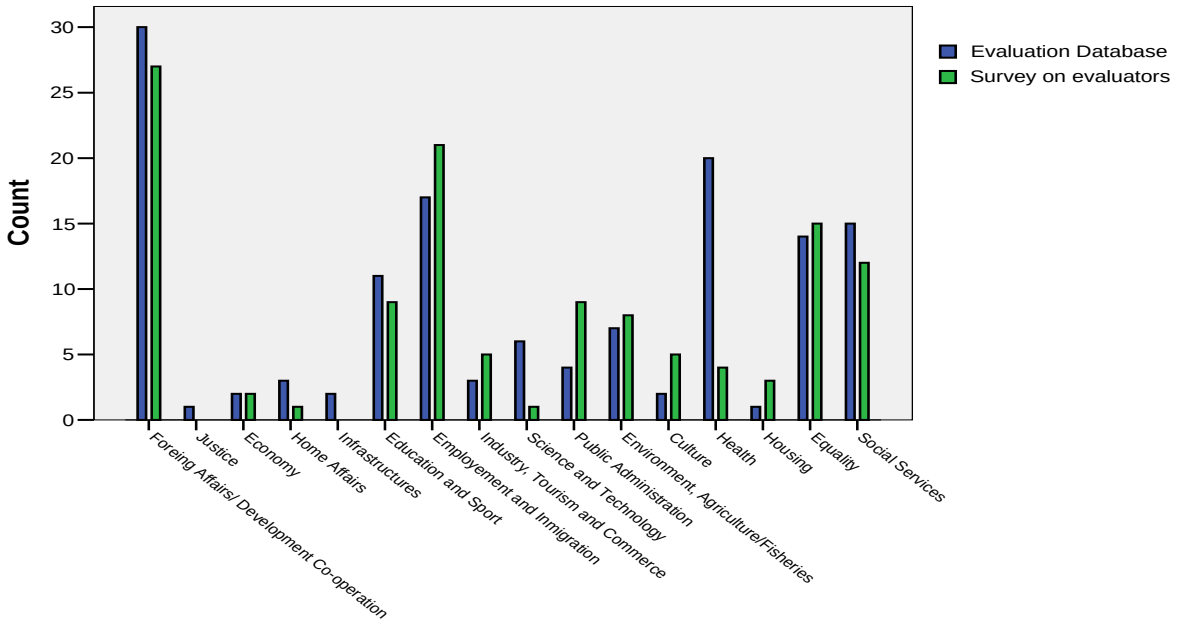

Figure 1. Frequency Distribution of Samples

The first conclusion that can be extracted from the graph is the fact that, basically, the two samples have a similar distribution[8] amongst sectors except for the category of Health, which shows a different behavior.

The second conclusion is that there is not the same volume of evaluations in the different sectors. Some sectors are subjected to a lot of evaluation, such as, for example, Foreign Affairs and Development Cooperation, Employment and Immigration, Equality and Social Services (there are discrepancies in Health). Other sectors, such as Defense, Justice, Infrastructures, Home Affairs, Housing or Economy are less evaluated or even not evaluated at all. This is the case of Defense, which does not even appear in the graph.

The absence of evaluation in certain sectors can be ascribable to the need for confidentiality in their policies. This could be the case of Defense, Home Affairs or Justice. However, this does not apply to other sectors such as Housing, Economy or Infrastructures, which require other possible explanations.

On the other hand, the trend that the data show is familiar. Following Leeuw et al's suggestion to use indirect evidence, Spanish bibliography generally provides more examples and references of evaluations on social intervention, drugs, health, education, gender and employment than on other sectors. This may also be the case in the bibliography for other countries. In the case of the USA, it is more usual to find references to social or educational programs than to other areas. Carol Weiss even states that some of such fields, such as education,

[8] The adjusted standardised residuals were analysed and they showed that there are no statistically significant differences between the pairs of categories in either sample, with the only exception of Health, where differences do appear.

poverty or crime prevention, have been key in developing the discipline of evaluation (1998:12).

All in all, the data from both samples show consistent and significant differences in the volume of evaluations according to sectors, the issue now is to find out why there is this disparity and whether this is in fact a feature of evaluation in Spain.

Explanation

Amongst the different factors[9] which have been considered as possible explanations, there is a variable which partially explains the disparity in the number of evaluations, and it has a bearing on the behavior of political decision makers when commissioning evaluations. This variable has been called "universal or selective nature of interventions."

