COLLOQUE SUR LES EXPERIMENTATIONS POUR LES POLITIQUES PUBLIQUES DE L'EMPLOI ET DE LA FORMATION

22 et 23 mai 2008

SOMMAIRE

Présentations académiques d'expérimentations

Ouverture

Antoine MAGNIER, directeur de la DARES

Monsieur le Haut Commissaire, Mesdames, Messieurs, je suis Antoine Magnier, directeur de la DARES, et je suis heureux de vous accueillir pour ce colloque sur les expérimentations pour les politiques publiques dans le domaine de l'emploi et de la formation. L'objet de ce colloque, vous le savez, c'est d'échanger sur les questions que peuvent poser les expérimentations, ainsi que sur les potentialités et les enseignements que l'on peut tirer des expérimentations qui ont été menées à l'étranger et de celles qui se mettent en place depuis peu en France.

Ce colloque intervient à un moment un peu particulier dans notre pays, alors qu'un plus grand consensus se fait jour parmi les décideurs publics, entre le gouvernement, le Parlement, les responsables des collectivités locales et les partenaires sociaux, pour développer l'évaluation des politiques publiques et ce notamment dans le domaine de la formation et de l'emploi. Ce colloque intervient également à un moment où le gouvernement a décidé de s'appuyer explicitement sur une démarche d'expérimentation, pour progresser dans le domaine de l'insertion sociale et de l'emploi, sous l'impulsion du Haut Commissaire aux Solidarités actives contre la pauvreté, Martin Hirsch, et sous la responsabilité également des Ministres en charge du Travail et de l'Emploi.

Nous avons organisé ce colloque avec l'appui décisif d'Esther Duflo, du MIT, et de Bruno Crépon, du CREST. Je souhaiterais les remercier vivement pour cette collaboration, ainsi que notre mission d'animation de la recherche, pilotée par Dominique Goux, qui nous ont préparé ensemble ces deux journées d'information et d'échanges. Je souhaiterais remercier aussi particulièrement les participants qui nous viennent de l'étranger, pour nous éclairer sur les expérimentations auxquelles ils ont participé ou qu'ils ont finement étudiées. Je ne doute pas que leur expérience et leur perspective nous seront utiles pour progresser dans cette voie.

Ce colloque se tient sur deux jours. Aujourd'hui, nous aurons la possibilité d'échanger, sur la base d'exposés et travaux de nature académique, sur un certain nombre d'expérimentations menées à l'étranger. Ce matin, dans le domaine de l'économie du développement et de l'éducation, sous la présidence de John Martin, de l'OCDE ; cet après-midi, dans le domaine du travail et de l'emploi. Demain, nous reviendrons dans un deuxième temps sur les principaux aspects méthodologiques, juridiques et éthiques, que pose la mise en œuvre des expérimentations. Nous évoquerons également les expérimentations qui sont développées en France et celles qui se mettent en place à l'initiative du Haut Commissaire.

Avant de lui laisser la parole pour lancer les travaux de cette journée, je souhaiterais vous transmettre également les regrets de Christine Lagarde, notre Ministre de l'Economie, de l'Industrie et de l'Emploi et ceux de Laurent Wauquiez, notre Secrétaire d'Etat en charge de l'Emploi. Ils devaient nous rejoindre aujourd'hui et demain, mais ils ne seront finalement pas en mesure de le faire. Ils le regrettent vivement et vous prient de les en excuser. Leur directeur

de cabinet, Thomas Fatome, nous apportera néanmoins leur voix demain, pour la clôture de ce colloque.

Voilà ce que je souhaitais dire en guise d'introduction et pour accueillir Martin Hirsch, notre Haut Commissaire aux Solidarités actives contre la pauvreté, un Haut Commissaire dont l'intérêt, l'enthousiasme et l'énergie qu'il déploie en faveur de la démarche expérimentale n'est plus à démontrer. Je vous remercie.

Retour sommaire

Martin HIRSCH, Haut Commissaire aux Solidarités actives contre la pauvreté

Merci beaucoup pour cette introduction. Merci de dire que nous avons de l'énergie. Merci aussi d'avoir organisé ce colloque, dont on attend beaucoup. Je remercie tous les participants étrangers prestigieux qui sont venus apporter leur expérience sur l'expérimentation et la démarche qu'ils ont pu faire dans leur pays. Cela éclairera bon nombre des démarches que nous essayons de faire.

Le rapport de la France à l'expérimentation est paradoxal et ambigu. Expérimenter, c'est un mot qu'on aime bien, mais qui, en réalité, colle assez mal à notre culture des politiques publiques depuis des décennies. Notamment dans les domaines social et de l'emploi, il s'agit plutôt de faire de grandes annonces, à travers le plan qui va résoudre tous les problèmes, de faire la loi générale qui va traduire ces problèmes, de la compléter par une panoplie de textes réglementaires, puis de changer cela dix-huit mois après, de l'évaluer une fois que le système précédent a déjà été changé, etc. On a peu de grandes politiques publiques qui se soient véritablement introduites par des expérimentations aussi contrôlées que celles que vous allez évoquer, et qui ont nourri la genèse de politiques publiques dans d'autres pays.

Pendant des années, les termes "expérimentation" et "évaluation" ont plutôt servi de prétexte ou de repoussoir. Notamment quand on dit : « On va expérimenter quelque chose », la réponse est : « Mais cela fait dix ans qu'on expérimente. Il faut une vraie politique. Ce que l'on vous demande, ce n'est pas de continuer à expérimenter, mais de prendre de vraies mesures ». De la même façon, quand on leur dit qu'il faut évoluer, ils disent : « Ne nous mettez pas des exigences supplémentaires », etc. Avec une sorte de tradition française d'avoir la science infuse plutôt que d'y aller par des méthodes de recherche, de politique fondée sur l'évidence, etc.

Ceci est en train de changer, y compris parce que l'on a touché au texte le plus sacré de notre pays, qui est la Constitution, sur ce point-là. Parce que l'une des raisons qui expliquaient le fait que la politique expérimentale n'avait pas long cours, c'était à la fois le fait que l'on est plutôt dans une culture bardée de certitudes et également parce que l'on interprétait souvent le principe d'égalité comme incompatible avec toute introduction progressive d'une innovation, toute comparaison, tout fait de pouvoir traiter deux populations de manière différente, y compris de traiter deux départements, deux régions, deux villes de manière différente. Avec une sorte de sacralisation du principe d'égalité qui rendait difficile une véritable démarche expérimentale. Donc, s'il y a cependant des politiques qui ont pu être initiées par l'expérimentation, c'était en réalité par le fait que l'on laissait éventuellement deux ou trois

avant-gardistes commencer dans leur département. C'est ce qu'il s'est passé, par exemple, pour l'introduction du RMI, il y a un peu plus de vingt ans, dans lequel trois départements – l'Ille-et-Vilaine, le Doubs et le Territoire de Belfort – ont mis en place à leur propre initiative, en utilisant les petites marges de manœuvre juridiques qui leur étaient ouvertes, la possibilité de commencer un revenu garanti puis, très rapidement, de venir au RMI. Mais jamais dans des conditions dans lesquelles on pouvait comparer, entre un territoire ou un autre, l'effet d'une politique par rapport à une autre. Cela s'est fait relativement rarement dans notre pays.

Cela a changé par la conjonction de deux phénomènes. Le premier est la décentralisation, qui a conduit à ce que des échelons locaux se voient attribuer des responsabilités supplémentaires et par conséquent aient l'envie de pouvoir tester telle ou telle politique de manière réellement innovante. C'est à cause d'elle que nous avons donc changé la Constitution, en 2003, pour permettre et inscrire dans celle-ci la possibilité d'expérimenter, dans une collectivité locale, une politique qui déroge au droit national. Le deuxième phénomène est le poids des dépenses publiques, notamment des dépenses sociales qui ont conduit, à un moment donné, à se dire que si dans les dépenses sociales et les dépenses publiques, on n'a pas les résultats attendus, il est difficile d'annoncer un surcroît de dépenses publiques et de dépenses sociales, sans être certain d'avoir un retour sur investissement.

Ce qui nous a rapprochés d'un certain nombre de pays qui ont plutôt une aversion à la dépense publique et qui se sont dit qu'ils ne pouvaient introduire un progrès, une innovation, un changement social que s'ils étaient capables de démontrer qu'il pouvait y avoir un retour sur investissement. Retour sur investissement, ce n'est pas forcément quelque chose de malsain, grossier et financier, mais il s'agit que l'effet social attendu soit cohérent avec l'effort public ou financier réalisé. Et s'il y a aujourd'hui toute une démarche d'expérimentation autour du retour à l'emploi, de l'accompagnement des demandeurs d'emploi, cela vient notamment du fait qu'on a le sentiment d'une discordance entre le montant des dépenses publiques et le taux de chômage, de la même façon que l'on peut dire qu'en France, il existe une discordance flagrante entre le montant des dépenses sociales et l'évolution de la pauvreté au cours des dernières années.

On peut dire la même chose en matière d'insertion. On a essayé de consolider à l'occasion du Grenelle de l'insertion l'ensemble des montants que l'ensemble des collectivités publiques consacre aux politiques d'insertion. Le chiffre s'élève à près de dix-neuf milliards d'euros par an, somme à la fois considérable et insuffisante au regard des besoins et des résultats que l'on obtient.

Donc, un certain nombre d'acteurs se sont dit que pouvoir changer les politiques sociales, les politiques publiques et les politiques d'emploi en ayant recours à de l'expérimentation rigoureuse pouvait avoir du sens et éviter à la fois les *stop-and-go* et la fuite en avant, qui fait que l'on change et que l'on ne sait pas pour autant comment cela évolue.

Alors, c'est une chose à laquelle nous nous sommes attachés, parce que nous avons considéré que les politiques à la fois d'incitation de retour à l'emploi et de minima sociaux ne produisaient pas les effets escomptés. Ainsi, il y a vingt ans, le RMI a été calibré en imaginant

qu'il n'y aurait que quatre ou cinq cent mille personnes qui en bénéficieraient. Or, aujourd'hui, plus d'un million en bénéficie. Les mécanismes d'incitation ou de retour à l'emploi tels que la prime pour l'emploi ont été faits à grand renfort de tambours et trompettes, en expliquant qu'ils allaient avoir un effet incitatif sur l'emploi. On y consacre près de quatre milliards d'euros et demi pour se trouver ensuite avec la Cour des comptes qui nous explique, et je la cite : « Ces politiques ne sont ni incitatives, ni redistributives », alors que c'était les deux effets recherchés.

