

CSS Commission on the “Evaluation of the Effects of Public Policies”

INTERNATIONAL WORKSHOP ON IMPACT EVALUATION
PRACTICE AND PROSPECTS

April 13-14, 2007
Verduno, Italy

A compilation of the experts' answers to the questions posed

A. WHO DOES REQUEST IMPACT EVALUATIONS AND WHY?

A.1. How often does motivation for requesting an impact evaluation stem from a genuine desire to improve a policy already in operation, by identifying “what works” (and what does not)?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
My impression, at least for France, is that this motivation corresponds to a minority opinion, not only in Ministries but also in public agencies. Promoters of programs (politicians and bureaucrats) are generally reluctant to request scientific evaluations, because they think that the results from evaluation studies could be interpreted as a criticism against their policy.	Main motivation / very often.	The general assignment for the IFAU is to evaluate labour market policy and (the labour market effects) of education policy. We have substantial discretion in choosing what policy measure should be evaluated. Fairly rarely do we get specific government assignments. If we do get such assignments, they are usually “follow-ups” (i.e. descriptions of what has happened) rather evaluations (i.e. an attempt to estimate the causal effects of participation.). In the Swedish case, it is therefore quite hard to answer the questions here, since one has to speculate as to why the authorities decided to give us the general assignment to start off with. I think you can rule out symbolic motivations. Presumably a desire for knowing what works (and what does not) is more accurate. I think policy makers have a fair understanding of what a “causal effect” is. But being labour economists we focus on employment outcomes. And if the results in the report do not fit their priors, they can defend the policy in terms of other “causal” effects which they think are relevant (e.g. redistribution). Apart from this general statement, I will only answer questions 2 and 5 below.	It is perhaps best not to be too cynical about calls for impact evaluations of a programme. Mostly those are motivated by a desire to see if the programme “works” and if it can be justified on value-for-money grounds. On a more cynical note, the following anecdote may be revealing. In the late 1980s, I asked a Deputy Secretary-General (who shall be nameless) from the Labour Ministry of an EU country why his Ministry funded no meaningful evaluations. If I can paraphrase his explanation, it would run as follows: 'Most of our programmes are lousy! They were dreamed up quickly to give the Minister some good news to announce at a time when unemployment is rising. We do not want evaluations revealing to the general public how bad our programmes are; we know this already!'. I should add that the civil servant in question has subsequently retired and his Ministry has timidly dipped its toes in the water in terms of funding a few evaluations. But it still remains rather unenthusiastic about funding rigorous evaluations.	"Improving policies and program" is a big motivator at programmatic level. Ministries and departments like the idea of using IE as a management tool for testing alternatives, choosing the best, and improve program effectiveness over time. E.g. South Africa	Comparatively rarely. Impact evaluations are quite difficult to undertake retrospectively due to the problem of defining a counterfactual especially if the policy is implemented nationwide	In the early days of social experimentation, studies focused on testing new bold ideas—modifications in the nations public assistance policies; strategies for reintegrating former criminals and drug addicts into mainstream society; or major changes in health insurance policies, for example. However, today social experiments are conducted for a wider range of purpose, including in many cases to provide policy makers and practitioners with guidance on the effectiveness of current practice and ideas for program and policy improvement. This change in strategy began in the late 1980s as we federal legislation enabled states waivers of federal employment and welfare policies, for example, in exchange for commitments to engage in rigorous study of the impacts of the state’s alternative strategies as contrasted with federal policy as usual. More recently, the federal government, and in some cases state governments, have been mandating tests of new program and policy initiatives as part of phased implementation plans. Prime examples are recent programmatic initiatives to strengthen families and reduce single parent child-rearing and their efforts to devise stronger programs and policies to reduce teenage childbearing and sexual health risks among teens and unmarried adults. In both cases, the federal government has supported major program development initiatives side by side experimental design program evaluation efforts aimed at determining the effectiveness of the model programs and developing guidance for program and policy improvement. Then, most recently, we have seen an explosion of efforts to use intervention research to guide the selection of programs and practices that are effective in the education—including research on particular curricula, pedagogical practices, professional development models, and supplemental services.

A. WHO DOES REQUEST IMPACT EVALUATIONS AND WHY?

**A.2. How often is evaluation used to test in advance the effectiveness of a policy by implementing it as a pilot test?
How common is the use of what in the US are called demonstrations?**

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
In France, evaluation has never been used as a pre-test of the effectiveness of a policy, even a recent law of the Parliament authorizes the implementation of experiments at a local level to identify "what works" (and what does not). This has something to do with the general reluctance to use randomised experiments for evaluating public interventions.	This is very uncommon in Germany. There are some "pilot studies" in "pilot areas", but this is really not the general case	Demonstrations/pilots are uncommon. But on some occasions, there have been some attempts to run a pilot. Policy makers have not waited for the results of the pilots. Rather they have rolled-out the program nationally before the results were known. Nevertheless, even if the programs were rolled out nationally, the pilots gives a window-of-opportunity for evaluators. A problem with the few pilots that have been run is that the implementation was not rigorous	The use of so-called "pilots" or "demonstrations" is essentially confined to the U.S. and Canada; more recently, it has been taken up in the U.K. (see, for example, the Pathways-to-Work pilots for recipients of Invalidation Benefits).	Pilot tests are fairly common, but often with limited external validity. We are trying to move pilot testing on national or program representative samples to ensure external validity.	Piloting has been the preferred mode of testing policies in the UK since 1997. Only one large-scale demonstration project has been implemented although a small number of random assignment experiments have been conducted. It is worth distinguishing between prototypes, where the decision to implement a policy (nationwide) and the research is to fine-tune the implementation from pilots which ostensibly are designed to determine whether the intervention being tested deserves to be implemented (nationwide)	Social experimentation in the U.S. began as "pilot tests" of innovative strategies and subsequently branched out to encompass evaluation of ongoing programs. Federally supported impact evaluations in the U.S. represent a mix of demonstration initiatives and program evaluations. Examples of recent and ongoing evaluations of ongoing program initiatives include evaluations of the following: Job Corps, AmeriCorps, and other alternative education, training, and work experience programs for disadvantaged youth; Head Start; 21 st Century Learning Center After School Programs; Teach for America and other alternative certification teaching certification programs; Head Start, Early Head Start, and various state funded pre-school educational initiatives; and evaluation state Title V) and federally (SPRANS) administered health and sex education programs for adolescents and young adults. There also are major efforts underway to evaluate truly innovative programs. Examples include federally funded marriage promotion and support program; strategies for effective use of technology in K-12 education; strategies for improving development of early reading skills among children from disadvantaged and non-English speaking backgrounds; strategies for promoting stronger character development among middle and high school youth; innovative approaches to reducing school dropout among high school age youth; and innovative approaches for smoothing the transition of youth from foster care into independent living as they age out of foster care. In the past 10 years, there has been a notable increase in the use by academic researchers of social experimentation to pilot test innovative strategies to address social, health, and education issues. This includes tests of family and individual therapeutic approaches to address behavioral issues among children and youth; of various approaches to teaching and learning among children of all ages; of mentoring, tutoring and various other strategies to improve social and educational outcomes of children and youth; and of new ways to mediate issues related to juvenile delinquency.

A. WHO DOES REQUEST IMPACT EVALUATIONS AND WHY?

A.3. How often does the motivation for evaluation stem from a more generic call for accountability for results? In this case, is the notion of “impact” used by policy-makers still equivalent to that of “causal effect”, or does it entail a broader (possibly vague) meaning?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>This motivation is more often invoked. But here the notion of impact is generally considered from a more global perspective, that corresponds (more or less) to a macroeconomic cost-benefit analysis. In this perspective, the microeconomic point of view, with its specific concepts (such as the counterfactual situation, the average effect of the treatment on the treated, etc.), is generally ignored.</p>	<p>The discussion centres around gross and net effects. Some politicians/policy-makers still look at gross effects of policies, e.g., 70% of the participants in a VT found a job afterwards. However, the concept of net effects and impact evaluation becomes more and more widespread</p>			<p>Accountability for results motivates treasuries but is a poor bargain for implementing agencies that need greater operational guidance (the how to of improving programs)</p>	<p>The meaning of this question is not transparent. Many UK evaluations are part of a process of negotiation with Treasury to elicit funds to put a policy in place. Evidence that the policy will work (and/or is cost effective) would form part of the case for funding. There is quite often a presumption that the policy will work that may well be based on prior ideology. The evaluation may cause policy makers to specify the 'theory of change' more clearly/explicitly</p>	<p>Increasingly, and perhaps almost universally, the call for impact evaluations in the U.S. tend to be calls for comprehensive studies in which evidence regarding the causal impacts of the intervention constitute and important, but by no means the only type of evidence demanded. It is standard for social experiments in the U.S. to include implementation and operational analyses, as well as impact assessments. Studies are generally designed to measure impacts for important subpopulations and/or in different settings and to relate the observed pattern of results to knowledge about program design, implementation, and operations.</p>

A. WHO DOES REQUEST IMPACT EVALUATIONS AND WHY?

A.4. How often the motivation is of a symbolic nature, i.e. the call for evaluation is a way for policy-makers to reaffirm their commitment to the underlying policy issue, although they have no strong interest in the actual evaluation results?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>This symbolic motivation seems to be often present. An example is the evaluation of the effects on the employment level of small-size firms of the introduction in August 2005, in France, of a new labour contract called the Contrat Nouvelle Embauche (CNE hereafter). The CNE contract is a long-term labor contract more flexible than the regular one, called contrat à durée indéterminée (CDI hereafter). A few months after the introduction of this new contract, the government asks a private survey institute to conduct an “evaluation” study. This institute asked 300 small-size firms whether they used this contract to hire new workers; approximately 30% responded “yes” and said that this new contract helped them to increase their employment level. The government used this result to affirm that this new contract was effective. Now, more scientific evaluation are conducted (I am in charge of one of these evaluation studies with two colleagues from INSEE), and their first results (to be confirmed) show that this new contract has a very small (even null) effect on the employment level of small-size firms</p>	<p>- After the reform of the legal basis for labour market policies in 1998 (see above) and especially the Hartz-reforms, output evaluation was made mandatory for all spending on ALMP. Hence, I would not say that evaluations are symbolic. Clearly, it is a different question whether politicians rely on the evaluation results when (re)designing policy</p>			<p>Political motivation is strong for communication to the public and answering criticism from the opposition. Election time.</p>	<p>Certainly there is a tendency for a pilot to serve as a statement that the government is interested in an issue and is doing something about it (without spending a large sum). However, evaluations tend to develop a life of their own. They may initially be symbolic but impending results tends to focus policy makers interests and concern. Early positive findings may well be used (illicitly) to justify full-scale implementation. Early negative findings may provoke a rethink, a change in the pilot strategy or the preparation of a defensive strategy in case the final results are also negative</p>	<p>There certainly are times when evaluations are commissioned for political reasons—they are sometimes the price for bi-partisan support for a policy or program. This was the case, for example, in some of our recent health and sex education policy initiatives and this was the motivation for recent evaluation of a major federal policy initiative to provide federal support for after school care for elementary and middle school children. However, more typically, I would say studies are motivated by interest in finding ways to improve existing policies—to find more effective ways to address social concerns.</p>