Universal and Selective Interventions

Universal policies or programs are those interventions which aim to benefit the whole of the population equally in a certain field. Examples include the public health system, public infrastructures, traffic management, and police services or national defense policies.

Selective or *positive discrimination* policies favor a certain social segment or group which differs from the rest because they share a certain characteristic or demographic, social or cultural condition. They are based on the fact that not all the citizenship enjoys the same options or shares the same difficulties. Subsequently, actions are required to give the disadvantaged more opportunities. Amongst the possible examples, there are literacy programs for adults, insertion programs for minorities, rehousing of shanty-dwellers or programs for minimum-income citizens.

Observing the data in Figure 1, it seems that selective policies undergo more evaluation than universal policies, regardless of the political sector in which the interventions are framed. This hypothesis was tested with both evaluation samples.

The evaluations in the samples were classified again according to their universal or selective nature. In this process, doubts emerged regarding some cases such as schools and some rural and local development interventions. In the case of the compulsory educational system, the whole population of a certain age must be enrolled in compulsory education, so it has been considered of universal nature, although it could also be considered a selective intervention. In the case of rural development projects, there were doubts about the actual characteristics of the interventions, and so, they were excluded from the analysis. The samples were reduced to 138 in the case of the data base and 122 in the case of the self-completed survey.

The samples were analyzed with a frequency distribution and the differences in "d" proportions were extracted as shown in table 1.

[9] One of the analysed factors was budget size by public sectors. The Spanish public budget shows that the frequencies shown in Figure 1 have no relationship with the budgets assigned to each area. For example, the areas receiving the largest proportion of the budget are Employment and Immigration, Economy, Infrastructures, Defence and Home Affairs. With the exception of Employment, there are hardly any evaluations for these areas in the samples. However, this can not be considered evidence, as the evaluation samples correspond to different administrations and periods while the public budgets are those for 2009.

Table 1
Differences in Proportions of Number of Evaluations Between Universal and Selective Evaluated Programs

	Universal evaluations over universal evaluated programs (%)	Selective evaluations over selective evaluated programs (%)	Difference in "d" proportions	*N*
Evaluation data base	31.4 %	68.6%	-37.2%	138
Evaluator survey	28.7%	64.8%	-36.1%	122

The two samples show a very similar behavior. The data shows that between 68.8% and 64.8% of the evaluations under study have been carried out on selective programs and between 31.4% and 28.7% have been carried out on universal interventions. The difference between them is about 36% or 37%. In other words, for every universal intervention evaluated, there are approximately 2.2 selective intervention evaluated.

The variable "universal or selective nature of the interventions" partly explains the disparity in evaluations found between the different sectors. If these two variables, "universal-selective" and "political sector", are crossed, a high association emerges between the two in both samples. These adjusted standardized residuals have also been cross-tabulated in order to see the association for each of the categories, the results have been summarized in the tables below[10].

Table 2
Political Sectors by Universal or Selective Strategies in Evaluation Database

Selective (Sectors presenting more selective strategies than expected)	Equal (No significant differences between strategies)	Universal (Sectors presenting more universal strategies than expected)
Foreign Affairs and International Cooperation	Industry, Tourism and Commerce	Economy
Employment and Immigration	Science and Technology	Education and Sport
Equality and Social Services	Health	Public Administration
		Environment, Agriculture and Fisheries
		Culture
		Housing

[10] The adjusted standardized residuals have been cross-tabulated. The chi-square test applied to the evaluation data base and the evaluator survey results in a significance level of $p = 0.00$. The variable categories have been classified according to whether or not they obtained a remainder over 1.96. Cases with values between -1.96 and +1.96 have been classified in the "Equal" category, considering that there were no significant differences.

Table 3
Political Sectors by Universal or Selective Strategies in Survey on Evaluators

Selective (Sectors presenting more selective strategies than expected	Equal (No significant differences between strategies)	Universal (Sectors presenting more universal strategies than expected)
Equality Social Services	Employment and Immigration[11] Foreign Affairs and Development Cooperation Justice Education and Sport Industry, Tourism and Commerce Public Administration Health Housing	Economy Home Affairs Infrastructures Science and Technology Culture

The sectors which emerged as the most subjected to evaluation in Figure 1 are classified in both samples or in one of them as those which include more selective intervention strategies than expected. There are two sectors which do not fit this pattern. Health, which already presented differences between the data obtained from the two samples, does not show a significant difference between one strategy or the other. The other case is Education, which is the fifth most evaluated sector and is classified in one of the samples as universal[12].