Quand nous avions réfléchi, avec un certain nombre d'acteurs, à proposer des changements assez radicaux dans la manière dont on appréhendait les minima sociaux et les politiques incitatives, on s'était bien évidemment demandé ce qu'il se passait dans d'autres pays et on avait essayé de tenir compte du *Earned Income Tax Credit* aux Etats-Unis, du *Working Tax Credit* en Angleterre, du début du *Kombilohn* en Allemagne, d'un certain nombre de politiques qui s'étaient mises en place et de se demander d'abord comment elles s'étaient mises en place, quel retour l'on pouvait en avoir et, si l'on choisissait de s'inspirer de cela, comment cela pouvait être introduit en France.

Quelque chose m'avait toujours frappé, quand on lisait les ouvrages des chercheurs et des universitaires en France, sur les politiques sociales, chaque fois qu'ils citaient des programmes d'expérimentation sociale, aucun ne se référait à des programmes français, toujours à des programmes étrangers. Nous avions donc proposé à la fois que l'on fasse un investissement conséquent et que l'on puisse commencer par des programmes expérimentaux. Nous avions commencé à y réfléchir avec un certain nombre d'universitaires français.

Mais c'est là que les ennuis commencent ! Dans ce cas-là, tout le monde est pour l'expérimentation. Tout le monde trouve cela absolument fantastique. Tout le monde trouve l'idée très bonne. Mais, quand il s'agit d'en faire la traduction concrète sur des vraies personnes, sur des vrais humains, sur des vrais territoires, avec de vrais élus, de vraies institutions et que l'on invite Marc Gurgand, Eric Maurin ou d'autres à leur expliquer les prérequis méthodologiques, les yeux s'écarquillent, on vous dit que vous êtes vraiment sympathique mais que si vous pouviez le faire dans le territoire d'à côté, cela ne serait pas forcément plus mal !

Alors, on a essayé toutefois, pour introduire cette logique expérimentale dans la conception de nos politiques publiques, de commencer par des compromis entre les exigences méthodologiques et ce qui commençait à être acceptable par les différents acteurs. Ce que nous sommes en train de faire sur le Revenu de solidarité active est une sorte de compromis, avec les bons et les mauvais côtés que cela sous-entend. A titre d'exemple, parmi les éléments que les chercheurs anglo-saxons nous indiquent comme nécessaires à la réussite d'une expérimentation, l'on trouve la possibilité de pouvoir aléatoirement catégoriser ou choisir entre deux échantillons que l'on puisse comparer. Alors, on dit : « On va prendre les allocataires du RMI, on va les tirer au sort », etc. Les éléments éthiques, voire juridiques, viennent s'y opposer. Alors, nous avons essayé de les contourner, tout en sachant que de ce fait on affaiblissait la force de nos protocoles mais que l'on pouvait acclimater les acteurs à cette logique. Pour l'instant, face à l'impossibilité juridique de pouvoir tirer au sort dans un

même territoire les différents publics, nous avons essayé de tirer au sort les territoires. On pensait alors avoir déjà contourné un obstacle ; à ce moment-là, les élus nous ont dit : « Tirer au sort les territoires, on est d'accord, du moment que vous tombez sur le territoire où je suis moi-même élu, pour la partie positive ». On a réussi, là aussi, à dire que l'on pouvait caractériser des territoires homogènes et à tirer au sort des zones témoins qui soient comparables dans leur dynamique d'emploi, de pauvreté, aux territoires auxquels allait s'appliquer le dispositif. On s'est habitués à contourner, ruser avec les différents obstacles, pour que la logique de l'expérimentation soit introduite dans ce pays et sur une politique sociale importante.

Au moment où l'on essaie de passer à la vitesse supérieure, tous ceux qui disaient : « L'expérimentation, on n'aime pas cela » nous demandent de la prolonger. Cela est bon signe, preuve que cette logique rentre dans la culture de notre pays.

C'est une chose sur laquelle on travaille. Si on ne s'y prend pas trop mal, le Revenu de solidarité active sera la première réforme sur laquelle on aura tant investi dans la phase expérimentale, dans le protocole d'évaluation prégénéralisation, dans les prérequis méthodologiques pour le faire, donc grâce à un certain nombre de chercheurs, dont François Bourguignon, qui interviendra demain et qui préside notre comité d'évaluation.

Parallèlement, on a essayé de pousser des acteurs à se lancer dans des programmes expérimentaux. Un des problèmes que l'on rencontre dans le domaine de lutte contre la pauvreté, de soutien, d'accompagnement des plus défavorisés par rapport à l'éducation, à la santé, à la mobilité, à l'emploi, c'est de montrer que si l'on fait un changement, il va produire de l'effet ou par ailleurs, démontrer que si quelque chose se fait à petite échelle, pour laquelle, en général, on trouve que cela est formidable, le passage à la moyenne échelle ou à la grande échelle peut se faire dans des conditions dans lesquelles on n'annule pas les effets que l'on avait à la petite échelle et qui peuvent être influencés par tel ou tel biais de sélection ou de moyens très concentrés qui peuvent être faits pour réaliser quelque chose à la petite échelle. Pour ce faire, on a lancé un appel à projets il y a six mois, en incitant des acteurs, en essayant de faire le "Meetic" de l'expérimentation sociale, en disant : soit ce sont des chercheurs et des universitaires qui ont des protocoles à proposer et on essaiera de trouver des collectivités d'accord pour être terrain d'expérimentation ; soit on a des collectivités, des entreprises, des associations, des acteurs locaux qui ont des projets et, dans ce cas-là, on essaiera de trouver des chercheurs pour pouvoir les aider à les mettre dans un protocole qui rende les choses véritablement évaluables et dans des conditions dans lesquelles on peut en tirer des leçons. On a eu un processus exigeant de sélection, d'appariement, de retravail sur ces projets, en ne visant pas l'exigence maximale et parfaite dès le début, mais en essayant d'avoir les choses permettant d'avoir un niveau d'exigence plus élevé que ce que l'on faisait habituellement.

C'est ce qui nous a conduits, après tout ce processus de sélection qui vous est rappelé dans votre dossier, à pouvoir retenir trente-sept projets dans des domaines de santé, de retour à l'emploi. Parmi eux, trois ou quatre qui nous tiennent à cœur et qui sont illustratifs de la variété : pouvoir évaluer la prévention des ruptures dans l'apprentissage en Corrèze, autour de Tulle. On est partis d'un projet de chercheurs qui étaient intéressés par ce sujet et d'une

mission locale pour l'emploi qui se trouvait confrontée à ces difficultés, ils se sont appariés pour pouvoir faire un programme. De la même façon que nous vivons dans un pays dans lequel on a toujours l'idée qu'il faut encourager le soutien scolaire des enfants défavorisés et puis l'on s'aperçoit que quand on met des actions de soutien scolaire, ce ne sont souvent pas les enfants défavorisés qui en bénéficient. Puis, quand on demande aux différents chercheurs ce qui leur paraît le plus efficace en matière de soutien scolaire, il y a une telle diversité qu'il y a très peu d'évaluations. Là aussi, nous avons un projet qui a été déposé par l'Ecole d'économie de Paris et qui cherche à mesurer l'impact du soutien scolaire et des actions d'information et de sensibilisation vis-à-vis des parents, sur les résultats. De la même façon que la région Ile-de-France a déposé un projet sur la manière dont la construction des programmes de formation professionnelle pouvait être modifiée pour pouvoir mieux tenir compte des besoins des personnes les plus éloignées de l'emploi, qui sont en général les moins concernées par les programmes de formation.

Donc, nous avons toute une série de programmes qui sont lancés et nous sommes prêts à continuer dans cette dynamique.

Je terminerai par deux remarques : l'une française, l'autre européenne. En France, nous souhaitons et nous avons l'ambition de pouvoir, si ce n'est généraliser, étendre cette pratique pour que, de plus en plus, les réformes sociales puissent se fonder sur des données acquises par l'expérimentation. C'est toujours le démarrage qui est le plus difficile, face aux impatiences des différents acteurs publics. Quand vous êtes élu président de la République, président du Conseil général, vous souhaitez pouvoir mettre en œuvre une nouvelle politique le mois d'après, en avoir des résultats l'année suivante, etc., en fonction de votre mandat. Donc, si certains, très sympathiques (hauts commissaires, chercheurs, etc.) arrivent en disant : « On a une idée pour votre mandat, c'est que vous allez lancer une expérimentation ; nous en aurons les résultats à la fin de votre mandat et votre successeur pourra les mettre en œuvre », là aussi on a un succès d'estime mais, en général, un enthousiasme relativement mesuré. Or, si l'on arrive à avoir un portefeuille de données expérimentales qui permette que dans deux, trois ou quatre ans on puisse, au moment où se font les discussions des grands choix politiques, éviter qu'elles se fassent dans le vide ou de manière idéologique, mais qu'elles puissent se fonder sur des données expérimentales solides, on arrivera, petit à petit, à changer progressivement le débat sur les politiques sociales et sur leurs changements.

Ceci suppose un certain nombre de choses, à la fois juridiques, c'est-à-dire d'aller plus loin dans les marges de manœuvre juridiques que l'on peut donner pour le faire – et je pense que l'on profitera des textes pour pouvoir le faire – et qu'il y ait de l'argent pour cela. A la fois de l'argent public, d'essayer de prévoir, comme le font certains pays, qu'une partie des dépenses publiques soit conservée et consacrée à l'expérimentation. Cela nécessite aussi que les acteurs privés et les fondations s'intéressent davantage à ce sujet, pour que l'on puisse mettre un peu d'argent privé pour soutenir ces programmes.