A. WHO DOES REQUEST IMPACT EVALUATIONS AND WHY?

A.5. Whatever the motivation, how common is to include an impact evaluation in the design of a programme, or at least to plan the evaluation before the programme is implemented?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>All examples of evaluation that I have in mind for France are ex post studies. The inclusion of an impact evaluation in the <i>design</i> of a programme, or at least to plan the evaluation <i>before</i> the programme is implemented, is <i>definitively</i> not a common practice in France</p>	<p>This depends very much on the policy. There is a recent debate about the in-troduction of UK/US-style in-work benefits like the WFTC or the EITC and there have been a good amount of ex-ante simulation studies to estimate the effects of different designs. For other policies, there are less concerns about possible effects and planning evaluations in advance. One major step, how-ever, was the fact, that administrative records have been made available for scientific research and that some resources have been spent to improve data quality.</p>	<p>This is as uncommon as running a pilot. However, my personal opinion is that this is one of the most fruitful avenues for getting good (better) evaluations. Politicians like to make things happen, i.e., they are not so keen on waiting for the results from a pilot. It is more likely that they will agree to implementing the policy such that it is possible to evaluate the new measure. For instance, a step-wise implementation across the country gives some scope for credible evaluation if treatment and control regions are chosen with some care. A further argument for step-wise implementation is that it gives the authorities in charge of implementing the policy some scope for fine-tuning implementation until the policy is run nation-wide. Whenever I encounter policy-makers, I try to convince them of the usefulness of this approach.</p>	<p>It is still quite rare for policymakers to include an impact evaluation in the design of a programme. This would be the ideal practice, but the typical case in many countries is to commission an evaluation ex-post, usually after the programme is up and running for sometime. This, of course, makes the evaluation often more difficult and costly to undertake and the reliability of the results less certain.</p>	<p>In the case of World Bank projects, we are challenging the standard ex post evaluation practice and replacing it with ex-ante design to ensure rigor of the evaluation and contribute to program design (by challenging notions). In country we use policy transitions and programmatic changes as an entry point.</p>	<p>As already noted, this has become the norm since 1997 across a number of UK departments</p>	<p>It is quite common to include evaluations in the design of programs of high national significance for which there is considerable disagreement about the likely effectiveness of a particular approach. So, for example, there was an enormous amount of social experimentation with different approaches to promoting economic self-sufficiency through employment during periods of rising welfare roles and stagnant child poverty rates in the face of rising public spending on welfare-to-work programs for poor, single parents. It was clear that the standard policy was not achieving the intended results. Thus, the government began inviting state level experimentation under conditions of rigorous evaluation. We are seeing similar trends beginning to emerge in the education.</p>

A. WHO DOES REQUEST IMPACT EVALUATIONS AND WHY?

A.6. How crucial is the presence of “checks and balances” between different branches of government in determining the demand for evaluation? How often are impact evaluations demanded directly by the legislative branch exercising its oversight on the executive branch?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
Over the recent years, the French Senate and Parliament have examined the effects of the French reform on the reduction of the weekly working time and on the reform relative to the change in the system and the time profile of unemployment insurance benefits. For that purpose, they have auditioned researchers having conducted impact evaluations of such reforms, and then they synthesised their judgement in public reports	The Hartz-evaluations were demanded by the legislative branch (Bundestag), but very often it is different branches of government demanding evaluations. It might also be the case that different branches (say the Ministry of Finance and the Ministry of Labour) ask different institutions to evaluate the same policy.	IFAU's general assignment comes from the government rather than the parliament. There are a few evaluation authorities in Sweden which have it in the same way. The initiative for specific assignments (when independent researchers bid for a contract) also comes from the government. Most often it is the Ministry of Finance which is demanding an evaluation. There is one exception from this general rule; this exception is the Swedish National Audit Office (“Riksrevisionen”) whose general assignment comes from the Parliament rather than (branches of) the government.	With regard to the key institutional actors in demanding evaluations, I would argue that the US Congress sets the pace here with respect to the evaluation of labour market/social programmes. It tends to systematically demand such evaluations and the Administration generally complies with such demands. Unfortunately, the legislatures in other OECD countries, or rather the Employment and Social Policy Committees in National Parliaments, are much less demanding. In some countries, an alternative source of pressure in favour of evaluations can sometimes be found in Finance ministries. They have a natural tendency to worry about value-for-money in terms of public spending and can bring this to bear on the spending ministries.	In African countries, the legislative is often weak and unable to put forth these demands. Budget negotiations within the executive is where IE results are starting to be used.	The legislative branch has no (or a very limited role) in this area. The National Audit Office is charged by Parliament to monitoring value for money of central government spending but rarely has the resources (or inclination) to conduct impact evaluations. The same is true of the Audit Commission which has a similar irresponsibility for local government expenditure. The checks and balances apply, as already noted, in negotiations with Treasury. In the past this might have entailed post-facto evaluations; more often now it involves prospective evaluations (pilots)	There certainly are checks and balances for some of the more high-profile studies—for example, those linked to federal funding streams. However, the standards for rigorous research in the U.S. are such that the checks and balances we see have their greatest impact on the packaging and release of the final results. In the end, publicly funded research must be made available, regardless of the findings. So, at best, the political process affects dissemination at the margin

A. WHO DOES REQUEST IMPACT EVALUATIONS AND WHY?

A.7. How often does the demand for evaluation stem from the interplay between different levels of government, where one level funds a policy or programme that is going to be implemented by another (usually lower) level?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
	Not often	This is the most common way (see above).	The EU Commission should be able to exert similar pressure via the EU Social Funds/Structural Funds. Indeed, it does demand evaluations of such spending. However, my impression is that the quality of many of these evaluations is very variable and there is very little follow-up, so it is doubtful whether the Commission is really as major a force for good evaluations as it could be or should be.		Money for special initiatives given by central government to local government may well include the requirement to monitor performance but this would rarely entail measuring the additivity of an policy intervention. Likewise, since 1997, government has been awash with performance targets but these involve monitoring rather than impact assessment.	Generally, most of the demand for rigorous impact evaluations comes from the federal government and from private foundations. However, there are instances where state governments get involved—a notable case being all of the welfare reform research in the 1980s and 1990s which tended to be co-sponsored by state and federal agencies (and not infrequently with some foundation backing).

A.8. Within the executive branch of government, the policy-makers requesting evaluations tend to be politicians (or political appointees) or rather high-ranking career officials? How do motivations and incentives differ between the two groups?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
It is difficult to give a precise answer, but in my opinion, the policy-makers requesting evaluations are more often high-ranking <i>career officials</i> , rather than <i>politicians</i> (or <i>political appointees</i>). The motivation of politicians is more strategic than scientific (or analytic), which explains that they are more reluctant to demand impact evaluations on the effects of policies they have introduced. In principle, high-ranking <i>career officials</i> are more disinterested	Once again, that depends on the type of the policy and the branch of government	I would suspect that the initiative comes from (high-ranking) civil servants rather than the politicians, even though the politicians must give the formal “go-ahead”.		Politicians are interested in point estimates that demonstrate success of their policies. Program managers/implementing agencies are more interested in testing ways to improve effectiveness.	Pilots will be commissioned by civil servants rather than politicians. The latter will subsequently be interested in the broad outcomes, civil servants in detail. (This might mean politicians are more interested in impact evaluations and policy makers in process evaluations. I deal with these issues in more systematically and in more detail in the papers that I have already sent to you	It is my sense that the requests tend to come from political appointees from within the executive branch, but more from the career staff within the legislative branch. Politicians are most actively involved in evaluation requests when they have reason to believe the outcome will support a particular political agenda.

A. WHO DOES REQUEST IMPACT EVALUATIONS AND WHY?

A.9. What is, on average, the awareness of the practical and conceptual difficulties in making causal inference among those policy-makers who specifically request impact evaluations?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
This awareness is very insufficient. French administrative and political elite has a very weak economic and methodological culture. The dominant culture comes from non-quantitative political science	It was very bad ten years ago, but is improving very much. The evaluation of the Hartz-reforms played a crucial role, since they have been also received a great deal of public (press) attention	In a comparison across countries, I would guess it is fairly well-developed in Sweden. It seems to me that the awareness is growing over time within the country.		Generally low to start. We develop awareness on a case by case basis through an approach that integrates learning with doing.	Variable. Partly there are distinctions between civil servants who are professionals, economists and social researchers, and those who are generalists. Economists might favor RCTs without thinking about the difficulty practicalities but are more familiar with econometric modelling, Social researchers often mistrust RCTs particularly if they lean towards interactionist sociology. Generalists are intelligent lay persons, who can recognise constraints when pointed out to them but are under pressure and constrained by the needs and constraints on their political masters.	There is high variability. However, most government agencies and political groups now have a reasonable degree of awareness, if not expertise, among them regarding what constitutes quality causal evidence.