On the other end of the spectrum, the least evaluated areas, such as Infrastructures, Home Affairs, Housing or Economy, are classified in both samples or in one of them as employing universal strategies. Justice only appears in one of the samples and does not show a clear trend and there is no information on Defense.

Taking into account these exceptions, there is a trend showing that the most evaluated sectors are those which base their policies more on selective strategies and, on the other hand, the least evaluated sectors are those whose policies are based more on universal strategies.

All in all, the conclusions that can be extracted from the whole analysis is that policies of a selective nature are more subjected to evaluation than universal policies (2.2 selective policies to one universal policy) and this affects the unequal distribution of evaluations by sectors, as certain sectors tend to favor one type of intervention strategy over the other.

Therefore, some aspect of selective intervention strategies must encourage evaluations or, on the other hand, universal strategies may imply a process which discourages evaluation. This is the new question which the data pose.

Discussion

The existence of trends which make some programs more evaluated than others is something that has already been described in evaluation bibliography. Carol Weiss explains that “In the 1960s

[11] The residual is 1.9, although very close to 1.96, it was decided classify it in the equal category.

[12] This category generated doubts as to where to classify school system evaluations. They were finally classified as universal.

and 1970s it was the new and innovative program that was evaluated. The hardy perennials went on without cavil or question, whether o not they were doing much good. And since the new programs of the period were programs for marginal groups, such as poor children, juvenile delinquents, and released mental patients, these were the programs that were scrutinized" (1991, p. 222).

Another example is that of tax expenditure programs in the USA, "unlike most direct expenditure programs, most tax expenditures are open-ended in terms of the amounts involved, they are not subject to annual competitive appropriations, they are permanently authorized, and, until recently, their effectiveness rarely has been evaluated" (Datto & Grasso, 1998, p. 1). Besides, the authors find that tax expenditure programs, at all levels (federal, state and local) "tend to benefit the wealthier, and direct expenditures tend to benefit the poorer".

In both cases, the most evaluated interventions are aimed at populations in difficult conditions or situations. Weiss explains that, partly, this is due to the fact that "programs with powerless clientele may lack the coalition of support that shields more mainstream groups. They may not have developed alliances of interest groups, professional associations, citizen representatives, and bureaucrats that will seek to reduce the intrusion of evaluation" (1991, p. 222).

There are certain similarities between what these authors explain and what this article puts forward with regards the Spanish case. According to the quoted references, the programs which are more prone to undergo evaluation are those which are aimed at the less wealthy population, marginal groups or the powerless. By definition, these programs only intervene in a certain section of society, which would classify them as selective policies.

The explanation which this article expounds is that universal and selective policies generate different reactions and, therefore, different evaluation demands from the whole of society, but also from the policy makers that commission the evaluation.

The fact that not everybody can benefit from a program or a public service may provoke wariness: *if I have the same rights as everybody else, then why can't I benefit from this integration benefit, that rehousing flat or those training courses?* The concern expressed is more than a simple theoretical doubt. It relates to the suspicion that the policy in question may be unfair and that the technical criterion behind the selection process may be arbitrary. There is mistrust that the person benefiting from public money may not fulfill the established requirements.

Mistrust may be so great that, sometimes, concern with "fraud" displaces the very aims of the policies and becomes its main objective. For example, in certain minimum income programs, the control system is so strict and bureaucratic that it consumes more resources than it should, hindering the achievement of the main aims[13].

The doubt about the truthfulness of the beneficiary's condition is not anything new, neither does it respond to greater political technicalities. In the case of policies to combat poverty, the same concern can be found at least since the 18th Century. At that time great

[13] There are exceptions to this logic and some decision makers understand that certain "fraud" percentages have to be taken into account (the estimation is about 5%) in order to avoid excessive bureaucracy and costs.

consideration was already given to distinguishing between the "real poor" and the "fake poor", regardless of the fact that all of them lived on the streets, begged and shared the same life conditions. Before giving any money or goods to a beggar, it was necessary to settle whether the person had really been thrown into precariousness or whether, on the contrary, they were able to work but did not want to:

> The reform of social assistance demanded, beforehand, that false beggars be unmasked, only in this way could every city support its own poor people: if there was some irrefutable way of establishing in each case whether it was real or feigned poverty. In consequence, the aim of many of the acts involved with this issue will be to show false beggars up, that is, all those who could work if they wanted to, in order to dedicate existing resources to assist and take care of the "real" poor people, that is, "those who can not earn their living" and "must [sic] really beg." (Cabrera, 1998,p. 30).