En France, on a l'ambition, dans les mois et années qui viennent, de pouvoir s'appuyer sur ces appels à projets, sur vos expériences étrangères, pour parvenir à le faire. De la même façon, au niveau européen, la présidence française va commencer dans quelques semaines et je fais

donc à présent le tour des capitales européennes, pour voir ce que l'on met dans l'agenda social. Quand on leur dit qu'on va mettre de la réglementation, un certain nombre de pays nous répondent : « Ne comptez pas sur nous là-dessus ». Un certain nombre de choses ne collent pas du tout par rapport aux ambitions des différents pays. Cependant, quand l'on discute, l'on s'aperçoit que les problèmes sont assez analogues. L'autre jour, j'étais dans une ville dans laquelle on m'a fait visiter le service public de l'emploi, l'action spécifique qui avait été émise pour le retour à l'emploi des chômeurs de longue durée, pour les immigrés et les personnes les plus éloignées de l'emploi, en m'expliquant que l'on avait mis cela en place parce que l'ANPE ne voulait pas s'occuper de ces publics. Je me croyais dans une province française et j'étais à Stockholm. Ils avaient donc, en Suède, la même problématique que celle que l'on rencontre chez nous, d'avoir effectivement un service public de l'emploi pas toujours concentré sur les personnes les plus en difficulté, dont les critères d'évaluation, etc., ne sont pas organisés autour de ceux-là. Quand on explique à ces pays que l'on souhaiterait pouvoir promouvoir des expérimentations coordonnées entre plusieurs pays européens, là, l'intérêt se manifeste et j'ai pu le voir la semaine dernière, en discutant avec le commissaire européen. Nous allons profiter de la présidence française pour pouvoir, d'abord, prendre un peu d'argent de la commission et lancer des programmes expérimentaux. On fera ce lancement à Grenoble, à la fin du mois de novembre, en faisant venir celles et ceux qui, dans les différents pays européens, universitaires ou opérateurs, s'intéressent au programme d'expérimentation, pour que l'on puisse lancer des programmes d'expérimentation à l'échelle de quelques localités choisies dans différents pays européens.

Vous voyez pourquoi l'on attend beaucoup de ce que vous ferez et direz pendant ces deux jours. Nous avons l'ambition qu'en France et qu'au niveau européen nous puissions tirer beaucoup des programmes d'expérimentation dans le domaine social.

Je vous remercie de votre attention.

Antoine MAGNIER

Merci, Monsieur le Haut Commissaire, pour ces perspectives éclairantes sur la conduite des politiques publiques dans notre pays et sur le développement des expérimentations, notamment celles de nature contrôlée. Notre colloque aujourd'hui a également une autre valeur expérimentale : il coïncide avec un mouvement social d'ampleur dans notre pays. Je ne sais pas si les intervenants suivants, Monsieur Banerjee et Monsieur Angrist, ont pu nous rejoindre. C'est le cas, je m'en félicite et je laisse la parole à John Martin, directeur du département de l'éducation, de l'emploi et des affaires sociales, à l'OCDE, qui va présider cette séance.

Retour sommaire

Expérimentations en économie du développement

John MARTIN, OCDE

Bonjour tout le monde. Je suis John Martin, le directeur de l'Emploi et des Affaires sociales à l'OCDE. J'ai l'honneur de présider les séances de la matinée. Avant de donner la parole à Monsieur Banerjee, je voudrais simplement vous faire part des règles. Chaque intervenant a droit à 45 minutes, pas plus. Ensuite, le discutant aura droit à une quinzaine de minutes, ce qui nous laisse normalement un quart d'heure pour les questions avec la salle. J'aimerais avoir le maximum d'interaction avec la salle. Je vais donc vous demander de poser des questions assez courtes, précises et de ne pas faire de longs commentaires ou des discours.

Abhijit BANERJEE, MIT

Thank you for being here, it's really exciting for us from the Poverty Action Lab to be with you. For us this is a very significant occasion. This colloquial is part of our expansion into various necessary areas and we feel very privileged to have an audience like this here, in this wonderful city.

What I wanted to talk about today was randomized evaluation in development economics. It has become one of the integral and most influential parts of development economics over a period as short as ten years. I think ten years ago it was not anywhere; it was very marginally on the map. Now I think it is almost to the point where people think that there may be too much of it. And so, it has had a very major influence on how development economics and more generally the economics of policy and economics of the poor are studied. Let us start with what is a randomized evaluation. Or rather, let us start with the even prior question, which is what are impact evaluations? The word evaluation is one of the most overused words in the world. People use it to mean many different things. An impact evaluation is an evaluation which tells you whether what you are doing has the particular impact you were expecting. In other words, if you give textbooks to children does it raise test scores? If it makes them better at playing tennis, that would not constitute an impact. As against that, what is another use of the word evaluation? That is in some ways important because that's what's contentious. An alternative sense the word evaluation is commonly given, is what we call a process evaluation, which is to say, did the textbooks actually reach the children? Did they read the textbooks? So in other words, asking the question, did what was supposed to happen, happen? And that is also a sense of evaluation. I think our main focus is on impact evaluation, and randomized evaluation is a very increasingly recommended standard and in many ways very influential technique for doing impact evaluations.

So why do evaluations? Why did we ever think this is going to be an interesting subject? I think that it is in some sense easy to sort of forget how specific policy is relative to most things that happen in the world. If you take toothpaste, the one thing you know about toothpaste is that you don't need an impact evaluation, why don't you need an impact evaluation? Because there is a natural mechanism for getting rid of toothpaste that people

don't like, so if your toothpaste comes in a beef flavor and people don't like beef flavored toothpaste, then beef flavored toothpaste will go out of business. There is a natural market test for many things.

What is distinct about a lot of social policy is that for two distinct reasons there is no market test. One I think because there is often no choice, you have to send your children to school, and there is only one school, if the school doesn't work then you still just send your child to that school. That's one reason. The other reason is that governments often want to make you do things that you don't want to do. The whole point of government interventions is to make you not smoke in certain particular private or public locations. There is no way you will get a market test of that. It exists partly because governments want you to do things that you don't want to do. And in that sense there is no natural way that, of course, people will not want to do it, and the fact that people don't like it is never going to be a good test of whether the social policies work.

You have to ask a different question, that is a reason why in the context of government policy, when we deal a lot with market failures and where precisely the government acts in ways to restrict what people might want to do, or to direct them in other directions, you would worry very much about something like relying on a market test, or on randomized evaluations. So coming to the point now, randomized evaluations are the social science equivalent of medical trials. It's pretty honest because that is where it came from; it was a test largely to imitate medical trials. In the medical trials what you do is randomly choose a bunch of people and give them a particular medicine, you give the rest a placebo and you look at whether the medicine made a difference.

The basic idea here is exactly the same as when you choose units to receive an intervention, where intervention could be a school in a neighborhood, a micro credit loan to a poor woman, it could be a particular form of advertising, social advertising to get mothers to breast feed their children. But it's something that is chosen—you know you can choose a unit—it could be a village, a neighborhood, a person, that is, who the target of the randomization is could vary. But you pick some unit at random and then you compare the outcomes from the unit with those who didn't get that treatment. And then you see if there is a difference. It's a very straightforward idea, you pick two groups at random, one of them gets the advertising to promote breastfeeding and the other group doesn't get the advertising to promote breastfeeding. And you look at whether when you promote breastfeeding, is it the case that breastfeeding goes up? Is it the case that as a result of exclusive breast feeding children's immunity goes up? etc. You measure a bunch of outcomes. That's an elementary and almost trivial concept.

What is remarkable is that until ten, twelve years ago this was almost never done. So how else did you do an impact evaluation? Well there are standard ways of doing it, some of you are very familiar with it; I still say it just for the record. One thing you could do is to compare the women who breastfeed their children with the children of the women who don't breastfeed, or who don't do exclusive breastfeeding. There you would see a difference, now would you attribute it to the exclusive breastfeeding or would you attribute it to the many things that

make some women do exclusive breastfeeding, and others not? There are lots of reasons why people behave in different ways. Those things have independent effects on what outcomes you are looking for. If one woman is doing exclusive breastfeeding and another is not, that is in itself a potential source of difference. In other words, we need to ask our question, why do we believe they would have had the same outcome absent the intervention. Why is that? If not, we cannot compare them. Because when we compare one woman who exclusively breastfeeds her child with a one who does not, our presumption is that these women absent that choice, that one choice, everything else they would have done the same. And that is just not particularly plausible. You see people do things for a reason, you know people who exclusively breastfeed, may be women who are also working less hours in a day, and can therefore spend more time with the children. And if you see the children's outcomes differ, how do you know if it's coming from the breastfeeding rather than from the fact that the woman has more time in hand and therefore can do other things with the children? Otherwise the child may be taken care of by her nine year old sister, who may not be as careful in many different ways. That's one way of doing it. And that has its obvious problems. The other standard we are doing it is to look at things before and after the intervention. So you did a breast feeding campaign, and you look at before you did the breastfeeding campaign what was the fraction of people doing breast feeding, and after you do the breastfeeding campaign what fraction is doing it? You compare them; the problem is that other things change over time. If you look at the increase in breastfeeding, it's hard to believe that most of the increase in breast feeding that's happened over the last fifteen years, since it's become a focus, was all because of campaigns. A lot of it also has to do with people themselves talking to other people, and all those other sources of difference.

It's hard to know whether the difference we observe is because we did a campaign or whether the difference came because they would have in any case, because things were changing. There is lots of dynamics in society, and so unless we are willing to believe that, other than your campaign, nothing has changed in the world, it's hard to do that. That's what randomized evaluations is supposed to solve, they're supposed to solve the problem of finding a comparison group, by simply saying look I'm going to randomly choose a group, because it is automatically statistically identical to the group that got the treatment. Because it was randomly chosen, that means there is no systematic difference between them, therefore we can always compare the outcomes. That very elementary idea, taken very much from the medical domain, is working for us.

Just to say a bit more about randomized trials, it should have been clear from what I just said, but maybe not as clear to those of you who haven't seen how much economic research gets done, is that one nice thing about what I just told you, is incredibly simple. We randomly choose one group, and compare outcomes. Conceptually at least, sometimes practically it gets hard to do, conceptual is simple to do, the advantage of that is that it is hard to argue with the results, you know where it came from, you know what exactly was done. So as a result, politically it is often extremely powerful because you cannot say, well you did this statistical technique to control for this or that, or the way you did it was not the right way to analyze the data, and if I analyze the data in some other way, we'll get some other conclusion. It's the

standard discourse and policy I want an answer, I analyze the data to get the answer. And there are many different ways to analyze data; as a result we can often get the answers we want.