B. WHAT DOES INFLUENCE THE DESIGN OF IMPACT EVALUATIONS?

B.1. For designing an impact evaluation, how critical is that the *objectives* of the policy or programme are clearly stated in advance? How critical is that they are expressed as *numerical targets* (e.g., x% reduction in the rate of...)?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>Stating clearly in advance the objectives of the policy or programme is crucial. Such a statement should orient the collection of the statistical information required for evaluating the intervention. It is also a way to obtain from the policy makers some sort of commitment on the anticipated effects of the programme (if the programme fails to produce these effects, the responsibility of the programme promoters is more clearly identified). Stating the objectives should also result in a more precise definition of the main expected outcomes of the programme (e.g. improving employability, increasing wages, reducing poverty or segregation, etc.). However, we should be careful: promoters of the programme (political parties, governments, public agencies, etc.) should not have in mind all the possible second-order or indirect effects of the programme (for instance, the school performance of children in poor households benefiting from a programme intended to reduce residential segregation). A preliminary consultation of experts should permit to include these indirect effects in the list of possible outcomes of the programme (and, consequently, in the list of variables to observe).</p>	<p>It is crucial that the objectives are stated in advance and that it is also clear, how these objectives can be operationalized, e.g., how to measure an individual "employability" (and not the actual employment chances). Expressing them in numerical targets is less common when talking about net effects.</p>	<p>The crucial aspect is that the targets are formulated in qualitative terms (e.g. the program should improve employment prospects).</p>	<p>The issues raised in these questions are crucial. It is vital that the policy objectives be clearly laid down in advance. It is much less critical to specify them in terms of numerical targets. Indeed, given the inherent uncertainty about the behavioural responses associated with most labour/social policy experiments, I would be very wary of any evaluation which sought to lay down precise numerical targets in advance. However, most of the key objectives should be capable of being quantified and those that can be should be separated from other, more qualitative objectives.</p>	<p>All our work starts with policy discussion, identifying "doubts" and capitalizing on those doubts to test viable alternatives. The evaluation will incorporate policy objectives but span beyond to incorporate evaluation of impact on other outcomes (desired/undesired; cross sector; etc)</p>	<p>Evaluation is meaningless in the absence of stated and measurable objectives. Assessment against normative exogenous objectives is a different exercise often pursued by academics. Evaluation forces, or at least force, policy makers clearly to state their objectives. UK policy makers are very familiar with numerical performance targets</p>	<p>It is important to understand the goals of the program or policy and how progress toward those goals could be measured, the intended target population for the policy or program, and central parameters related to the implementation of the policy or program. It is not necessary to stipulate numerical targets. However, it is important in the study design to have a sense of what size impact (on key outcomes) would be expected and what the smallest size impacts are that would be of practical significance for policy. The latter information is important for designing an efficient, adequately powered study.</p>

B. WHAT DOES INFLUENCE THE DESIGN OF IMPACT EVALUATIONS?

B.2. In defining the “treatment” whose effects are to be evaluated, how much attention is devoted to identify the treatment <i>actually received</i> by the beneficiaries, as opposed to the statutory treatment? Is <i>implementation analysis</i> used in identifying the treatment actually received?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>I am not sure to understand the question. But, if you have in mind something like the difference between the general definition and objectives of the programme (e.g., increasing the human capital of unskilled unemployed workers) and the practical way it is implemented for instance by local agencies (that may modify the volume of hours of training and thus to act on the quality of the training programme, because of different local resources), this is obviously an important issue. The programme is generally defined as an overall framework for public action, but its practical implementation often implies some local variability that has to be taken into account when evaluating its effectiveness. This variability is a potential source of identification for the statistician and omitting it could result in biased, or at least imprecise, estimates of the effects of the programme. Estimating that the “treatment” has no effect could be due to a composition effect if it is implemented with low resources in most cases. Trying to estimate heterogeneous effects (due to a variability of the treatment intensity) is then crucial.</p>	<p>The kind of treatment received is usually quite detailed recorded in the administrative data (up to 1,000 options to specify for the local caseworker). Implementation analysis is usually more often used to examine qualitative effects and the interplay between the different actors.</p>	<p>In every evaluation produced by the IFAU, we do are best to identify the treatment actually received (sometimes this is not possible despite the fact that we do our best).</p>	<p>It would be ideal when designing the programme to consider alternative ways of delivering the services/ incentives in question so that the evaluation could highlight the relative efficiency and cost-effectiveness of the alternatives. This would enable policy makers to gain valuable insights on the how part of the evaluation question. A recent example of such an implementation evaluation is the one which was undertaken recently of alternative approaches for monitoring the job-search behaviour of JSR clients in the UK. You put your finger on an important point when drawing a distinction between the actual treatment given to programme participants, as opposed to the statutory treatment laid down in the legislation/programme design. It often turns out in practice that there is a divergence between the former and the latter, and this divergence is often very important in understanding why the programme outcomes sometimes fail to live up to expectations.</p>	<p>This is a risk. Mitigation includes i) working closely with implementing agency; ii) using monitoring systems and iii) placing a staff in the field to monitor actual implementation, and help guard against contamination throughout the evaluation cycle</p>	<p>Treatment actually received is vitally important although the distinction is often the focus of accompanying process evaluations rather than being disentangled in the core impact evaluations. Most big evaluations are multi-method combined impact and process evaluations in recognition of the importance of implementation issues</p>	<p>It is very important to measure the nature of the treatment and to be able to contrast this with the counterfactual condition against which the treatment is being compared. This can be easy or hard, depending on the nature of the policy of program being studied. An important aspect of understanding the nature of the treatment is to ensure that there is comparability in the measurement of the intervention. Often times, the intervention is simply an “add-on” to existing policies or practices and not expected to alter the status quo. However, in other cases, the focal policy or practice could supplant or complement the intended comparison condition. The evaluation design should include strategies for being able to fairly compare and contrast the intervention and control conditions in ways that relate to the program and policy goals.</p>

B. WHAT DOES INFLUENCE THE DESIGN OF IMPACT EVALUATIONS?

B.3. In general, how often is an implementation analysis conducted <i>in conjunction</i> with an impact evaluation? In which sense can the two be “integrated”, given the different methods by which they have to be conducted and the different units of analysis? Could you provide examples of such an integrated approach?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
I am not an expert in the methodology of implementation analysis, but I think that it should be conducted <i>as often as possible</i> in conjunction with an impact evaluation. As I said before (see my response to question B.2), an implementation analysis should help to identify more precisely the effects of a heterogeneous treatment. In some cases, it should be conducted <i>before</i> the impact analysis, since the existence of such a heterogeneity is not clearly predictable. An implementation analysis helps then to construct the list of variables that affect the intensity of the treatment (for instance, the number of classroom hours, the number of trainers per trainee, etc.) and that should be precisely observed.	As said above, in the evaluation of the Hartz-reforms it was mandatory to conduct implementation and impact analysis. However, the degree to which the two approaches were really interlinked differed. In an ideal setting, implementation analysis can be used as a preliminary step for impact analysis. For example, identify the factors which influence the participation decision in qualitative interviews (with potential participants, caseworkers, and providers) and use this information to model the selection process in the impact evaluation.	The most common approach at the IFAU is probably to have the implementation analysis and the actual evaluation in a sequence of reports. One reason for choosing this approach is that the implementation analysis can be conducted in, more-or-less, real-time. While for the evaluation you have to wait for outcomes to be observed. Since policy-makers usually want quick answers, it is a good idea to release the implementation analysis earlier than the full evaluation. Two examples are: 1. The evaluation of employment subsidies. (There were two reports – only available in Swedish – on the implementation (e.g. “Forskningsrapport” 2001:9): The evaluation is published as IFAU WP 2004:18); 2. Subsidized career breaks (The implementation analysis – only in Swedish, although with an English summary – as “Rapport” 2004:6. The evaluations are published as IFAU WP 2005:17 & IFAU WP 2005:18)			Combined evaluations are the norm. Examples are included in the papers already sent to you. As I indicate in those papers there is a major challenge in integrating both elements constructively. Some of the problems are practical - for example time scale and sequencing. Others reflect different mind-sets although many evaluators in the UK are at least open to different approaches	In most major impact evaluations supported through the federal government and in an increasing proportion of those supported through foundations and other private sources, implementation analysis is integrated into the overall evaluation plan. In order to integrate the impact and implementation evaluations, it is important that the evaluation teams are highly overlapping—the impact evaluators need to understand (if not have experienced) the context for and nature of the program/policy implementation and the implementation researchers need to be fully aware of impact evaluation questions and findings as they emerge. Implementation research is most valuable when integrated with the impact findings. The research on programs designed to improve the life courses of teenage mothers and their children is a really good example of the power of well integrated study designs. http://aspe.hhs.gov/hsp/isp/tpd/index.htm ; http://www.mdrc.org/publications/145/execsum.html Another example, is the study of the Upward Bound college access program. This program was found in an early study not to be meeting its goals. As a result, the U.S. Department of Education used the findings from the comprehensive evaluation (impact and implementation and process study) to reform the program in ways that seem promising. They are now testing the effectiveness of the updated intervention strategy. Here is the link to the original study: http://www.mathematica-mpr.com/education/upbound.asp . The evaluation of the current program model is being conducted by Abt Associates: http://ies.ed.gov/ncee/projects/evaluation/upward.asp