Updating the terms, the same concern appears in contemporary programs dealing with unemployment benefits, rehousing or home care, amongst others. What is actually behind this distinction is the persistent doubt and suspicion of fraud that seem to accompany programs granting goods and services to only a portion of the population.

These reactions become even more extreme when social and personal difficulties are not so obvious. Examples can include the social development problems of "immigration children", lower social classes or inequality between men and women generated by the gender system[14]. These cases sometimes provoke vehement reactions and arguments which even deny the existence of differences or, if they are recognized, defend not investing public money in readdressing these imbalances.

Although this interpretation is focused on the Spanish case, similar reactions to positive discrimination can be found in other countries. In the USA social unrest about positive actions can reach such a degree that this kind of laws has even been banned in certain places. In 2006 at Michigan there was a public consultation through referendum about the legality of this type of strategies. 58% voted against their use and, as a result, the use by public institutions of positive discrimination programs "based on race, gender, ethnicity or national origin" has been declared illegal. (Holusha, 2006). Some months later, the United States Supreme Court declared the positive discrimination policy aimed to promote racial integration in schools unconstitutional (Monge, 2007).

Reflecting upon opportunities and about the efficiency of positive discrimination does not seem to be the issue at hand. What concerns people is the strategy, how it is done, regardless of how satisfactory results may be and how much they may contribute to reducing social inequalities. On the other hand, universal policies do not raise that legislative zeal, even if they deepen and broaden the difficulties they were intended to eradicate.

Evaluation is a useful and legitimate tool to find out more about public intervention and to account to the citizenship for its performance. However, the evaluation system is not neutral and,

[14] System of gender means: "Just like every society has its system of production, there is also a gender system, which is the aspect in social life which organises relationships between men and women" (Britt-Marie Thuren, 1993,p. 97).

as House states, evaluation can be politically driven (2006).

Commissioning an evaluation can also be politically driven, and is part of the political influence game. Patrizi and McMulla (Henry, 1998) found out as a result of their survey carried out on leaders of foundations that one of the main priorities to finance evaluations is the influence they have on public policy. In this sense, Weiss understands that "a political statement is implicit in the selection of some programs to undergo evaluation and others to escape scrutiny and Chelimsky considers that in some cases "study questions may hide a partisan or ideological purpose" (1998, p. 404).

Therefore, it is likely that more evaluations are commissioned in cases in which it is interesting, motivating, convenient and necessary to refute or confirm the efficiency of the program in front of the decision-making social actors. Thus, it is not strange that in this context evaluation is used to analyze those policies that generate most concern.

In practice, if a certain intervention does not raise doubts or questioning, it is likely that the actors will not demand any evaluation, unless it is requested by other instances. Therefore, if selective actions generate concern, they are more likely to call for more evaluations than policies of a universal nature, which do not inspire the same kind of suspicion.

The paradox in this situation derives from the coexistence of unevaluated universal interventions worth millions and more humble selective programs which are evaluated time and time again. In Spain costly projects such as dual carriageways in areas of low traffic density, tunnels, burying roads, frequent refurbishment of urban furnishings and landscape, state subsidies for car purchases, various public TV and radio stations, institutional presence in cultural events or construction of large emblematic buildings are not only not evaluated (except for auditing and accounting controls) but neither is there any manifest public demand requesting that they are. Nothing is known about their results, appropriateness, prioritization of social needs or about the opportunity cost of such investments.

On the other hand, certain plans and programs such as those aimed at people with minimum incomes, dependent people, prostitutes, unemployed people, training for certain collectives, social reinsertion or subsidies for artistic creation, regardless of their cost, tend to be under greater evaluative pressure.

In any case, evaluation should not be a punishment, as having evaluations is good news. In theory, their constructive effects should be felt in the form of more sensitive and responsible actions by professionals and decision makers. According to this view, it is policies of a universal nature that suffer, because they do not benefit from the feedback provided by evaluation like the others do. However, despite the fact that in absolute terms, systematic evaluation is profitable, there are certain risks in the unequal distribution of evaluation which are discussed below.