One big advantage of a technique which basically says I will first announce this is the outcome we are looking for, then I'm going to randomize, and then offer it to some places and not others, and then we'll see what happens. At the end of the day there are no fancy statistics. The results are almost as they are, you see them or you don't see them. That makes them politically very powerful. One of things that is an interesting story, is the story of this Mexican program that has now become very influential in the world, the program called Progresa. It was implemented in the last two years, not of the current Mexican government, not of the previous one, but the one before, which was kind of worried that they would lose the election. Which they did. So one of the things they did was to say, look, we're going to lose the elections, we want the program to survive, so let's do it as a very public randomized trial, before the program goes to full scale. And because they did a very public randomized trial, and the results came out, they invited international scholars to come and do it, the results looked very credible and positive. When the new government came, the new government said look, how do we deal with this? In the end the evidence had been sufficiently publicized that they didn't actually think they could repeal the program, so what they did instead was the next best thing, really, they renamed it. They called it Oportunidades. So Mexico now has a well known program called Oportunidades which is *mutandis mutandis* Progresa. One of the advantages of having very clear evidence is that, at that point, there is not much that you can argue with.

Now one thing that you might be thinking about at this point is how do you randomize, how does it work in practice and how do you go to people and say well can you randomly change your program and do they listen. It's a real question, because I think the first time I tried to explain to an organization that they should randomize their program, they thought I came from Mars. It took some time persuading. It took six months of long and very repetitive conversations to get to the point where they actually thought it was a good idea. The basic idea, once you know how to say it, is that it is actually very easy to do. The reason that it is easy to do is that actually very few programs go to scale in a day. Most programs expand over time and some expand even into areas with fairly large areas for a year, two years, and three years. Many programs, unless they are programs that are implemented bureaucratically from one office, tend to take a long time to be implemented, and because they take a long time to be implemented, there is actually a window of opportunity, the window of opportunity is that you can say look, you are never going to reach everybody in the next six months. So let's just pick the people you are going to reach in the next six months randomly. And that's politically often a very easy argument to make, because you know it's not going to reach everybody, it's actually fair that this is a lottery that determines where it goes. And people are often quite persuaded, that look this is a very fair way to do it. And then once you get past that mark then you are okay. That's the key point, once people recognize that its fair, then it is actually very easy to get. Sometimes you don't even need to do that, there are other ways of doing it.

This particular way of doing it is actually, using the expansion is a good idea for other practical reasons, that's almost always the best time to do an evaluation. If you're going to expand, that's when you want to know if it is worth expanding or not, or how you want to change the program. If that's the right time to do it; it also has the right property that you can get to know the answer. A match of the two is very good, and it is often quite practical to do that other ways of doing it. If I have time I might come back to those.

If I was giving this talk six years ago, I think this is where I would have stopped, I probably would have said more things, just to fill up my time, but basically this is where I would have stopped. And what would have happened at that point, I would have said randomized trials would be established as a useful policy tool, that would be good but that would be it. I think what has happened in the last six years is that it has entirely changed our understanding of what these trials are good for. That's one of the most heartening things that have happened, what I didn't expect was that in some ways these trials are now playing a very different role from what they were doing even a half a decade ago.

In some ways I think that randomizing has changed the way we are doing economics, and they have changed the way we are doing economics in a number of different ways. This is what I think we had not understood, maybe some people have had more insight than me, but I certainly had not fully understood the power of this methodology till it became available and people started using it. So let me try to explain what makes it really exciting, at least for me and I think for the community as well. The four different things that are done, one is an increment in measurement, and I'll say something about that, and another is asking a different set of questions to be asked, in particular a special case of refining our questions and finally challenging our theories.

So one advantage of randomized trials is that you give people the treatment, therefore you know who got treated. When you study most policy interventions you actually don't know who, you just know in this particular area this particular policy was announced in 2004. That's the kind of information you usually have. Whereas, in randomized trials, you are doing the randomizing; you know that this village got this on this day. What that does is it tells you who you have to measure, it is much better identified. And when you are looking at a broad policy that the government changed at some point or the other it affected a broad set of people. You know exactly who is treated.

Usually the problem of measurement is that it is expensive. And therefore you didn't know exactly who was treated, you would need to measure everybody in an area. Let's say this district got this policy in this year, I want to look at its impact. I have to measure everybody. Not everybody in the district actually got it, only a few people got it, but the policy moved to that district. That is all I can observe. I cannot observe who exactly were the people who were covered by the policy. Whereas, when I do randomized observations I change the policy, or whoever I am working with changes the policy, so I know exactly where it had an impact. That means I can pick where exactly I am going to do the measurement very precisely. I identify the people who I am going to measure much more precisely than I would in a standard policy analysis. I know exactly the people who are affected, that is very important

because then I can do very fancy measurement because I am not measuring a million people, I am only measuring five thousand, I can therefore proportionately spend more money per head in doing measurement. What does that allow? It encourages creative measurement; let me give you an example.

In post-conflict Sierra Leone there are a lot of projects going on funded by the World Wide Bank to help people in local communities learn to live together. This is after a long civil war, so the idea is you have dialogs between people, joint projects, etc. And the question is does it work? How do you know that people have been reconciled, it's a hard question. How do you measure reconciliation, if you go and ask people, it's hard to get them to tell you. If you ask people if you hate your neighbor, it's unlikely that they are going to tell you the truth. If you ask them do you love your neighbor, they'll probably tell you yes.

These are very hard questions to ask, so one idea that came up was the following; so instead of asking them these questions, why don't you basically, ostensibly independently of this reconciliation effort, give every village that is in the Treatment, when you are doing the reconciliation, something useful. The example that came up was a cassava grater; a cassava grater is something you use to grate cassava. You need it about once a week for a few hours, the rest of the time you don't need it. So now that you have given it to them, you can observe what they do with it. And you can ask the question, do they share this cassava grater with the people from the previous enemy group, or do they not. If they reconciled, they would share the cassava grater with people from the previous enemy group. If they are not reconciled, they would not share it. That's a kind of measurement that you could not imagine trying to do with a data set that comes from some government that collects a large data set on some generic intervention that happens somewhere. It's because we know exactly what the question is, what we are measuring, we know who are affected, and we are measuring it, that we can do it.

Second advantage of randomized trials which is related, we control the treatment. In other words, one of the things we can do here is we don't therefore just do things that make policy sense, usual policy analysis; you look at a policy change. When the government changes policy it doesn't change one thing, it changes how the school is run, the management changes, the textbooks change, the teacher training changes, a whole bunch of different things changes, because it makes no sense for the government to change one thing at a time. So in the world, the government changes things all together. When you want to learn, you want to learn which of these things worked. You want to know if whether it was teacher training that was important, or textbooks that was important.

How do you separate them? Well in normal policy all these things move together, therefore you cannot separate them. In randomized trials we can pick what varies, we can always choose, this set of schools will get only textbooks, this set of schools will only get teacher training, this set of schools will get teacher training plus textbooks. Therefore we can look at the effect of whether textbooks will generate education. Is it teacher training? Is the combination of the two better than each individually? We can ask all those questions, we can break the question up into the individual components. As a result, because we can pick the intervention and we can try it out very locally that encourages people to try out things that are

unlikely to succeed. Because you know that this is not some broad policy that has been implemented across a lot of places, it's something that you are controlling, it's very local, you are trying it out in one place so you know you are going to know the outcome precisely fairly quickly.

A study on immunization was carried out in India. Immunization for children is one of the universally agreed upon development deliverables. Almost nobody will disagree with that, yet if you look at the full immunization rates in this part of India where this study was done it was around four to five percent and that was a massive government failure. And when you discuss with people, there were two answers that people suggested. One was that the government system for immunization has collapsed, and the second was that people don't want to immunize out of some political, social or religious resistance to it. So in one experiment, you made a commitment to deliver immunization to fixed date in every month in a camp. If you were a sober organization which was making a plan, this is what you would have decided to do. In addition, because you could do everything, add little pieces and control them. They were willing to do one more intervention which everybody thought would not have any impact, or people were not so optimistic about, which is offering a kilo of dried beans to every mother who brought her child to be immunized. This is the kind of thing that if you were running a live program you would never experiment with, it seems so unlikely.

So when you give the camp the immunization delivered predictably, the rates go from four percent to fifteen percent. That's a three times increase or a bit more, maybe four times. If you would give them the dried beans, it goes from fifteen percent to forty-five percent. So in other words, the reliable delivery is good, the dried beans are superb. You get a ten times increase in immunization rates. As a result of delivery, this is the kind of thing unless you were in this experimental mode where you could control exactly what the treatment is, you could vary it locally, you would never be willing to do. It just is not cost effective to do it, you would not try it out. And because a) you try to do it, b) you could measure it, you can learn things you did not expect to learn. The other related points is that because we can ask very specific questions we can now actually know whether what we thought was a generic intervention, is actually a generic intervention or whether it is little pieces of amorphous matter.

Instead of asking should the government spend more on education, you could ask the question, should the government spend more on textbooks, or on teacher training? If we should be spending more on teacher training, should the teacher training be of this kind or that kind? You can keep refining your questions. And you might say why does it make a difference? Maybe they make small differences but is it worth learning then? The next picture is trying to answer that question. What the next picture does is that it asks the question. These are from different successful randomized trials, each of which was considered to be a success. In other words, they increased attendance. Each of them was aimed [at increasing] school attendance. This is the cost of a year of education induced by these interventions, each of which is deemed to be a success, so all of them had a positive significant effect on the outcome. If you look at the screen, what is remarkable is that the red outcomes, which are from Mexico, are on a different scale, they are so big that if I put it on this scale you couldn't see the stuff on the left.

These are all the costs of the same thing, what is the effect at purchasing power parity? So correcting for difference in the prices in different countries, the difference in cost of different ways of increasing school attendance, the cheapest costs three dollars and twenty-five cents, that's giving children deworming medicine, that's the cheapest way to get increased attendance. That's so small you almost don't see it. Then you go up and you see that it is stuff that is school meals gets it for you at thirty dollars and then you can go to school uniforms which cost a hundred dollars and then you can keep going. Then there is the Mexican ones that costs thousands of dollars.