B. WHAT DOES INFLUENCE THE DESIGN OF IMPACT EVALUATIONS?

B.4. How much attention is devoted to <i>understand the causal mechanisms</i> that generate the effects of the policy vs. to <i>estimate of the direction and size of the causal effects</i>? Which methods can be used to understand such mechanisms?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>In general, an insufficient attention is devoted to understand the causal mechanisms that generate the effects of the policy. Let us consider the example of job-search assistance programmes. In this context, the “structural” interpretation of a positive effect of the programme may be difficult: is this effect resulting from a change in the worker’s search effort and effectiveness, or is it the effect of a change in the arrival rate of job vacancies offered to the trainee, even if her search intensity is unchanged? A more precise observation should help to understand such mechanisms</p>	<p>This is still an under-researched area; a lot of space for (also scientific) improvement.</p>	<p>The primary focus is on estimating the causal effect of the policy. Of course, we are also interested in identifying the causal mechanism. In general I can think of four ways to understand the causal mechanism. 1. To set up a (theoretical) model. From the model you can in principle derive the “moments” you think you are estimating (e.g. a differences-in-differences parameter). In general this moment will reflect a combination of structural parameters. Prior knowledge can be used to assign plausible values for some of the structural parameters. The estimated moment(s) can be used to pin-down remaining parameter(s). 2. Implementation analysis can be combined with impact evaluation s.t. you get an idea of what the policy did to the participants. 3. Surveys (i.e. asking more detailed questions) can be combined with impact evaluation to pin-point plausible causal mechanisms. 4. One can examine different sets of outcomes. If you e.g. estimate the impact of the policy on employment probabilities, wages and annual income you can say useful things about the causal mechanism. For instance, if labour market training has a positive impact on annual income but no effect on the outflow to employment and wages it means that the duration of the match increased</p>	<p></p>	<p>A lot of attention is going to the latter, but we are trying to move more in the direction of the former, mainly to address the limitations IE faces in generalizing results and therefore the need to better understand context-specific and other factors that contribute to the results. In the case of e.g. the program of evaluation of the Malaria Booster Program, we are discussing survey instrument design issues to carry out determinant analysis. Considerations for cross country analysis include harmonization of survey instruments and their coverage, as well as the number of cases in each “program of evaluation”. Within country, issues might include sub-regional representativeness of sample and number of regions covered.</p>	<p>See previous answers. A lot.</p>	<p>It is easier to estimate net impacts of interventions with a high degree of confidence than it is to document reliability the mechanisms through which effects occur. The state of art in program evaluation entails working from a program logic model to the design of an evaluation that will test with rigor the questions of causality. Then, depending on the resources and the core study design, the evaluation design team will develop a plan to incorporate examination of supplemental impact questions (for example, for subgroups or for intermediate outcomes/mechanisms) and implementation and process analysis questions that can help interpret the study findings in ways that smartly guide their application to policy and practice. In some cases, it is easy to incorporate tests of the mechanism through which impacts occur and, in other cases, this would require substantial additional data collection and analysis. For example, if home visitor services were expected to impact child well being through a sequence of intermediate accomplishments like increasing the mother’s receptivity to parenting education, then changing the way she interacts with the child, it would be necessary to add an expensive layer of data collection with the mothers and observations of parenting practices to rigorously document the mechanism through which child outcomes did or did not change. Studies sometimes do and other times do not have the resources to investigate with causal analysis (as opposed to descriptive, process analysis) the mechanisms of change. Some of the home visiting research in the U.S. includes fairly rich analysis of the causal mechanisms (for example, see the work of David Olds http://pediatrics.aappublications.org/cgi/content/full/114/6/1550). And, in other cases, the research has focused limited the impact analysis to the main outcomes of interest and examined the mechanisms more descriptively through implementation and process data (for example, see the results from a federally supported home visitor demonstration program focused on teenage parents http://repository.upenn.edu/dissertations/AAI9965502/).</p>

B. WHAT DOES INFLUENCE THE DESIGN OF IMPACT EVALUATIONS?

B.5. How much attention is devoted to the <i>participatory</i> aspects of evaluation? Is there an active involvement, in defining evaluation criteria and design, on the part of the beneficiaries of the policy, directly or through their advocacy organizations? Should these participatory concerns be seen as crucial for <i>impact</i> evaluation?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>At my best knowledge, most of the (statistical) evaluation studies that have been conducted in France ignore such aspects. This insufficient attention tends to create a popular feeling (sometimes amplified by press articles) against the “technocratic” aspect of evaluation studies. Such a feeling could be limited by promoting participation of potential beneficiaries in the design process of evaluations. This aspect is even more crucial in the case of social experiments, where the administrators and the evaluators of the programme should be sure of the consent of potential beneficiaries. However the main problem is to identify their advocacy organizations (associations, unions, etc.).</p>		<p>No involvement. In most cases, I don’t think it is called for either. It is not obvious that participants have the same objectives as society. Of course, there are exceptions. Suppose you want to introduce a new tool at the public employment service (PES). Now, PES officers must, at least, accept this new tool. Otherwise, the policy is bound to be a failure.</p>		<p>In my experience there is very little of this going on. The exception are mixed methods evaluations that use qualitative mechanisms to inform the design of the evaluation and quantitative survey instruments, e.g. Ethiopia women's productive activities</p>	<p>Participatory approaches are varied and contested. They are rarely used in public policy evaluations if defined as programme beneficiaries taking control of the goals and execution of the evaluation. If the term is taken to mean that the views of the various policy actors, including beneficiaries, are sought, then this is the norm.</p>	<p>Traditionally, there has not been a lot of engagement by the advocacy groups in the design of evaluation. However, more recently, we have seen more of this. For the most part, the involvement of such groups is limited to helping frame the important research questions and to helping anticipate the types of implementation and process data that will be important for understanding and explaining findings should they turn out particular ways. A wonderful example of this type of broad participation in evaluation design was the recent, highly controversial study of a new federal policy to fund abstinence education through state administered block grants. This policy was highly controversial, with extreme views on the left (against) and on the right (in support of) the program. Congress mandated that the U.S. Department of Health and Human Services commission a rigorous, experimental design evaluation of the programs. The evaluation contractor and the U.S. Department of Health and Human Services engaged in extensive consultation with various constituencies to understand the outcomes of interest to various groups and their expectations regarding the possible ways the programs might help or hurt young people. These consultations also served the purpose of involving these groups in reviewing the evaluation design to ensure that, in the end, all understood and agreed that, regardless of the study findings, they would represent a fair test of the policy. Some background on this evaluation can be found at: http://www.mathematica-mpr.com/publications/PDFs/firstyearabstinence.pdf and http://www.mathematica-mpr.com/publications/PDFs/evalabstinence.pdf. There are other less dramatic examples where the engagement of the constituent groups in the evaluation design process was more focused on securing their buy-in for the randomization and data collection. One such example is Minority Female Single Parent Demonstration, where it was critical to convince program providers to over recruit prospective participants in order that excess demand could be assessed and that a true control group could be established. Information on this project can be found in Gordon and Burghardt: “The Minority Female Single Parent Demonstration: Short-Term Economic Impacts. A Technical Research Report. Into the Working World Series. Lessons from Research.”</p>

B. WHAT DOES INFLUENCE THE DESIGN OF IMPACT EVALUATIONS?

B.6. How much reliance is there on the <i>experimental design</i> for impact evaluation? Even when not utilized for ethical or practical reasons, is the experiment still considered an ideal benchmark? How widespread is the use of non-experimental methods that try to mimic the logic of experiments as opposed to methods that are model-based?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>Social experiments are still very rare in France. Up to recently, the opposition to social experiments was very strong, partly for ethical reasons, but also because social experiments are subject to several drawbacks (imperfect compliance, possible substitutes to the treatment for persons who are randomised out, Hawthorne effects, etc.). Very recently, CREST-INSEE (Centre de Recherche en Economie et Statistique, INSEE, Paris) was chosen to conduct a randomised experiment that is designed for evaluating a job-search assistance programme offered by the Public Employment Service to unemployed workers. Non-experimental methods that try to mimic the logic of experiments are still very infrequent. One recent example is the collection by the statistical department of the Ministry of Education of data that will be used to evaluate a programme (called “Collèges Ambition-Réussite”) that will give more resources to junior high-school localized in disadvantaged urban areas. For this evaluation study, the high-schools composing the control group were chosen to be as comparable as possible (on the basis of a propensity score) to the “treated” high-schools.</p>	<p>Experiments have not been used very often in Germany (in analogy to pilot studies). But clearly, the idea of “statistical twins” is very appealing to policy-makers. Most evaluation studies within the Hartz-evaluations used some sort of matching. But this clearly also depends on the questions posed and the data available.</p>	<p>In Sweden, experimental designs are as uncommon as elsewhere in Europe. There are a few examples, which have mostly been run locally. There are also a few recent and upcoming monitoring experiments (in parental leave insurance, unemployment insurance, and sickness insurance), which have been initiated by the IFAU and a government commission. Among evaluators, I think the experimental design is still the ideal benchmark (although no panacea). Policy makers understand this, but are generally not willing to initiate experiments. Non-experimental methods (such as matching) are gaining ground and are probably more common than regression in program evaluation (Although there is also some scepticism, particularly in the research community).</p>	<p>With respect to experimental methods, it is the case that they are rarely used outside the US and Canada. The UK is a partial exception to this, but only recently. Most other OECD countries which engage in evaluations are unwilling to invest in experimental methods, essentially on ethical grounds. Another explanation is that they lack familiarity with this approach and do not have the research infrastructure in place to apply large-scale social experiments - they do not have the local equivalents of the MRDC and Mathematica in the US or the SRDC in Canada. Happily, there has been much investment in non-experimental methods as a substitute for experimental methods, and much greater understanding of their strengths and limitations. Consequently, there is a growing reliance in many European countries on the use of non-experimental methods. See, for example, recent evaluations in the Nordics, Netherlands, Switzerland and Germany.</p>	<p>Experimental design is becoming more and more common (example the IE program in South Africa, experimental and at scale). Exceptions are general eligibility programs with variable pick up rates where we are using encouragement/promotion strategies; programs that target all the population below a threshold, where we use RD; or large infrastructure projects where random assignment is not possible where we might use pipeline or matched comparisons. We still use experimental methods to test alternatives within these programs, e.g. Ethiopia rural electricity.</p>	<p>Research to answer this question directly is about to be published by the UK Home Office. Quasi experimental designs are commonplace; true experimental designs rare. The evaluation community is divided by discipline and policy area. My guess would be that the majority view is that experimentation is a waste of time and energy; epistemologically inappropriate; practically very difficult. But there are strong advocates</p>	<p>There is now widespread agreement among the U.S. evaluation and policy community that nonexperimental evaluation methods generally do not yield reliable estimates of program and policy impacts. However, there also is increased awareness of the conditions necessary for experimental-design evaluations to generate unbiased, consistent, and reliable estimates. Most federal agencies now insist on well-designed experimental design evaluations whenever possible. They look very carefully at the designs for non experimental evaluations. Propensity score methods for creating control groups has been widely discredited as a reliable alternative, as have various methods of statistical matching, and various instrumental variable methods. See for example, Michalopoulos et al. http://www.mitpressjournals.org/doi/abs/10.1162/003465304323023732 and a reading list on the topic at www.gse.upenn.edu/pimfer/docs/Readings_Propensity%20Scoring%20Methods%20for%20Impact%20Evaluations.pdf. The favored alternative to explore currently is the regression discontinuity design. However, credible regression discontinuity designs require that some very specific conditions be met in order for them to yield credible impact estimates. For example, this method is being applied in a study of an early reading intervention sponsored by the U.S. Department of Education. The linked presentation by Howard Bloom illustrates the types of issues that must be addressed to determine whether or not regression discontinuity design methods are likely to be credible in a particular circumstance http://www.gse.upenn.edu/pimfer/docs/Bloom%20RDD%20Talk%20for%20UPENN%2011-27-06.pdf. There is a sizeable empirically-based literature pointing out the limitations of non-experimental methods that entail various forms of regression-based modeling, including use of instrumental variables and statistical matching. A meta analysis of the findings from studies that have attempted to replicate experimental research findings, ex post, using non experimental evidence (Glazerman et al. 2002) raises major concerns about the merits of the myriad methods commonly used as “second-best” strategies (see http://www.mathematica-mpr.com/publications/PDFs/nonexperimentalreps.pdf) .</p>