Risks

An evaluation exercise must be based on the possibility that the assessment may be negative. Therefore, the more evaluated programs are more exposed to this type of judgment than programs which are not evaluated. Using Weiss's words "the evaluated program has all its linen, clean and dirty, hung out in public, the unanalyzed program can tuck its secrets

away in the bureau drawers" (1991, p. 221).

When negative results are *hung out* for the public to see it is easier for the next step to be restriction, denial of funds or closure of the program. There are many examples of this. In the European Commission, the Evaluation Unit in the Directorate General for Budget argues in favor of the usefulness of evaluation processes by pointing out that they served to argue in favor of closing down a school-breakfast program (School Milk Measure)[15]. In the USA, when Ronald Reagan "took office his agenda was to curtail government, and evaluators were asked to find inefficiencies" (House, 2006, p. 120). G.W Bush announced that he would cut down or eliminate over 150 government programs that were not getting results. As Bush warned, those programs that cannot be held to account for good performance will be reduced or eliminated (Renger, 2006).

This is not to speak against the closure or cutting down of programs which are not working, but to point out that a greater exposition to public analysis renders selective policies more vulnerable. Meanwhile, universal policies, which are less evaluated, stand apart from these dynamics without being questioned. Going back to examples from the USA, Monnier (1992) explains how evaluation was used by the Nixon administration to put off the execution of social measures imposed by Congress and cut down the number of innovative social programs.

In the face of these situations, it is not strange that the reactions to evaluation are avoidance and armor plating whenever the decision makers behind these interventions think that the outcome of evaluation will not be positive (Newcomer, 2004). The phenomenon detected by Chelimsky relates to this. Between 1980 and 1994 the author perceived an increase in "secrecy and classification" regarding information in a large number of agencies. "The irony is that the threat to validity posed by classification often has little to do with national security but rather with the dark side of agency independence (2008, p. 407).

Just like certain evaluation methods may have perverse effects (Leeuw & Furubo, 2008), the same is true of the unequal spread of evaluation, which can have a penalizing effect on certain policies.

Taking into account the current situation of evaluation in Spain, the risk is that selective public actions or those based on positive discrimination may be penalized as a result of being more exposed. The consolidation of this dynamic could contribute in the long run to shape a certain political conception which, supported by a technical discourse, would promote linear interventions in which everybody receives the same, regardless of the difficulties or social needs they may suffer.

Besides, the relationship between strategies (universal and selective) and political sectors, the fact that Justice, Infrastructures or Defense tend to act by means of universal strategies whereas Social Affairs, Labor, Immigration and Equality need to selectively address specific groups implies, indirectly, that the latter political sectors, associated with social welfare policies, suffer a detriment.

The effect of evaluation, as it stands, seems to contribute to a technocratic State

[15] Case quoted by Eduardo Zapico in his paper on the Evaluation in the European Commission in the VI Seminar on Evaluation Experiences. Master on Evaluation of Public Programmes and Policies of the Complutense University of Madrid, 28/09/09.

model. Strong in Justice, Defense, Interior and Infrastructures but weak and undercapitalized in the social field. In this kind of State, if intervention on a specific collective is considered technically indispensable, it will probably be undertaken in a limited, restricted and "scrutinized" fashion with the help of various mechanisms including evaluation.

Concluding Remarks

Although the final conclusions of the article suggest that it may be possible to generalize them to other realities, it is necessary to support this hypothesis with specific data in order to find out whether this is also true of other countries. There are indications that seem to point in this direction. The studies by Leeuw et al, Nioche and Poinsard and the Request for Proposal show at first sight similar trends to those found in the Spanish case.

The first recommendation is an invitation to researchers to analyze samples, evaluation data bases, registers or evaluation requests in different countries in order to find out whether similar processes are taking place.

The second recommendation implies investing some efforts in correcting the current uneven application of evaluation. The identification of some perverse effect should not result in less evaluations being requested. On the other hand, it is necessary to increase the domain of evaluation so that all policies, both universal and selective undergo the same feedback and learning process.

Establishing a monitoring system for the spread of evaluation in the different political sectors. The monitoring system can be based on periodic sampling of evaluations taking the different administrative units as population for research on evaluation contracts offered by the different administrations or in bibliometric studies according to the suggestion by Leeuw et al mentioned above. This monitoring function could be taken up by general governmental units or bodies in charge of evaluation policies. In the case of Spain, a clear role might be played by the National Agency for Evaluation (AEVAL) or similar regional government bodies.