The point I'm making here is that things that people think work, things where the results are all seen as positive, the difference of a factor of thirty between three dollars and a hundred dollars, is completely standard. That's a fact we didn't know. So we knew that spending money on education is important, where we want to go from there is to say which of these elements of spending money on education is really important, we want to identify that. That's a much harder question unless you are willing to do experiments. And because we can do experiments, we see that actually it makes a huge difference. If you got the right one of these we could save four times, ten times, fifteen times as much money, spend one fifteenth as much money as we would have spent otherwise. And we have done these for other things, increasing test scores, etc. You can get the same kind of thing a difference of a factor of five is easy, so in other words, in the world, some successful things cost one euro and other successful things cost five euro's and we need to know which one. These are not big differences, this is not a difference of spending money on education, versus spending money on health. This is within education, even within inputs, which inputs you should buy. I think we have changed which sense of details matter, how much they matter, and how we should look at this problem. It would be great if we found that it doesn't matter spend it on anything and you get the same return, but this turns out to be not true. It turns out to be completely false, you spend it on one thing you get a return that is five times as big as if you were spending it on something else. That in a sense has both a) undermined our confidence in the macro numbers on the subjects, and b) it has made us much more aware of the importance of the details.

One last thing, the last thing is that partly because of what I just said, that you can vary the treatment, you can choose what you are looking at and partly because when we do experiments, the interpretation of the results of the experiments per se don't rely on any theories. They come from just the fact that we randomize, we didn't need to know any economic theory. We would need economic theory to design the experiment, we would need economic theory to take away lessons from the experiment, but the experiment itself, whether there is a difference or not, just comes from randomization. As a result, it gives us a lot of power in challenging because in some sense it doesn't rely on theory to generate its results. We can use it to challenge even our most fundamentally held theories and sometimes we start learning about where our theories stop.

I'll end in a sense by telling you about one last experiment by some of my colleagues in the laboratory. This experiment is about fertilizer. They start with an experiment which is relatively standard: it shows that if you give fertilizers to farmers for free and you encourage

them to use it they get high returns. Hundred and fifty percent to five hundred percent returns on fertilizer, very big returns. Yet nobody uses them, essentially a quarter of the population uses them. One possibility was that people just didn't know about the results, that this is so good. That was the first hypothesis. You start with that hypothesis, testing theory one. You go there you do this experiment and you look at whether the people who live next to the person who got the five hundred percent returns, do they start using fertilizer. The answer is no. Doesn't seem to affect them. Then you can ask the guy who you had given free fertilizer first time do they try it out themselves the next time. They got good results. They learned, and the answer is no they don't seem to. There is some increase in usage. You get maybe from a quarter to about thirty-seven percent or something but that is it.

It's still the case that even among the people who have seen the results, you don't really see a big impact. These people go back to not using it. What do you think the problem is? When you ask people, they say we do not have any money, which is of course true. In some sense they are poor people, this is all in Kenya, and the farmers in Kenya are poor but nevertheless, the fertilizer is very divisible so you don't have to use, if you use fertilizer on ten square meters you get ten square meters worth of increased output. So if you can't buy a lot of fertilizer, you can buy a little bit. So that seemed implausible. They could have some money; they may be poor but may be able to save some of their money to invest in it. So it seemed like it was more of a problem about savings. So one standard theory in economics these days is that people have self control problems, they would like to do things but they can't, they can't save because something shows up and they feel tempted and they spend it. These people offered them a deal. They said, when you sell your crops you have some money, and you are worried that you are going to spend it by the time you need the fertilizer. So instead of buying the fertilizer when you need it, buy it now. We'll have a futures contract, where you buy the fertilizer now, but we will deliver it to you later. That way you don't use up the money. It turns out that the people said fine, okay, and it was very popular. Suddenly people started using fertilizer which seemed to confirm that people have a big self control problem. Except that it left a new mystery. People didn't actually do what people expected, they expected people would say ok I don't want the fertilizer now bring it to me on planting time, when I need the fertilizer, because if you give it to me now, I will sell it. I will sell it because I will have the same temptations I have always had. In fact they said give it to me now, they took the fertilizer, they didn't sell it. So at this point you can come up with a theory, and they have come up with a theory; but the point is that facts keep coming up that challenge our biases about the world. We start with theories, and the theories can be directly confirmed with facts, the beauty of this is that because in some ways we don't operate in a world where the theories are imposing structure on us, we can keep looking for things that might surprise us.

Just to conclude, I think we have learned a wonderful learning tool. We are certainly excited that we now have access to it. We are sort of like children who have found a new toy. And we are really incredibly excited about all of the things we could do with it, and I hope I have been able to convey to you some of the reasons we are excited. And hopefully some of you will join us and some of you will get excited with this, and even if you are not, thank you for listening to me.

Retour sommaire

Jakob SVENSSON, Université de Stockholm

Thank you very much. It's actually not that easy to be a discussant on a presentation when you basically agree on everything that has been said. But I will try anyway. So I am going to start off by reemphasizing some of the key points in Abhijit's presentation about why we should be more involved in projects and doing randomized evaluations. To me and to more and more people today, this is not so much a question, this should be obvious to many that we need to know more about how our project works, in particular in the developing countries where there are limited resources and all of the almost unlimited needs, and conflicted interests. It seems to me almost impossible to be able to advise policy makers if you don't know the impact of different policies. We might put different weights on these different impacts, but we still need the information about the impacts.

Of course there are different ways to do impact evaluations but as Abhijit has stressed, I think it's clear here that randomized evaluations have some very clear advantages. And one of these advantages that Abhijit mentioned is that when you do more of a traditional type of non-experimental approaches, the focus in the research is to a large extent actually trying to convince the reader or trying to convince the audience that you are actually measuring the impact in a correct way. So if I was to present the paper using non-experimental approaches, maybe fifty percent of my presentation would deal with that particular question. But if you do randomized trials and you do them well, that is not an issue anymore. So you can focus on analyzing and understanding the problem at hand as well, and the discussion that will follow will be a very different discussion.

So a discussion using a non- experimental approach will deal a lot with the way the data has been analyzed about identification techniques and so forth. Why the policy to be made has been based on a randomized trial will be more about what the impact of this study really applies. I think that is a very important point that Abhijit mentioned. I think the second big point is this focus on new questions, trying out new ideas that have not been tried out in a systematic way. I think for that reason today development economics are pushing the research frontier in various dimensions forward, which typically was not the case if you look back some ten years ago, when development economics was more applied using research tools from some other parts of economics. This is the second key point that I want to mention. I would like to add a few things as well. Abhijit made this comparison to medical trials, and the impact evaluations of social policy done by economists in particular use the same "gold" standard method as is done in medical field trials, but the focus is different. I think that is important to stress. The focus when U.S. economists and social scientists study these types of questions even in the health sector is on some behavioral questions rather than the direct impact of a particular drug. For instance, there is a huge amount of community-based randomized field trials in medicine. They basically address the impact of the drug or some treatment practice, and if I simplify somewhat typically, those evaluations are done with the constraint that the hospital workers are actually competently doing what they are supposed to do. And the economists compare its advantages to ask the question, how do we ensure that these health workers are actually competently doing what they are supposed to do? And in

both cases, the input can be health outcomes or health utilization. I'll come back to that a little bit later.

Another thing I want to stress is that randomized evaluation were, at least a few years back, predominately used to assess impact of a specific input, or a mix of inputs, broadly defined, it could be textbooks, it could be drugs, it could be additional teachers or, as Abhijit mentioned, potentially a combination of them. But going back to what Abhijit started off, the differences between process evaluations and impact evaluations, you can also do impact evaluations on processes and Abhijit did mention one of these, the conflict reconciliation study, there are also a couple of other studies being done, and presently are on the way. It's just like participation on grassroots monitoring and similar process type of policies. And more broadly I think while the focus in the beginning was very much on input it has shifted to study more incentive types of problems, like the study on fertilizers in Kenya for instance.

Let me give you one example of one such process evaluation, one involving myself. The latest figures are that approximately eleven million children under five die each year, about half of them die in Sub-Saharan Africa. And more than half these children will die of diseases that could have easily been prevented or treated if the children had access to a small set of inexpensive services. It is important to stress that these services have been proven to be very effective in community based medical trials, but the problem is they are not used in a systematic way. So why is that? There are different answers, one is lack of resources, one may be a lack of knowledge that the health workers don't know what they are supposed to do. But another, I think, has to do with processes or local institutions. Maybe it's in effect a system monitoring or weak accountability relationships. So in this study we were trying to address the following question: can we strengthen the users' role in a rural setting in Uganda and have them participate more actively in monitoring health care providers. If we can do that, will it have an impact on health outcomes? So we connected randomized field experiments where, in just one sense, local NGOs information about how things seem to be working, and how things should be working according to the government set plan for service delivery at this level. And then these local NGOs also organized accountability meetings in the communities. One year after this type of meetings were organized, we came back to the communities and collected a new round of information to be able to assess the impact of it, and we find strikingly large effects: a year after the intervention, the treatment are more involved in monitoring their providers, and the health facility staff appear to exert higher effort to serve the community, and most strikingly we find large effects on both utilization and health outcomes. That actually compares favorably to some of the more successful community based trials in medicine.

As I said, the difference here is where we focus on trying to get health workers to do what they are supposed to do, while the medical trials focus on putting new inputs on new treatments, given that the health workers are doing what they are supposed to do.

Let me end by a couple of things that I think are important to keep in mind, at least when you are arguing for a more evidence-based of policy agenda. I think, at least if you look a few years back, this legacy there could be bias in the type of projects being evaluated to which

projects that are to a process easier to evaluate. And that doesn't mean of course if you are interested in basing your policy on things that seem to work you cannot only look at projects that have been credibly evaluated but you have to, of course the argument is you run around and do other things that you know anything about, but rather that you need to also do evaluation on the broad type of different types of policies. For instance, in social service delivery, when it comes to education, health and so forth, there have been a number of studies on adding new additional input such as changing the mix of races at schools. Abhijit talked about some of these, and the cost benefit analysis of these different interventions, interventions that happen at the school level.