B. WHAT DOES INFLUENCE THE DESIGN OF IMPACT EVALUATIONS?

B.7. When the experimental design is used, is the random assignment mainly done with respect to <i>sites/areas</i> or with respect to <i>individual units</i> (persons, households, firms)? How are ethical objections overcome in the two situations?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>In the experiment conducted by CREST-INSEE to evaluate the job-search assistance programme offered by the public employment service, the random assignment was done with respect to persons. No serious ethical objection was raised; which may be considered as the sign of an important change in opinions regarding social experiments (some workers' unions were strongly in favour of this experiment).</p>	<p>Not really an issue in Germany.</p>	<p>As noted above, experiments are uncommon in Sweden. If they are to become more common, the realistic option (in most cases) is site randomization, because such randomization is less subject to ethical concerns. Nevertheless, in the recent and upcoming monitoring experiments randomization was conducted at the individual level. To a great extent this was made possible by the nature of the policy investigated. Cheating on government benefits is viewed as unacceptable. Therefore, randomized monitoring (at the individual level) was palatable. In principle one can also conduct random allocation within a sub-set of the population to reduce ethical concerns. Consider the following strategy, which is applicable to a program where the # slots is lower than the # potential participants. Case-workers get to sort individuals in to three categories w.r.t. to the potential gains from a program. Say that individuals are sorted into "Gainers", "No-gainers", and "Don't know". Randomization is then conducted within the set of individuals where case-workers don't know whether the individuals would gain from the program. In such a situation, randomization is less arbitrary than any other conceivable allocation mechanism. No ethical concerns can be raised.</p>	<p></p>	<p>Randomization is politically viable when there are budgetary constraints (once reasons for targeting are exhausted, all targeted get fair chance to participate); roll out capacity constraints (later beneficiaries are the control for today's beneficiaries), or doubts regarding "best" alternative (random assignment of treatment 1 and treatment 2). The unit of experimentation is program dependent. Clearly randomization at individual/household level is less expensive and done whenever feasible.</p>	<p>The answer differs by policy area. Schools can easily be randomly allocated; labour markets less so on grounds of expense and practical definition. On ethics see (already sent to you): Walker, R., Hoggart, L. and Hamilton, G. with Blank, S. <i>Making Random Assignment Happen: Evidence from the UK Employment Retention and Advancement (ERA) demonstration</i>. London: Department for Work and Pensions, Research Report No. 330, ISBN 1 84123 981 X., 2006.</p>	<p>Increasingly, evaluators decide to use individual or group randomization based on the nature of the intervention. If an intervention is targeted at individuals and unlikely to have a "group" response, there is the possibility to conduct a meaningful evaluation using individual assignment to treatment or control condition. For example, this likely would apply to a situation where the study was addressing mental health conditions or special learning needs. On the other hand, when the program or policy to be studied is expected to have "contagion" effects or be affected by the setting, it is most appropriate to design the evaluation around the assumption that the "affected" groups will be randomized to condition. An example would be testing a new method of mathematics instruction or testing a new approach to case management with welfare clients. In the former case, for example, it is not likely that teachers can avoid using what they judge to be "best practices" with all of their students, regardless of their study condition. In the latter, it likely would raise ethical concerns if case managers applied different standards for services provided to clients with similar needs. Ethical considerations are similar in the two cases. One needs to consider whether there is the expectation (or even a reasonable probability) that harm will result to individuals as a result of their random assignment. Most often, the reason to do a study is because it is hoped (but not known) that a policy change or new program will improve outcomes for individuals. And, often times, the alternative to doing the study is to implement the new policy or practice whole-sale or not to implement it at all. In such a case, it is generally deemed ethical to randomize "eligible/appropriate" individuals or groups to what is hypothesized to be the better condition.</p>

B. WHAT DOES INFLUENCE THE DESIGN OF IMPACT EVALUATIONS?

B.8. When an experimental design is used, especially when random assignment is done with respect to individual units, what are the *operational procedures and practices* adopted to assure a sound implementation of the design? Are they reasonably successful in avoiding the potential threats to randomization coming from the behaviour of the service providers and of the treated and/or control units?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
	Not really an issue in Germany.	As noted, experiments are rare. The only case where I have some first-hand knowledge was conducted in cooperation with the Social Insurance Board (SIB). In this case, considerable efforts were undertaken in order to gain acceptance at the top level as well as the lower levels of this authority. A program operator was appointed at the SIB. A call centre was instituted to take care of queries from individuals who had been informed that they were going to be monitored. There was no interaction between program administrators and treated treated/controls so this was not an issue in this case. In yet another case randomization was based on whether individuals were born on even or odd dates. This has the advantage of making clear to everyone who was eligible for treatment. The drawback is that may give rise to compensating behaviour on the part of the service providers.		We find that implementing agencies that understand why they are randomly assigning treatment are less likely to contaminate results. The operational usefulness of the evaluation, the buy-in of the implementer and their ability to affect the design/priorities questions are critical issues.	See previously cited paper	The mechanics of conducting randomization are fairly well established. Many of the hypothetical threats to randomization can be addressed through education of the program staff and to adaptations of the randomization procedures to the needs of program operators. It is customary among the leading evaluation groups in the U.S. to tailor any experimental design evaluation to address local concerns. For example, there is rarely a hard and fast rule about the assignment ratio between the intervention and control condition. This is allowed to vary in ways that addresses the joint needs of the program (for full- but not over- enrollment) and the study to avoid assignments with probabilities or 1 or 0. Program/policy implementers also need to understand that in any experimental design evaluation the assignment status of an individual or group remains constant throughout the study period. This, any attempt to serve the control group would, if the intervention is truly effective, bias the study results against finding an effect. Once program staff members understand this, they tend to avoid offering the services to the control group. An intermediate control is to institute monitoring of program participation to identify “cross-overs.” And, the ultimate control is that the evaluation will report on “cross-overs” and conduct a sensitivity analysis to address possible bias resulting from it.

C. WHERE DO THE DATA FOR IMPACT EVALUATION COME FROM?

C.1. How often is the evaluation design to be <i>data-driven</i> and how often is the data collection to be <i>design-driven</i>?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>In France, evaluation studies are mainly data-driven. These studies make use of either survey or administrative individual data. These data have two types of sources: 1) surveys collected by INSEE (<i>Institut National de la Statistique et des Etudes Economiques</i>, Paris) and statistical departments of the main ministries (Education, Health, Labour, etc.), 2) administrative files produced by Ministries, public agencies and other institutions (fiscal authorities, Social Security, public employment service, etc.)</p>	<p>With the introduction of the Social Code III in 1998 the awareness about data requirements to conduct thorough evaluation analyses has been increased. Following that, administrative records not only have been made available for scientific research but also the quality of these data has been improved sub-stantially. Hence, there is now quite a good basis for evaluation research. However, depending on the research question, additional data/surveys are necessary. Two examples: 1) Administrative records only cover employment which is bound to social security contributions. Hence, if you are interested in self-employment, additional data is needed. 2) When evaluating the effects of programmes for people with strong placement restrictions (e.g. due to long-term unemployment, drug use, etc.) it might be clear from the start, that the information in the administrative data is not sufficient. Hence, in the de-sign phase one could either choose estimators that take that into account or try to get these information while collecting the data.</p>	<p>In practically all cases, the evaluation design is data driven</p>	<p></p>	<p>For prospective evaluation data collection is design driven. Retrospective evaluations are data driven</p>	<p>You already have examples cited in my papers. Since 1997, a great deal of evaluations have been design-driven (although usually of a quasi experimental design). Post hoc evaluations of existing policy are inevitably data-driven</p>	<p>It is common in the U.S. for evaluation to be sensitive to and capitalize on readily available data. However, it is not common for major federally-sponsored evaluations to be overly constrained by administrative data. An example of a series of evaluations that were conducted using largely administrative data is the many welfare-to-work studies conducted in the late 1980s and early 1990s (see, for example, Gueron, J., Pauly, E., et al. (1991) <i>From Welfare to Work</i>, New York, NY: Russell Sage Foundation). More often, there is substantial data collection with the individuals participating the study sample, as well as collection of administrative data and observational data to assess the nature of the intervention and the control conditions. For a good discussion of program and implementation data collection see Werner, Alan (2004) <i>Implementation Research</i>, http://www.urban.org/expert.cfm?ID=AlanWerner.</p>