The monitoring model may raise alarm if there are sectors which are rarely evaluated. General bodies, like the AEVAL in Spain may lead evaluation processes in such sectors or encourage them by calling the attention of the actors in charge of evaluation in each political sector.

The choice of what is to be evaluated must be made by making the decision criteria explicit. This motivation can be included in the commission or the terms of reference.

In order to avoid some of the perverse effects, the commission can establish certain recommendations in the evaluation contract. The following aspects might be recommended in their terms of reference:

- Evaluations must contemplate an analysis of the context.
- The results must be examined in relation to other evaluations in order to prevent them from leading to negative judgment without evidence from the theoretical and practical context.
- A careful and, as Weiss advises (1991), critical view point must be applied to the standards upon which the judgment rests. If evaluation is carried out according to models with standards, the choice of these standards must be justified. On the other hand, the

quality of a certain intervention can not rest solely upon its outcomes (Greene, 1999). In order to contextualize and understand the results it must also be possible to obtain information about processes and structural elements.

In countries where the evaluation is in an emerging phase, like in Spain, it is particularly important that evaluation is not seen as an enemy but, in association with a constructive methodology, as help rather than control. There might be two recommendations in this sense:

1. Rewarding organizations which are capable of engaging in the transparency implied by evaluation and of applying the resulting recommendations regardless of whether the results are positive or negative. This is particularly important for NGOs or organizations which depend on subsidies. Organizations which commission evaluations such as the Spanish Agency for International Development Cooperation (AECID) can include in their commissions the commitment not to penalize organizations which open themselves up to evaluation.
2. Making subsidies or donations conditional upon evaluations being completed. In the case of organizations which are recurrently not evaluated, the granting of subsidies or donations must be put to question. This measure is applicable to all intervention sectors, but particularly to those which have not developed evaluation as much, such as Defense, Justice, Infrastructures, Home Affairs, Housing or Economy.

References

Ballart, X. (1992). *¿Cómo evaluar programas y servicios públicos?*. Madrid: MAP.

Bustelo, M. (2002). ¿Qué tiene de específico la metodología de evaluación? In Bañon, R. (Ed), *La evaluación de la acción de las políticas públicas* (pp.13-32). Madrid: Díaz de Santos.

Bustelo, M. (2004). *La evaluación de las políticas de género en España*. Madrid: Catarata.

Bustelo, M. (2006). The potential role of standards and guidelines in the development of an evaluation culture in Spain. *Evaluation, 12*, 437-453.

Bustelo, M, Díez, M. A, Izquierdo, B., & Ligero, J. A. (2006). *Building capacity for evaluation. The case of Spain*. Paper for European Commission.

Cabrera, P. (1998). *Huéspedes del aire. Sociología de las personas sin hogar en Madrid*. Madrid: UPCO.

Chelimsky, E. (2008). A clash of cultures. Improving the "fit" between evaluative independence and the political requireiments of a Democratic Society. *American Journal of Evaluation, 29*, 400-415.

Christie, C. A. (2007). Reported influence of evaluation data on decision makers' actions. *American Journal of Evaluation*, *28*, 8-25.

Comisión para el Estudio y Creación de la Agencia Estatal de Evaluación de la Calidad de los Servicios y de las Políticas Públicas. (2004). *Informe 4 de Octubre 2004*. Madrid: MAP.

Cohen, E., & Franco, R. (1993). *Evaluación de proyectos sociales.* Madrid: Siglo XXI.

Datta, L.-E., & Grasso, P. G. (1998). Editors' notes. *Evaluating tax expenditures: Tools and Techniques for Assessing outcomes. New Directions for Program Evaluation, 79*, 1-9.

Díaz-Puente, J. M., Cazorla, A., & Dorrego, A. (2007). Crossing national, continental, and linguistic boundaries: Toward a worldwide evaluation research community in journals of evaluation. *American Journal Evaluation, 28*, 399-415.

Díaz-Puente, J. M., Yaqüe, J. L., & Afonso, A. (2008). Building evaluation capacity in Spain: A case study of rural development and empowerment in the European Union. *Evaluation Review 32*, 478-506.

Fernández-Ramírez, B., & Rebolloso, E. (2006). Evaluation in Spain: Concepts, contexts, and networks. *Journal of Multidisciplinary Evaluation, 5*, 134-152.

Freeman, H. E., & Solomon, M. A. (1981). The next decade in evaluation research. In R.A. Levine, M.A. Solomon, G.-M. Hellstern and H. Wollaman (Eds.), *Evaluation research and practice. Comparative and International perspectives.* Berverly Hills, CA:Sage.