In some sense maybe the biggest problem in many developing countries is not the lack of resources or the mix of things at the school or health level. It could be, as you saw in the cost benefits estimate that Abhijit showed, that there are some very cheap ways to actually boost for instance school enrollment, but there could be other problems that has to do with more processes or institutional failures in developing countries. I think that the one of the key constraints that if you look at the impact of policy, and policy in social service sectors, has to do with implementation, and implementation at all levels, implementation at the school level or at the clinic level, but also implementation of the project or of the policy at different levels of the government hierarchy. So why is it the case, even when inputs are available, the teachers and the health workers are not present at school, or even when they are present they don't teach or treat patients?

And moreover a country's ability to improve service delivery outcomes, as I mentioned, is not only determined by what happens at the school level, but sometimes and maybe even primarily it's determined by what happens at other different layers of the public service delivery system. And one implication of this system is, since implementation of social service delivery is often problematic in many countries, often plagued by inefficiencies and corruption, interventions that focus on improving governors of social services, maybe them in particular, may be a very cost effective way to improve service delivery outcomes. The good news is that such governors interventions can also be evaluated with randomized trials. This is not against using experiments, but rather that we need to take a broad approach, a holistic approach to what type of interventions we should evaluate. And as I mentioned there are already a number of these evaluations that have been done, and more in the pipeline. Thank you.

John MARTIN

C'est le moment de quelques questions dans la salle. Je vous demande de vous identifier et de poser une question assez concise, sans faire de long commentaire.

If not, let me, at least, put a question to our two speakers. I think it's been clear that, in the world of random assignment and experimental methods, there is more interest now in the question of how things might work or not work. But is it so easy in fact to always design satisfactory random experiments where the issues are more complex than simply just a two-by-two comparison of say, whether I put money into school uniforms or whether I put it into school meals, I would get better outcomes? Let us say, for example, I am looking at the

design of a system of employment services to job seekers, where there is an issue of sequencing and there is an issue of the different roles that case managers and the job seekers themselves play. So I can end up with very complicated multiple-choice questions about sequencing and design and it is not as clear to me, by the way, that random assignment or experiments are necessarily the most effective way to try to answer those more complicated questions about why things turn out the way they do. So I will be interested in hearing your responses to this question. First Abhijit, then perhaps Jakob, would like to say something. In the meantime, I hope there are some more questions from the audience.

Abhijit BANERJEE
I agree with half of what you said, I think that it is certainly true that the more complicated the question the harder to design an experiment. On the other hand, more complicated question in the particular sense you are describing, the harder to learn anything from any non experimental data. In this sequencing after all, if it varies, first it doesn't vary a lot in the world because there is an accepted model of doing things, therefore you don't see very much variation. And then if there is variation, you don't know whether the variation came from some management initiative which also changes fifty other things. So I take your point but I don't know if there is a choice, I don't know there are really wonderful non experimental methods for answering these questions.

Jakob SVENSSON
I agree with Abhijit's point. I maybe would add that one of the advantages with randomized evaluations is that you can actually experiment with these types of purposes. We have the same problem in medicine, when you look at evaluations of drugs and many other treatment practices. It's actually the case that medical researchers don't really know the mechanism, don't really know the exact combination of drugs, so they use experiments to try it out first on non-humans, of course. But then when they are somewhat more confident they turn it into impact evaluations on humans. In principle I don't see why the same approach cannot be used when it comes to social experiments.

Etienne EISENMANN
Vous avez fait une comparaison particulièrement convaincante avec les expérimentations pharmaceutiques. L'expérience montre que même quand on a fait une étude sur cinq mille patients, lorsque l'emploi du médicament est généralisé et que cinq cent mille patients ont été traités, des effets secondaires non connus précédemment peuvent apparaître. Alors, en matière sociale, est-ce que vous disposez d'un recul qui permette de comparer le résultat de l'étude expérimentale aléatoire, basé sur des méthodes aléatoires, avec la généralisation à très grande échelle à toute une population ?

John MARTIN
Thank you very much, I think that is a very pertinent question about how one goes from small-scale demonstrations, to general, population-wide inferences, particularly when one is talking about social experiments. Abhijit and then Jakob, if you want to add something.

Abhijit BANERJEE

In some ways that's the core challenge, not just of randomized trial but of all social science research. Wherever you think you have reliable methods for assessing policies through the variation of policy that allows you to study that is almost always relatively local. There is no perfect solution to the problem. Clearly, you want to know whether this result would generalize to a larger population before you take it to a larger population. One advantage of experiments over other methods is that with most other methods, you don't actually know who are the people whose behavior or whose particular outcomes were changed by the policy. You can't identify them. That is the same point I was making before about measurement. Whereas with experiments you know actually the affected population, that's a big advantage. You can start by saying, look, this result is exactly for this population. Which population I can characterize in a great deal of detail, the advantage is I can start thinking of process by which, okay then I can say, well, the reason why this may not generalize is that this population has these characteristics. And I can look for other populations with different characteristics through the experiment, part of the advantage is knowing what the limitations are, is that we can guide that process in a specific way. That's not to say that if there are sometimes opportunities to do a large scale reliable non experimental study, when that happens, which is rare, I would say those are great and would solve the problem of scale, in that it is absolutely true, they are very rare, to the extent that they are available. I wasn't about to argue against them, it seems there is in a lot of places the trade off is between something which you may not have any faith in, and something you have some faith in but maybe only true in this particular population. But you know what this population is, and knowing the population gives you a lot of power. Because then you can start to say, well if I had to generalize it to these people what are the key differences and you can build on that. I don't know that there are any perfect solutions; this is a very old problem in social sciences.

Jakob SVENSSON

If anything, I think the importance of replicating studies in different contexts and settings would be one way to, at least, try to get at that issue. And I think that is what is actually being done. Some of the interventions that Abhijit talked about some of the simpler interventions have been actually replicated in different settings. But it is a problem.

Un intervenant

My question concerns the black box that you sort of defended, saying there is no theory behind this. In a sense you are just testing an experiment, but is that really true? And I was wondering why so many who promoted these methods are against a theory based approach, because looking at for instance at the grassroot example you still have a theory that there is a problem in the incentives. If you look at the Progresa program, which is a very good program, and very interesting results, you have a theory of cash transfers to mothers would actually bring children to school. If you look at the warming example, you have a theory that more healthy children actually learn better, so why this opposition against theory based approaches? Thank you.

Abhijit BANERJEE

I think quite the contrary, I'm a devotee of theory. It's the fact that the empirical identification of the result doesn't rely on a theory, it makes it more useful in testing theories. The problem often in testing theories is that you cannot maintain an assumption of a theory, which is sort of how you interpret the data. Therefore there is a significant part of the theory you cannot test because that is the maintained assumption. So I, as a devotee of theories, particularly like the fact that you can test theories without assuming other theories. Absolutely in Progresa there is a very specific theory that in a sense is being tested. What is an interesting exercise there is that the design of the test did not require that theory. The theory that we tested, we take as given that the experiment results are reliable in respect to other theories. You could have multiple theories there of why Progresa has an effect, and we could start designing building on the original Progresa experiment variations to figure out which of the theories. The reliability of the estimate doesn't rely on a theory. That's precisely why I think as somebody who's a long time devotee of theory, its exciting in a sense, in a sense maybe I didn't speak. But I am sure I certainly didn't intend to say that this is against theory. Its precise attraction is that it liberates theory from the weight of being used to identify empirical results. That makes it less testable.

John MARTIN

Thank you very much, this now closes the first session this morning.

Retour sommaire

John MARTIN
Thank you very much, ladies and gentlemen. We are delighted to have as our next speaker, another MIT economist, one very well known for working both in the fields of education and labor and social policy, Joshua Angrist. After Joshua has spoken, Marc Gurgand will make some comments, and then hopefully, we will have a few minutes for questions before lunch.

Expérimentations en économie de l'éducation

Joshua ANGRIST, MIT

Thank you very much. Thanks to Esther for organizing this and to the hosts. It's wonderful to see so much interest in randomized trials come to Europe and France. I'll take my time to talk about some specific projects that I've worked on. Let me just say by way of introduction, I'm a labor economist, and labor economists are often interested in education, which we see as falling under the general heading of human capital. And as part of that agenda we look for ways to make education more effective, more efficient, cheaper, and we also study the economic returns to education which is a big part of the human capital story, the earnings consequences of additional schooling or higher schooling or changes in school organization. And I've done a number of additional projects that are related to that. A little bit unusually for me in the last decade or so I've had the opportunity to do some actual randomized trials. And the things that I'm studying in these trials are what we call incentives for achievement, achievement awards, or achievement incentives.

When I started working on this about ten years ago, what I'm talking about today is culled from two projects, two really ongoing projects. One with Victor Lavy a colleague of mine in Israel, and one with Phil Oreopoulos and Dan Lang, colleagues in Canada. When we started on this, particularly when Victor and I started on this, in the late ninety's it was very controversial. The idea that you could get better outcomes in schools by directly rewarding students, but this idea is now catching on, probably not because of our research but hopefully we can claim a contribution. The idea that we want to make school work pay is growing, and one of the best examples is coming down the pike shortly, a very ambitious scheme in New York City schools that will pay students for performance in elementary and middle schools.

Another thing that is in the same spirit, is a plan to pay students who take AP tests, AP tests in America are Advanced Placement Exams that high school students take to get college credit, and a big concern there is that already non white students are not likely to take those tests, so there is an interest in boosting that. And Dallas and some other cities are piloting programs to pay students to take those tests. The kind of mother program for these incentives programs in some ways, though it doesn't involve random assignments, it's not really where the evaluation comes from, but is part of the policy background, is what we call the Georgia Hope Style program. Georgia Hope is a program that many American states have imitated which gives tuition scholarships at state schools, that is, public institutions of higher education to students that do reasonably well but not necessarily great. They might get a B average, which is pretty good but not outstanding. This is an innovation because schools and universities have

always rewarded—this is true at the high school level and the post secondary level—have always rewarded very top performers with scholarships and prizes. But the new development here is to push these awards down to the sort of good students who meet the standard but do not necessarily excel. So that as I say is an innovation.