C. WHERE DO THE DATA FOR IMPACT EVALUATION COME FROM?

C.2. When no primary data collection is conducted, are the data obtained mainly from <i>administrative sources</i> or from existing <i>household or firm-level surveys</i>? What are the constraints on the access to these data imposed by confidentiality concerns? How are these concerns addressed when the data are to be used for evaluation/research purposes?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>There are no constraints on the access to data coming from household surveys or labour force surveys. The access to data coming from firm-level surveys or firm-level administrative files that are used for the evaluation of reforms intended to act on the firm labour demand, is restricted. The access to such data is easier for statisticians working as civil servants either at INSEE or in the Ministry of Labour, it is much more difficult for econometricians or statisticians who are members of academic research teams, because of confidentiality on firm information. This restriction concerning firm-level data is problematic since it limits the possibility of counter-evaluations.</p>	<p>In recent years data are mainly from administrative sources. Access constraints are quite high, evaluation results have to be of “general interest to the public” to get a permission. Data has to be stored on close-ended computer systems, etc.</p>	<p>In practically all cases the data come from administrative sources. There are no real constraints. The data used contain no personal id:s and researchers must sign an agreement on confidentiality where they agree not to reveal the identity of the individual. If they break this agreement they may be charged (i.e. it is a legal offence). Certain types of information are more “sensitive” than others. There are additional constraints when you want to use information on country of birth, health, and information on political preferences, for instance. In general, these constraints are surpassable for researchers.</p>	<p>One thing is clear from the recent evaluation literature. Countries that have very rich administrative data sources that can track individuals' labour market histories over time, that can be interfaced with tax/benefit records, and provide a lot of information on individual/family/household characteristics are extremely well-placed to implement state-of-the-art non-experimental methods to evaluate policies. This is possible in the Nordics and is becoming easier to compile/use in a few other countries.</p> <p>However, privacy considerations, lack of a unified statistical system and failure to give sufficient weight to evaluation/research considerations mean that many countries do not have such rich data sets available nor are they likely to have them in the near future.</p> <p>Household surveys are less useful for these evaluation purposes because they are typically not interacted with tax/benefit system records and they generally contain little or no information on participation in specific labour market programmes.</p>	<p>In education, we are using administrative data + test scores in South Africa. Generally administrative data systems are poor, however, and do not cover non-beneficiary population. Existing surveys have been used a lot in non-experimental evaluations. Confidentiality concerns are upheld but are not a constraint. Basic constraint is representation of beneficiaries in the nationally representative survey samples. Oversampling has been done successfully in the case of the labor survey in e.g. Argentina to look at the impact of the Trabajar program.</p>	<p>Legislation allows administrative data generally to be used for evaluation purposes. These data may be released for public analysis if anonymous and individuals cannot be interviewed. Outsiders contracted to undertake evaluations for government can be treated as if civil servants.</p>	<p>The standard in the United States is to obtain active consent from participants in any evaluation. When active consent from individuals is not obtained, the analysis generally is restricted to use of group-level data.</p>

C. WHERE DO THE DATA FOR IMPACT EVALUATION COME FROM?

C.3. How the data needed for impact evaluation fit into the *overall statistical information system*, particularly the component of the system based on *register data*? Are register data able to provide detailed background information on life histories of individuals and/or firms?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
Register data are not always able to provide such information (because they have not been collected for evaluation purposes). However, it has been sometimes possible to add some questions in frequent household surveys (such as the <i>Labour Force surveys</i>) or to collect some further information in administrative files (such as the files of the Public Employment Service, called <i>Agence Nationale pour l'Emploi</i> or ANPE in France) which permit to evaluate the effects of some newly introduced reforms or programmes.	Yes! After some improvements in recent years this is the case. Restrictions apply, e.g., concerning self-employment.	Usually, there are very few restrictions on merging different registers for research purposes. Thus, if a particular register does not provide sufficient background information, it is generally possible to collect such information from other registers.			I do not understand the question	In experimental design studies, it is quite possible to conduct a highly credible study without access to baseline data. For example, in any study of an educational intervention (say a new approach to mathematics instruction) where academic achievement is the primary outcome of interest, it is possible to simply compare outcomes for the intervention group with those for the control group, controlling only for intervention status. Similarly, in many of the welfare demonstrations of in the late 1980s and early 1990s, the evaluations had a very limited set of baseline data on sample members and outcome measures (for example, quarters of covered employment and reported earnings) for the intervention and control group members. These data were adequate to address the most important questions. Where they fell short was in addressing questions about the mechanisms through which change occurred and the differential effectiveness associated with particular background characteristics.

C. WHERE DO THE DATA FOR IMPACT EVALUATION COME FROM?

C.4. Are the data collected for evaluation purposes subsequently made available for further use to allow replication of the evaluation results? If so, is that done in terms of production of <i>public use files</i> or <i>via some procedure of restricted access</i> to qualified researchers?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
Generally, they are available for further use (for instance, household surveys collected by INSEE, or administrative data on unemployed workers and unemployment spells collected by the Public Employment Service). Household surveys such as the LFS are public use files, while the access to administrative data files (which are property, for example, of the Public Employment Service or the Social Security administration) is restricted to qualified researcher.	Yes, with restricted access to qualified researchers.	Sometimes, we conduct surveys to learn more about what, e.g., employers think of a particular labour market program. We always make the analysis data sets available for replication. We generally don't make the original data available for other researchers. The reason for this is that the subjects have been asked to fill in the questionnaires for a specific purpose. If these data are to be used for other purposes, we must ask for the individuals' consent.		We are working to make all data publicly available.	Most government surveys are placed in public archives in an anonymised form. Replication is rarely undertaken but secondary data analysis is relatively common although the data are vastly under-utilised	Increasingly, data are made available through public use files. However, in some cases, there is a considerable time gap between the completion of the initial study and the issuance of public use files. And, in some cases, data access is restricted to approved individuals with approved study plans. Examples of public use data sets are those for the National Supported Work Demonstration, the Tennessee Class Size Experiment, the New Chance Demonstration, and the Teach for America Evaluation. Examples of some of the publicly available data bases are provided on the following website: http://www.gse.upenn.edu/pimfer/docs/Web%20Resources.pdf . Other major studies, such as the Moving to Opportunities Study of housing subsidies data are still available to only a select group of researchers (see, for examples, the discussion of data access at: http://www.huduser.org/publications/fairhsg/MTOData.html).

C. WHERE DO THE DATA FOR IMPACT EVALUATION COME FROM?

C.5. Which share of the total evaluation costs is usually represented by the costs of data collection?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>It is difficult to evaluate precisely the cost for data collection, since data are very infrequently collected exclusively for an evaluation study. In general, costs associated with data collection are much higher than costs associated with evaluation <i>per se</i>, which are proportionally negligible.</p>	<p>When relying exclusively on administrative data, the costs are usually rather low. For example, within the evaluations of the Hart-reforms, one project part was to provide the data for the involved researchers. That is, a team of the Federal Employment Agency was paid to do that by the Federal Ministry of Labour and no costs were charged to the evaluators of the different programmes. Once you have to rely on survey data, things look different, costs for data collection can easily sum up to 50% of the total budget.</p>	<p>On average, during 2002-2006, we have spent 4.6 % of the total budget on (all kinds of) data purchases at IFAU; 1.5 % of the total budget has been spent on actually collecting data.</p>		<p>Data collection is the big budget item representing anywhere between 2/3-4/5 of the costs.</p>	<p>This is a largely meaningless question because it depends on the size of the programme being evaluated. I have heard '10 per cent' cited as the norm but I cannot validate this.</p>	<p>Data collection generally consumes the largest share of the evaluation budget. However, data collection costs will vary greatly by the nature of the study setting (for example, school versus home), the degree of clustering of the study sample (for example, concentrated geographically versus dispersed over many sites), the duration of the follow up period and number of distinct observations, and the mode of data collection (for example, in person, phone, mail, or administrative sources).</p> <p>There are many ways to control data collection costs. These strategies include employing cost effective sample designs (for example, strategic clustering of the sample; or balancing the intervention and control groups to minimize overall study costs, including programmatic and data collection), through adopting creating plans with respect who is followed up with what frequency, through mixed modes of interviewing (for example, phone interviewing with field follow-up and through in-depth interviewing for random subsamples and administrative data collection for the full sample), and through random block design surveys (to reduce the time required to complete an individual survey).</p>

D. HOW ARE THE EVALUATORS CHOSEN?

D.1. What is the prevailing practice in selecting evaluators? Does it favour “Evaluation Units” within the government agency in charge of the policy, or specialized government agencies, or outsourcing to private (or university-based) research organizations?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>In France, evaluation studies have been most often produced by INSEE (Institut National de la Statistique et des Etudes Economiques, Paris), CREST-INSEE (the INSEE research centre), or the statistical department of the Ministry of Labour (DARES). They have been very rarely done by private or university-based research organizations, because of the difficulty to have access to appropriate data, which are generally produced by INSEE or by the Ministries of Labour and Education.</p>	<p>Outsourcing to university-based or private research institutions (DIW, IZA, ZEW, RWI, etc.).</p>	<p>The evaluation of labour market policies is mostly conducted by specialized gov't agencies (The IFAU). On average – if we restrict attention to proper evaluation – I would say that it is a combination of specialized government agencies and university-based researchers.</p>	<p>I remain very wary of in-house (public-sector) evaluation agencies. There is a natural tendency to censor negative findings, to pass off unscientific outcomes as "true" outcomes and/or to avoid serious evaluations altogether. I can provide examples from Australia and France to back these concerns.</p> <p>At the same time, it would be unreasonable to argue that all evaluations should be done by agencies at arms-length from government agencies. For one thing, most of these agencies are subsidized to undertake the evaluations and/or hired by a government agency so they may be susceptible to indirect influence.</p> <p>I think it is vital to insist that all of the evaluation data/analysis be made available to outside researchers. In this way, the threat of replication will serve to maintain a degree of honesty and independence on the part of the evaluators.</p>	<p>IE teams include government and evaluators. Evaluators are usually outside resources from WB, research institutes and academia.</p>	<p>Outsourcing via competitive tendering is the norm for major evaluation. There is also a significant amount of monitoring and modelling done by specialists in government .</p>	<p>The most common method for selecting evaluators is through competitive bidding on a scope of work generated by the government. These competitions favour research firms (and sometimes academic institutions) with prior relevant experience. However, there is a fairly large set of credible competitors and rarely does any firm have a “lock” on a contract.</p>