Furubo, J.-E., & Sandahl, R. (2002). A diffusion perspective on global developments in evaluation. In J.-E. Furubo, R.C. Rist and R. Sandahl (Eds). *International Atlas of Evaluation* (pp.1-23.) New Brunswick, NJ: Transaction Publishers.

García-Sánchez, E. (2005). *La evaluación de programas de reforma educativa en España. Tres estudios de caso desde un enfoque de metaevalaución.* Madrid: Editorial de Universidad Complutense.

Garde, J. A. (2006). *La evaluación de políticas públicas y su institucionalización en España.* Madrid: MAP.

Greene, J. C. (1999). The inequality of performance measurements. *Evaluation, 5*, 160-172.

Halusha, J. (2006, Novembre 9). Voters in 7 states back ban on gay marriage. *The International Herald Tribune,* p. 7.

Henry, G. T. (2002). How modern democracies are shaping evaluation and the emerging challenges for evaluation. *American Journal of Evaluation, 22*, 419-429.

Henry, G. T., & Mark, M. M. (2003). Beyond use: understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation, 24*, 293-314.

House, E. R. (2006). Democracy and evaluation. *Evaluation, 12*, 119-127.

Leeuw, F. L., & Furubo, J.-E. (2008). Evaluation systems. What are they and why study them? *Evaluation, 14*, 157-169.

Leeuw, F. L., Toulemonde, J., & Brouwers, A. (1999). Evaluation activities in Europe: a quick scan of the market in 1998. *Evaluation, 5*, 487-496.

Lipsey, M. W. (1995). What do we learn form 400 research studies on the effectiveness of treatment with juvenile delinquents? In J. McGuiere (ed.), *What works? Reducing Reoffending* (pp.63-78). New York: John Wiley.

Lipsey, M. W. (2000). Meta-analysis and the learning curve in evaluation practice. *American Journal of Evaluation, 21*, 207-212.

Monge, Y. (2007, June 29). Golpe en EEUU a la discriminación positiva. *El País*, Internacional.

Monnier, E. (1992). *Evaluación de la acción de los poderes públicos.* Madrid: Instituto de Estudios Fiscales.

Newcomer, K. E. (2004). How might we strengthen evaluation capacity to manage evaluation contracts? *American Journal of Evaluation, 25*, 209-218.

Nioche, J. P. et al. (1984). *L'Evaluation des Politiques Publiques.* Paris: Economica.

Pazos, M., & Zapico, E. (2002). Program evaluation in Spain: taking off at the edge of the twenty-first century? In J.-E. Furubo, R.C, Rist and R. Sandahl (Eds.), *International atlas of evaluation* (pp. 291-306). New Brunswick, NJ: Transaction Publishers.

Renger, R. (2006). Consequences to federal programs when the logic-modeling process is not followed with fidelity. *American Journal of Evaluation, 27*, 452-463.

Rutter, M., Giller, H., & Hagell, A. (2000). *La conducta antisocial de los jóvenes*. Madrid: Cambridge University Press.

Shadish, W. R., Chacón-Moscoso, S., & Sánchez-Meca, J. (2005). Evidence-based decision making: Enhancing systematic reviews of program evaluation results in Europe. *Evaluation, 11*, 95-109.

Stake, R. E. (2006). *Evaluación Comprensiva y evaluación basada en estándares*. Barcelona: Grao.

Stame, N. (2008). The European Project, federalism and evaluation. *Evaluation, 14*, 117-140.

Stufflebeam, D. L., & Shinkfield, A. (1995). *Evaluación sistemática. Guía teórica y práctica*. Barcelona: Paidós.

Thuren, B. M. (1993). *El poder generizado.* Madrid*:* Instituto de Investigaciones Feministas.

Tobal, C. (1982). *Guía para la formulación y evaluación nacional de proyectos de desarrollo rural integrado*. Washington, DC: OEA.

Varone, F., Jacob, S., & De Winter, L. (2005). Polity, politics and policy evaluation in Belgium. *Evaluation, 11*, 253-273.

Weiss, C. H. (1991). Evaluation research in the political context: sixteen years and four administrations later. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century* (pp. 211-231). Chicago: The National Society for the Study of Education.

Weiss, C. H. (1998). *Evaluation*. New Jersey: Prentice-Hall.