The other thing that I say is that these programs are increasingly likely to talk explicitly about financial awards without necessarily, what I call, the fig leaf of scholarship. We give you money. From an economic point of view that's just fine. Here is the first of the experiments that I'll talk about and I really don't have time to go through every number as much as I love these numbers and I would like to show them all to you. But I'll try to give you the flavor and the highlights of each of the experiments so I'll race through some of the stuff, skip through some of the stuff. The first experiment, as it says on the slide, an update because we have been pursuing it for long time and we have been getting new results, was a randomized trial that Victor Lavy and I did in Israel for achievement in high school and the reason its an update is for a long time, we didn't have any post secondary outcome. In other words we had the outcome we targeted which was a matriculation exam that is very important in Israel that I will talk about in a second.

But now we have some new data that follows our subjects over a longer period of time and whether they go to college and whether they applied for further education. I'll show you some of that. And the other one is the project with Phil Oreopoulos on a combination of incentives and services and financial incentives and academic support services for college achievement. So that's for freshman students. Briefly to highlight some of the interesting findings that come out of this work, one very interesting thing is that girls react to these programs and boys do not, and that was very clear in both of our studies. And it seems to parallel a fairly common finding that girls are wiser and more likely to take advantage of these relatively generous incentives, whereas boys press on and do what they do, and we're not quite sure on how we can get them to do anything different, at least they don't seem to care very much about what we offer them. Other findings related to program design are whether the subjects have a kind of a plausible chance of achieving an award. If the awards are too far out of reach, the motivating power of the intervention seems to be very weak. That's probably not surprising. And something we have in the post secondary study is an interesting variation where really the financial incentives have to be combined with some kind of support services to get anything.

And then I can make the broader point about these interventions, which is the reason we know so much about them in a kind of varied and interesting way, boys versus girls, the structure of the programs, is because we have done these experiments, they are well-controlled trials, reasonably well-designed not perfect and they allowed us to compare some things and give us a sense of what works and what is worth pursuing. Of course we don't stop here, but we take the findings from these studies and then we try to check it, it could be a chance finding, maybe something went wrong, and we would like to also refine it and think some more about policy design.

Let me now describe a few of the specifics on the study that Lavy and I did on high school. This may be most relevant for French policy makers because Israel has in its post secondary system some parallels with the French system, it's closer to the French system than to say, the American system. Because Israel has, like the French *baccalauréat*, a matriculation certificate which is the key hurdle for high school students and if you want to go to University and get certain types of jobs. Israel also has compulsory military service, to get a better placement in the army you would want to have your *bagrut* which is the Israeli version of the *baccalauréat*. There are enormous national disparities in the likelihood of getting that certificate. About half of Israeli high school seniors end up getting one, but there are big regional disparities and there are big ethnic disparities. Israel has a large Arab minority where *bagrut* rates are very low, and then among Jews there are Jews of European and Asian African origin. Jews of European origin are much more likely to get their *bagrut* than others. So there is a policy interest in trying to improve rates for groups that don't do very well. And the government has spent a lot of time and money; some thought maybe not as much thought as time and money, trying to come up with programs that might increase *bagrut* rates. Most of the programs that the government has tried, Israel has tried, is remedial services, support services of various kinds, some moral suasion, and that sort of thing. Nobody really knows what works, because most of that stuff does not get evaluated, but there is kind of a casual sense that, not that it has done any good, because these disparities persist and seem to change very little from year to year or cross-sectionally.

Victor Lavy and I proposed the simple idea that we would just pay people to take this exam and then we would pay them a lot if they actually passed it. And the reason why it is a plausible and interesting thing to do, is if you look at the data, and I wouldn't be surprised to see something similar in French data on the baccalauréat there would be a fairly large group, not everybody who fails of course, but a fairly large group where you say that kid with a little bit of extra effort might have passed that exam. He was within spitting distance of passing the exam. If he had studied a little harder or something else had gone his way, that person would have gotten certified. About twenty percent or thirty percent, depending on how you do it Israeli *bagrut* takers would be in that kind of marginal category. It makes sense to think about not only services but actually motivating the student to clear that hurdle. That was our plan. We did a number of demonstrations. We did a small one which was a random assignment at the individual student level, which didn't show much but it was a very small project and had some statistical issues. Then we did a larger one which was a school-based randomized trial, so that's what, I don't know if this was mentioned earlier in the discussion, that's what you would call in medicine a group-randomized trial or a community trial, where the unit of observation is a cluster instead of individuals and in this case the units of observation for random assignment were high schools and the target population were the forty worst high schools in the country, based on their 1999 *bagrut* rates. These are schools that on average have a *bagrut* rate of about twenty percent; we threw away schools that had no *bagrut* takers ever, because perhaps they can't be helped. But there was this fairly substantial number of schools to draw on where the *bagrut* rates are low but not hopeless. And so the way random assignment worked in that case, is we paired schools based on 1999 *bagrut* rate which is ultimately the target of the intervention, that's the thing we want to get higher, to get more

kids to get their *bagrut* or their *baccalauréat*. And then do a random assignment, which amounts to tossing a coin within the pair. This is common in the field of group randomized trials to actually do some matching and then to do the random assignment. There are various technical reasons for that. There was some heterogeneity in the population that we studied, I'll just mention, that in the forty schools, ten are Arab schools and then in Israel the Jewish system is divided along religious lines. There is a secular public system and a religious public system; we have some representation from all those strands in our study.

The other thing that is interesting about the study design which often comes up in randomized trials and sometimes is raised as an objection to random assignment is that you can't get everybody to play along, and that is certainly true. In this case, the way we did the random assignment is we had administrative records, of course we know what the history of outcomes of these schools is and we picked the schools for random assignment. Once a school has been selected for treatment, we then have to approach the administrator of the school, the principal of the school, and say would you like to be in this study. Would you allow your students to be in this study, where we are going to pay them if they succeed in getting a *bagrut* and that is indeed what we proposed to do. If a school was treated, every student in that school would be high school seniors, twelfth graders, who gets a bagrut will get a check for six thousand shekels, which was about fourteen hundred dollars. And some principles did not want that for whatever reason they object to it on principle, or there was an administrative hurdle which had to do with the level of cooperation, the way they interacted with us, they had to provide us with certain records and stuff.

Of the twenty treated schools, five did not want to be in the study. And from a research point of view that is just fine, the way we fixed that is with instrumental variables where we used the original random assignment as an instrument for the compliance. In practice, we basically can ignore that we understand that the treatment effects are diluted by non-compliance somewhat, but there is no reason why compliance issues, or voluntary opt-out, either in part of individual subjects or on the part of a cluster need compromise a randomized trial. That is something that's well understood today, that probably wasn't understood as well say thirty years ago at the dawn of this kind of social experimentation.

So let me skip ahead a little bit. Just to give you a flavor—there are way more numbers in this talk than we will discuss. This just gives some descriptive statistics you can compare here. The national *bagrut* rate is higher than the fifty percent that I mentioned before, for reasons that have to do with when it was measured. And then here you see things about our schools, a third of them are Arab, and ten percent are religious. And this is some of the family background of the kids. These are lousy schools in terms of their socio-demographic and socio-economic characteristics. You can see the average education of fathers is ten years, whereas Israel nationally is twelve. These kids come from bigger families than the country overall, and so on. You can see that it's a worse off population that we're studying.

The basic results are reported here in the form of Treatment-controlled differences that have been adjusted for co-variants, which isn't really necessary because of random of assignment, but for various reasons, it turns out to be a good idea. And then there's various sets of co-

variants, some school-related, and some kid background-related, like the education of the father, and so on, as controls. None of that seems to matter. The key finding is here. Here's boys, here's girls, these are just two different types of estimators. You see there's a 20% *bagrut* rate for the boys. You see there's a controlled difference. It's a very small number, and that's the measure of the statistical significance, so, for those of you who remember how that works, you'll see that these differences are not significant for boys, but for girls, there's about a ten percent difference, depending on how you do it. A little more, a little less, in favor of the Treatment group, and it's a little more than double the standard error.

So that's a key finding right there. The treated girl rates go up quite a bit. Overall, the treatment doesn't do much, because it's a mixture of ten percent on girls and zero percent on boys, but if you look separately at the girls, you start to see something. And something else that's interesting here is that if we break it up by who is doing relatively well in the previous year, and this relates to the fact that I mentioned before that you have to pick people who have some hope of responding to your intervention. These are girls who are doing relatively well. They're in the top half of last year's promotion and they get a very big Treatment effect, and their overall chance of succeeding is fifty percent. The group that's in the bottom half, they're really hopeless, they can't be helped, incentives will not do anything for them, so there's no Treatment effect for them.

So we kind of succeed in finding the right subset of the population. And in some sense, we should have thought about this a little more ahead of time and designed the experiment a little differently. One of the reasons that you do experiments is that you learn a lot about the process that underlies what it is that you are studying and you do a better job next time.

So one other thing I'll show you about that—this is a key question for people who study incentives. Economists weighed in and said we are going to pay people to do this or that, we are going to pay people to do well on tests. And sometimes that works and sometimes it doesn't, but if it does people will then ask you, you know what did you accomplish by getting somebody to do well on a test? The test, per se, is not that important, now in this case the one response to that, and I think this would apply with the French *baccalauréat* is that this test actually is important, it's a key screen or a key hurdle or credential that you need for other important things. But it might be that if I pay somebody to get this credential, they'll get the credential and stop there, because their not really motivated. And of course I'm not continuing to pay them, the labor market may reward them, but I'm not rewarding them anymore. So we would like to show that this particular intervention is worth something to you later on. That is a little harder to do from the view of the researcher because you have to wait a long time to find out what happens to somebody later on. You have to have some sort of scheme to track people. This was a problem for Victor and me because we basically lost track of our subjects. They were in school, they were in high school in 1999, and to get kind of the "later on" outcome in Israel, you have to worry about they finish school, they spend a little time goofing off, then they go in the army, they go in the army from three to five years, they spend a little more time goofing off, then maybe they go on to college.