D. HOW ARE THE EVALUATORS CHOSEN?

D.2. What are the dis/advantages of having the evaluation conducted by special “Evaluation Units” located <i>within</i> the government agency in charge of the policy? How long is the process for setting up such Evaluation Units?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>I think that the main disadvantage is a potential lack of political autonomy with respect to the government, which can put some pressure on the researchers of these units. However, in France, such units have reached a good expertise in the use of modern statistical evaluation methods. That is their main advantage.</p>	<p>A disadvantage is clearly the “credibility” of the results. One might fear that research is not as independent as it should be when “evaluation units” are based within the government agency. Furthermore, since a highly specialized staff is required, costs will be considerably high and it is not clear whether this might be financed within the permanent staff.</p>	<p>The main disadvantage concerns independence. It is not good practice to have the same agency formulating, implementing, and evaluating the policy. Another issue concerns proper research standards. The standards of the academic community may be very different from the standards of the government agency. The main advantage is easy access to information on implementation, data, and the workings of the policy.</p>		<p>Disadvantage: capacity constraints and skills. Advantage: relevance, ownership, feedback into policy design</p>	<p>In Britain, the results might not be trusted; 'spin' is a devalued term that has infected the collective consciousness</p>	<p>This is generally not relevant in the U.S. The General Accounting Office and Office of Management and Budget conduct process and implementation evaluations and monitoring studies. They also have conducted studies of the quality of contracted research. However, I am not aware of any instance where they have conducted an intervention study.</p>

D. HOW ARE THE EVALUATORS CHOSEN?

D.3. What are the dis/advantages of having the evaluation conducted by <i>another government agency</i> (including public bodies in charge of the oversight of government activities: e.g. the Government Accountability Office in the US or the National Audit Offices in European countries)?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
The main advantage is a greater distance to ministries, and a high visibility (and a high impact) in the public debate. Public agencies closely related to ministries can be suspected of a lower degree of independence.		Advantages: greater potential for independence (compared to 2), a long-run commitment gives better opportunities to build up expertise in the field (compared to 4), greater opportunities for influencing data collection and implementation (compared to 4). Disadvantages: actual independence may be an issue (relative to 4), funds may be locked-in an activity that eventually degenerates to something distant from research standards (relative to 4).		Advantage: independence; Possible disadvantage: access to information, conflicting motivation, lack of feedback		See question 2 above.
D.4. What are the dis/advantages of having the evaluation outsourced to <i>private research organizations</i> (or university-based research centres) through <i>ad hoc</i> contracts?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
The main advantage is to produce a second evaluation, that can complete or even contest the results of evaluation studies produced by public agencies. In particular, university-based research centres should have an easy access to data collected by public agencies for evaluation.		Invert answers under 3. In addition, competition should improve research standards provided that the outsourcers can recognize good research.		Advantage: quality/rigor, independence. Disadvantage: if the work is not done in close collaboration with implementing agency then it may become an academic exercise	Lack of communication about policy development and the politics of the evaluation means that evaluators may not address the most appropriate policy questions in the most appropriate way. See my publications. Trust is vital commodity	The advantage is that there is competition among firms to come up with the most creative and efficient evaluation design. Firms also can be held accountable for their performance and for cost containment. The government manages control over the process of dissemination, but not whether the product can enter the public domain. The policies and practices the government is interested in studying changes quite significantly over time. Thus, if the government were to do its own research, it would make sense to have a shared research capability. However, it is likely that a government research entity would be less capable of the size and skills of its evaluator pool quickly enough to efficiently respond to shifting priorities and evolving technical requirements of the field. The private sector has been quite responsive to the shifting needs for evaluators with necessary substantive and technical expertise.

D. HOW ARE THE EVALUATORS CHOSEN?

D.5. To what extent is such outsourcing motivated by the need for “independence” on the part of the evaluator, or instead by the need for specialized methodological expertise and operational flexibility in conducting the evaluation?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
Both arguments are equally important.	Mix of both	Can't provide an answer to the general question. Within the realm of labour market policy, the IFAU is viewed as an independent research institute, so independence is not an issue (as yet). When there has been outsourcing in this field it is either because the IFAU has declined to do it (e.g. because our budget is too limited) or because policy-makers have felt that we have a too limited perspective (that of economists).		Need to draw a difference between external and internal evaluation. The objective of the former is accountability. The objective of the latter is managing by/for results. Internal evaluation with top quality outside resources is a good way of getting rigor without sacrificing policy relevance.	Independence is important long-term for reasons of credibility but most politicians are not around long-term. The centrality of evaluation means that there is insufficient expertise available within government departments.	Both factors are important. Contracted research increases the appearance and reality of objectivity. But, as noted above, in the U.S., the private research sector has greater flexibility and incentive to adjust its workforce to meet the needs of government agencies.
D.6. In the case of outsourcing, are competitive procurement procedures followed to select the evaluator? Are there difficulties in administering effectively a competitive procurement for evaluation contracts on the part of government agencies?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
In this case, competitive procurement procedures should be imperatively implemented. My personal experience as a reviewer of research proposals and as member of evaluation committees conducts me to think that there are no inherent difficulties in organizing such procedures.	Yes, competitive procurement procedures are followed to select evaluators. Since there is usually a need for highly specialized methodological expertise (which is usually not available in the government agencies), the agencies use a scientific coordinator who helps them deciding between the different bids. It has been proven as crucially important to have a coordinator who knows the field.	Yes, competitive procurement procedures are usually followed (researchers get to bid for the contract). I think the main issue is whether government officials can judge whether the proposal is of high quality.		Intellectual property need to be addressed. No researcher will want to work on design and compete to implement. In our experience, we hand pick evaluator and procure data collection competitively; or procure competitively a package. The latter only in cases where design has already been developed by internal resources.	I deal with these issues at length in the publications you have	The process works quite smoothly. There are policies and procedures in place to expedite contracts when necessary through the use of “master ordering agreements” or “approved vendor” lists. Furthermore, there are incentives in many procurements to encourage bidding by new contractors, particularly small, women-owned, and minority owned businesses.

D. HOW ARE THE EVALUATORS CHOSEN?

D.7. During the implementation of policy subject to an impact evaluation, is there a complete separation between the agency in charge of the policy and the evaluator (either a government agency or an external evaluator), or is there significant <i>interaction</i> between the two parties? If some interaction takes place, what are its typical features?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
There should be a separation (which is a guarantee of independence) during the reviewing process of the research proposal and during the evaluation of the quality of the results. However significant interaction should take place in two distinct phases: 1) the agency in charge of the policy should help in constructing the data and making them available to the researchers, 2) it should also participate in the interpretation of the results and suggest how to disseminate them.	Yes, there is interaction to clarify the needs of the agency and/or adjust policies and evaluation design.	Excluding randomized assignment, there is not much contact between those in charge of implementing the policy. When there is need the agency in charge of the policy may be contacted such that the evaluators get the institutional detail right. They are also encouraged to provide feed-back on the analysis prior to publication. But, they are not allowed to have any say on methods, data used, as well as outcomes to be analyzed.		Full interaction from policy design onwards, including discussion of priorities & questions, development of evaluation design, implementation, capacity development of implementing agency through evaluation cycle (know why evaluate, what they'll get out of it, what to ask and how to use the results, and operational implications).	There is usually significant interaction arising from the reporting of interim findings, project management requirements including in-field design issues, policy briefing etc. Working practices vary between government departments	Generally there is not complete separation. However, this varies considerably from contract to contract and among agencies. In most circumstances, critical firewalls are erected. However, there are many instances where it is desirable to engage the support of the government agency in the evaluation task (for example, in site recruitment and in dealing with potential violations of intended intervention implementation).
D.8. Does the interaction possibly involve also the beneficiaries of the policy (or their representative bodies – e.g. as in the case of industrial policies)? Do these interactions possibly affect the policy and the evaluation design itself, and in which direction?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
Yes, the interaction should also involve the beneficiaries of the policy (or their representative bodies), especially when defining the general framework of the evaluation study (contours of the group of beneficiaries, possible sites, list of relevant covariates and outcome variables, etc.) and interpreting the results. I don't think that such a participation could affect significantly the evaluation design. On the contrary, it could help to reach a larger consent regarding the usefulness of the evaluation.		No.		no example	My guess is that this is rare. However, I have included representatives of disability rights groups on advisory groups when evaluating the impact of activation policies aimed at recipients of incapacity benefits	A good evaluator will insist on designing a study that will be impervious to the agency's involvement. That is to say, critical firewalls will be in place. It is important for the contracting agency, along the program or agency partners, to buy into the importance of an objective, rigorous evaluation. The biggest threats to the integrity of any evaluation are (1) evaluators who lack the technical expertise to design and carry out a rigorous study to address the questions of interest and/or (2) evaluators who themselves have personal biases or interests that interfere with their commitment to carrying out an objective, rigorous study.

E. HOW ARE IMPACT EVALUATION RESULTS USED?

E.1. Is there a positive relationship between the quality of the evaluation and the likelihood of it being used? In case there is such a relationship, which is the likely direction of causality? Better evaluation favours utilization, or is the prospect of utilization that stimulates better evaluations by motivating the evaluators?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>The link is not straightforward. A pessimistic view could be: <i>only "politically" useful evaluations are used</i>. Scientific quality is often something completely orthogonal. Considering the situation prevailing in France, I would say that the effect of the prospect of utilization on the stimulation of evaluators is more probable.</p>	<p>The "evaluation market" is highly competitive in Germany. A high degree of competition between different research institutions but also private consulting firms to get evaluation contracts ensures (mostly) a high quality. Clearly, the likelihood of utilization rises with better evaluations.</p>	<p>In general, I think the answer to the 1st question is yes, at least in the long run. (In the short run it is not obvious that policy makers can differentiate between good and poor evaluations. But in the longer run, the information gets to them). I would say causality primarily goes from better evaluations to utilization. The other route is not unimportant, but most researchers mainly care about what the research community thinks of the evaluation.</p>	<p>Unfortunately, it is not obvious to me that there is a strong relationship between the quality of an evaluation and its impact. For this, to be true, there would have to be a better appreciation among the informed public, policy makers and politicians as to (a) what constitutes a quality evaluation; and (b) a good sense of how one should interpret the findings. We are a long way from this ideal in many countries, especially those where the evaluation culture has hardly taken root. Even if the evaluation culture has taken root, there is strong political pressure to go for short-run evaluations and attempt to declare the policy a great success. However, it is the case that many policies will have their full effects only over the medium or long-run, say 6-10 years, and not many evaluations have such a longitudinal perspective. This time perspective matters: witness the evaluations of training programmes which often yield low or negative returns for the first few years, and only yield positive returns after 3-4 years or more. Another example is the SSP in Canada. The first years of evaluating this demonstration showed positive outcomes. However, once the time limit of five years for receipt of the earnings supplement was passed, the evaluation results revealed a rather different story</p>	<p>High correlation between institutional buy in and utilization. Correlation with quality depends on how much investment has gone into client capacity/awareness.</p>	<p>Good evaluative research makes it easier to defend unwelcome findings. However, the link between quality and usage is not necessarily close as I discuss in the papers that you already have. Timing, receptiveness of policy makers and the usefulness (politically and in policy terms) are often very important</p>	<p>Increasingly, stronger evaluation are leading to program and policy change. For example, there has been a major overhaul of Head Start in response to mediocre impact findings and the Upward Bound program altered its targeting based on the findings of a rigorous evaluation. Certainly, the convergent findings from the many rigorous studies of welfare reforms that occurred in the late 1980s and early 1990s were very instrumental in the 1996 bipartisan welfare reforms. It is, of course, quite likely that had there not been a near total switch to random assignment design studies the findings would not have converged to suggest that more paternalistic policies toward work expectations and more generous and supportive policies toward work-related child care and other supports would promote self-sufficiency and reduce welfare costs, while not exacerbating poverty. See for example the Association for Public Policy Analysis and Management Presidential Address of Judith Gueron (http://www3.interscience.wiley.com/cgi-bin/abstract/103524279/ABSTRACT?CRETRY=1&SRETRY=0).</p>

E. HOW ARE IMPACT EVALUATION RESULTS USED?

E.2. Are there significant examples in which “interim” impact evaluations have been used to modify the implementation of a policy or programme? Are there significant examples in which “final” evaluations have been utilized in the decision to dis/continue a policy or programme or to alter its design significantly?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
<p>Unfortunately, I cannot find such examples in France.</p>	<p>Yes, some of the Hartz-reforms have been adjusted based on the interim re-ports of the evaluation teams. The degree to which the final reports are used depend on the view of the spectator. It is far to say, that not all political deci-sion are evidence-based.</p>	<p>There are a few examples where interim evaluations have been used to fine-tune the details of the program. There are some minor programs that were terminated as a result of evaluations (and follow-ups) made by the IFAU. I give five examples of some significance where evaluation reports had a direct impact on policy. (1) Subsidized career breaks were abolished by the current government (see IFAU WP 2005:17). (2) The income ceilings in unemployment insurance and sickness insurance were harmonized to avoid flows between unemployment and sickness insurance that were due to benefit arbitrage (see IFAU WP 2002:6). (3) There was a shift in emphasis from labour market training to employment subsidies partly as a result of evaluations made by the IFAU; according to the evaluations employment subsidies seem to work, but the evidence in favour of labour market training is more dismal (see IFAU WP 2004:18 on employment subsidies; IFAU WP 2002:4 for an overview of evaluation results). (4) The possibilities for qualifying for a new period of UI benefits via labour market programmes were removed (see IFAU WP 2001:5). (6) The current government has decided to centralize the provision of active labour market programmes for youth partly as a result of IFAU evaluation reports (see IFAU WP 2006:6).</p>	<p>As for examples of whether evaluations have an impact on programmes, let me cite two examples, one positive and the other negative. When the national evaluation of the Job Training Partnership Act (JTPA) revealed that it had failed to provide earnings gains to disadvantaged youths, the US Congress eliminated almost all JTPA funding. However, the political process is not necessarily symmetric in the sense of rewarding successful active measures. Denny et al. (2000) highlight the fact that a number of successful active labour market policies in Ireland were either eliminated or run down in scale in the second half of the 1990s whereas some unsuccessful programmes were expanded.</p>	<p>Miguel and Kremer 2004 Kenya deworming study resulted in national adoption of deworming in school as part of Kenya Universal Primary Education strategy. The U.S. Job Training Partnership Act 's evaluation provided justification for the US Congress to cut JTPA budget by 80 percent (from \$540M in 1994 to \$110M in 1995). Energy Access Expansion evaluation is testing elasticity to connection barriers on a six months schedule with the purpose of providing the Electric Company with feedback on optimal connection subsidy.</p>	<p>These are illustrated in the publications you have and also in: Walker, R., 'Welfare policy: tendering for evidence', Pp. 141-166 in H. Davies, S. Nutley and P. Smith (eds.) <i>Evidence and Public Policy</i>, Bristol: Policy Press, 2000.</p>	<p>Indeed, evaluations have guided policy decisions. Generally, policies and programs are not altered in major ways on the basis of a single evaluation (for example, see the discussion of the myriad studies with convergent findings that were so powerful in the 1996 welfare reforms). However, welfare and support policies for teenage parents have changed quite dramatically in the U.S. following research that demonstrated that access to family planning services and knowledge about contraception were not, by themselves, effective in stemming repeat pregnancies and births by young unwed mothers. Time limited welfare is a policy based directly on evaluation evidence that it would encourage work-ready individuals to leave welfare and those not work ready but with employment potential to avail themselves of employment training and supports to prepare themselves for work. Evidence of the ineffectiveness of para professional home visitation services for single and high risk mothers led states and federal agencies to abandon its use. And, evidence of the success of employment supports for mentally retarded young adults has led to program expansions and experimentation with variations on this theme.</p>

E. HOW ARE IMPACT EVALUATION RESULTS USED?

E.3. If there are no examples of such direct or instrumental use, is there evidence of a more indirect utilization? Of what Carol Weiss calls “enlightenment” function of evaluation?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
		The current government has decided to introduce UI replacement rates that vary by duration partly as a result of research at IFAU. Also it wants to use monitoring and sanctions to a greater extent. See, e.g., IFAU WP 2003:5 and IFAU WP 2005:13 on these issues.			I would propose a distinction between evaluation and policy research; the former is more likely to be used directly, the latter indirectly. In the US research by Davis Greenberg suggests that administrators at state level are more likely to make use of process evaluation findings than the results of impact studies	See response to question 2 above. Between the first bit of rigorous evidence and a major act of adoption, there generally is a period of “enlightenment” or expanded interest. The current frenzy in the U.S. over the expansion of Pre-K programs and the call by James Heckman, Henry Levin in and others for massive increases in investment in children during their infant and pre-school years is based very heavily on the “enlightment” from studies of two small-scale high-intensity interventions conducted more than 30 years ago—Perry Preschool and Abecedarian. In both cases, the results of these studies are more relevant as evidence of the potential to alter the life course of children raised in very impoverished circumstances than they are evidence of the expected magnitude of effects we could expect from implementing similar interventions today. See, for example, James Heckman’s recent editorial in Education Week http://www-news.uchicago.edu/citations/07/pdf/070319.heckman-ew.pdf .

E.4. Are there some general rules or accepted practices for the dissemination of the results of impact evaluations? Do the media tend to utilize those results?

Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
When the evaluation is conducted by a public agency or “evaluation units” in ministries, the results are generally published in public reports. The media (in particular, newspapers) tend to utilize those results. Some recent examples in France: evaluation of changes in the unemployment insurance system, of implementation of job-search assistance and training programs, payroll tax subsidies, etc.	(Nearly) All of the results are published (either in printed form or on the web). Depending on the policy under evaluation, the media shows a high in-terest.	It is part of the assignment of IFAU to make the evaluation results known and understandable for policy makers and the media. As a general rule, we write popularized reports (in Swedish) in connection to every research paper that we publish. We then decide on whether the results are so interesting that we should make them known to the general public (via press releases) or whether we just write-up an executive summary of the paper. The media tend to use the results, in particular if there is an implicit critique (e.g. a particular labour market program does not improve employment prospects) of the current government policy.		Imbedding evaluation within the implementing agency work is best way to ensure use of results. The purpose of dissemination should be to affect others into using evaluation and forming their opinion of what might work in their context. Results have to be extracted, in small and digestable bits.	There has been a trend towards the automatic publication of evaluation findings although the timing and mode of publicity often remains in the hands of policy makers/politicians. Automatic publication increases credibility and aspirations to the independence of evaluations.	Until the past few years, there were no widely accepted standards for the reporting of impact evaluation findings. However, beginning in the late 1990s there has been increasing consensus regarding the standards for reporting research findings and there are evolving standards for the synthesis and disseminating for findings for purposes of guiding policy and practice. One of the most recent and influential such initiatives is the What Works Clearinghouse established by the U.S. Department of Education (http://www.ed.gov/about/offices/list/ies/ncee/wwc.html). Another prominent site is the Campbell Collaboration (http://www.campbellcollaboration.org/).

E. HOW ARE IMPACT EVALUATION RESULTS USED?

E.5. What is the <i>perception of the utility</i> of impact evaluation among policy-makers? And among the evaluators themselves? And among the media and the general public?						
Fougère (France)	Caliendo (Germany)	Fredriksson (Sweden)	Martin (OECD)	Legovini (World Bank)	Walker (UK)	Maynard (USA)
I think that this perception has changed progressively over the recent years. The general feeling is more and more favourable.		it is not for me to answer. In an evaluation of the IFAU (conducted in 2004) there were attempts to look into this question. In general, the message was that policy-makers take the results seriously. The Results from IFAU-evaluations matter particularly for the longer-run emphasis of labour market policies. Among the media (and the general public) the results are taken seriously (although I think that they are in an even worse position than policy makers to judge the credibility of the evaluation results)			Mixed as indicated above. Again the report due to be published by the UK Home Office is of especially relevance. I suspect that the time (especially) and resources required together with the limited range of questions that can be addressed by impact evaluations will not endear them to most politicians and policy makers. Treasury officials may take a different stance since the requirement for evidence of effectiveness is a good rationing device and reduces policy risk. I doubt if the general public have given the topic much thought and the media likewise. As indicated above, and also by the contents of the journal 'Evaluation' suggests that the evaluation community is deeply divided	There is great diversity of opinion about the value of impact evaluation research. In part this diversity mirrors the diversity in the quality of the research itself. Much of the existing research base is not of high quality—a fact that is quite evident, for example, when reading the intervention reports on the What Works Clearinghouse website. It is not uncommon for the clearinghouse to have found only a handful of studies meeting their standards of evidence from among dozens that have been done. There certainly are many senior policy officials in Congress and in the Executive service who understand and value rigorous research, as evidenced by the number of congressionally mandated studies, the role of evaluation in legislative and policy debates, and the high level of government investment in rigorous evaluation.