So there is a long period, but luckily this paper was rejected in a couple of journals, so we had time to keep working on it. And by the time we got a more favorable response, and people were a little more sympathetic to this kind of thing, enough time had passed so we could ask whether these students were likely to go to college. The other thing that we figured out, which we didn't realize when we were studying this, most of the data in this experiment, or administrative data, the *bagrut* of course is this big national test, its not based on a survey. It's this thing that the education ministry runs, so anybody with an ID number, you can find their *bagrut*, so that's inexpensive and attractive because you don't lose anybody and it doesn't cost you anything once you get the cooperation of the relevant people. But we didn't know that we could track people into college without surveying them, we were having a hard time finding them, and then we realized that the Israeli National Insurance System knows who is in college where, for various reasons that have to do with the way social insurance taxes are paid, and so very recently we were able to get that data. So for a very low cost we can now find an interesting downstream outcome which is if we paid you to pass your *bagrut* did you actually end up going to get more secondary school? And we would be very happy to find out if that is true, and so we made that link. It turns out that there is some evidence that it is happening.

Now you have to look at it a little bit the right way, of course its only for girls, there is no point to look at boys because boys didn't respond to the first thing. The boys were not more likely to pass the tests, so nothing else of interest happens to them either. So we are done with the boys. But the girls, particularly the relatively high scoring girls, in our low scoring population, the relatively high scoring girls might end up doing better. And there is some evidence that they do, this is not an elite population so they are not going Israel's elite research universities, Hebrew University, Tel Aviv University and so on, but they are more likely to go to some kind of post secondary program and we see that when we measure it using a fairly inclusive definition (down here). This includes teachers colleges, and what are called practical engineering colleges which might train you in with some certificate in software, as well as all academic colleges, including non research.

For the relatively good girls, that is the girls who are categorized as most likely to have responded to the program, there is a Treatment effect of about twelve percent in this schooling outcome. So it does seem that for a substantial portion, one way to describe it, how many of the extra *bagrut* certificates that we created by paying students to get them—w e paid students to pass their *baccalauréat*—how many of those bac were converted into enrollment into higher education? And turns out that number was about two-thirds or three-quarters, so that is pretty good, a pretty good yield. Of course, later we would expect to find, though hopefully we won't have to wait and keep working on this paper, but we would expect to find this education will also translate into higher earnings. That's a flavor of the way my colleagues and I look at that kind of experiment, the way we do the analysis, and the way it's designed.

Let me now describe something else, which has some similarities and some differences, this is a randomized trial in the post secondary environment. And we did this at a state school in Canada. Again, not an elite institution, though it's a very different population than the Israeli population, because these are college students, but it's not a very selective campus. It's kind

of more of a community college, if you know what that means in the North American context. A big concern in this kind of environment is that students in college, this will not shock any professors of faculty here, but students in college don't always finish on time, and sometimes they fail courses and they have to take them over, and they spend years and years in college. I'm sure this doesn't happen in France, but it happens in Canada, that people might take more than four years to finish their B.A.

In fact on campuses like the one we are studying, this is also true in public schools. In the United States, even though it is supposed to take four years to get a B.A., after six years only about three-quarters of the students have succeeded in doing that. So there is a strong policy interest in making that happen more quickly, reducing drop-out rates, getting students through the system quickly. There are various reasons why you would like to do that from a simple economic public finance point of view. You are subsidizing time in school for these students, so it's costly that they take a long time, but arguably it is not so good for them either.

The traditional approach to retention and failure to progress in a timely manner through college is a series of academic support services that almost all campuses in North America offer, where you get supplemental instruction, you get some advising and that sort of thing. It's relatively uncommon to offer the broad mass of students in those schools some kind of financial incentives to move along, there aren't many programs like that. There are programs that offer scholarships, again to relatively high achieving students; the best students of course have access to financial aid which is to some extent conditional on their performance. So we introduced a multi-pronged intervention for this population which included both incentives and services. It was a relatively complicated intervention as opposed to the Israeli one. And the Israeli one in the end we were just writing checks or somebody is writing checks, but we don't actually have to worry about providing services here. Here we ran a kind of academic support system which supplemented the existing services. We called it the student support program. It included both additional classroom instruction time, that's kind of in the mold of what is already being done on many North American campuses, and it included a program that was a little more unusual, which is peer advising. That is where upper classmen, older students are put in touch with new students and make themselves available, in a very active way. They email the students; how are you doing, can I help you, and so on. Every few weeks, they try to arrange meetings and give advice. The subjects here are freshmen, they are incoming students. So that was the service side of the program, and then the financial side of the program which was what my coauthor and I were most interested in. The findings come out not quite the way we thought they would. We called it the student fellowship program. It's a program that offers merit scholarships for maintaining a solid GPA; GPA is your grade point average, in your first year of school and progressing on to your second year of school.

So we had three treatment groups. Let me skip ahead. One group got services, one got incentives and one got both. So let me talk a little bit more about the incentives. The incentives were anywhere from one thousand to five thousand Canadian dollars, and Canadian dollars are now better than real dollars, because the dollar is so weak, the U.S. dollar. So quite a large fellowship amount for meeting your GPA target, the GPA targets that we set for this were broadly speaking what you need to stay in good standing academically.

But we adjust the targets a little bit based on where you were when you came in. If you came from a weaker high school background, if your GPA in high school was lower, your target was a little lower. We did that to save money; we had higher targets for the best students. So the expected award rate was about seven or eight percent for the five thousand dollars and about a quarter of the students would get a thousand dollars based on past data. And there's a chart there about how the awards were set. The awards were specific to what kind of student you were when you came in terms of your high school.

The way this experiment worked, let me talk a little bit about the design, I don't have too much more time, so I won't go again through every detail. The experiment, again we make maximum use of administrative data, we don't want to do too much surveying, though we did some. We identified the newly admitted students with the cooperation of the institution that's hosting the experiment, of course. We know who is coming to that campus and we are authorized to get in touch with them. And we take everybody who's coming, who's not, in the upper quartile, so the really top students are not eligible for our experiment because too many of them are going to get awards and it's going to get expensive. But we still have a large population that we can study. And the random assignment took that population. We offered two hundred and fifty of those students, so here the random assignment is at the individual level, not clustered.

Two hundred and fifty are offered the service package—this is in addition to whatever services they get—two hundred and fifty are offered the awards and that means that they are told if they do a certain level of achievement, they'll get paid. Of course, they are not offered awards up front, but they are promised awards, and a hundred and fifty are offered both, we call that the SFSP. Everybody else is a control and that is broken up by random assignment in the beginning so the groups are comparable. Again this program has the feature, all that we are randomizing is the offer of Treatment, we can't compel you to be in our study. Of course that wouldn't be ethical, even though the study is not very burdensome, but even so. If you want to be left alone, we'll leave you alone.

The way that it worked is you were randomly selected for treatment, and then we contacted you as you were doing all the kind of electronic paperwork of being a student. When you are a new student in a university, you know you fill out lots of forms, which is now done mostly on the Web of course. And the intake for our experiment becomes part of that process and you have to then sign a consent form saying 'I'm willing to be in the study'. And you're told which of the three treatment groups you're in. The controls are told nothing, the controls know nothing of the study. And this already generates some very interesting results; many people want to be left alone. If we offer you the services, fifty-two percent agree to be in the study; if we offer you a promise of a fellowship eighty-six percent agree to be in the study; and if we offer you both seventy-six percent agree to be in the study.

You can interpret it in any number of ways, clearly people are being offered extra services are quite likely to say leave me alone, maybe they don't understand that it doesn't obligate them to use the services. People who are offered money, for the most part are happy to be offered money, but some of them also want to be left alone, and if they didn't agree, so there are

fourteen percent of the people who were offered money, who don't agree to be offered money and later on if they get high grades we don't give them money anyway, too bad for them, but there aren't that many of those.

What's interesting is you can get people to agree to services by offering them money, so in the double package the consent rate goes from fifty-two percent to seventy-six percent, when you offer both services and incentives. On the other hand you could look at it and say people are less interested in money if you also threaten to give them services. Another interesting thing here that is key for us, and this relates to the boy-girl thing I was talking about before.

Girls are always easier to deal with than boys. Here, the consent rates are always lower in every Treatment group for the boys. Forty-five percent service versus fifty-seven percent for the female. Here seventy-nine versus ninety-one. Seventy-one versus eighty-one. So girls are much more likely to accept the prospect of having extra services or agreeing to some kind of financial compensation for doing well in school. So that's an interesting pattern which foreshadows the results of our study. The results of this study can really be summarized in two ways. First of all, these are Treatment effects and up here you have descriptive statistics for the control groups. This is a score of zero to a hundred and these are the standard deviations of that score. Services alone did nothing. Fellowships did something in the first semester. This is a small significant effect. The combination of services and incentives has the biggest effect.

But, again, boys got nothing out of this program on any dimension. They're all zero here and insignificant. It's the girls that responded and it's particularly the girls who responded in this group. This is the combination of services and incentives. So we have a nice result in the sense that it's clear. But again, it's discouraging, because we haven't done anything for boys and girls anyway tend to do better in college and in North America today there are way more girls than boys in college. So it is the boys in some sense that are the challenge, they're the ones that need to be helped.

The other thing that happened that's interesting—and we're following up on this—now, this is a different outcome, this is the first year GPA. The early incentive effect—this was in the Fall term—basically disappeared for girls as well, and the only program which had a lasting effect is the program which combined both services and incentives.

On the one hand we helped only girls, and we only helped girls in a way that kind of carried forward to their GPA's, when we provided additional support services as well as financial incentives. The last piece of that story is that boosts to female performance from that combination actually persisted beyond the program, so after the program ended and we were no longer paying people to do well in school, the girls that got that combined treatment continued to do well in school. So that's a nice finding that is related to the finding I mentioned earlier, the Israeli study where we had a long term outcome, because what it showed was is not only did we pay people and motivate them but in response to our payments they acquired some sort of skills that have to do with schoolwork or organizing time or something like that. Or maybe just confidence, but whatever they got out of it, stayed with them and they continued to do better in their sophomore year even though there was no

incentive system in place anymore. So I think I'll finish there, that's just a quick taste of some of what I do in my spare time.

John MARTIN
Thank you very much indeed Joshua. We now turn to Marc Gurgand, from the Ecole d'Economie de Paris.

Retour sommaire

Marc GURGAND, Ecole d'économie de Paris

This is obviously a very interesting set of papers which has strong but intriguing results. The first thing I want to say maybe is to put this into a general context, which is that of schooling policies in general where I think there is a lot of debate and general uncertainty about what are the most effective policies. It is a debate in the political area but it is also an academic debate. I think this is a very typical of a situation where it is important to have robust evidence, transparent and convincing evidence. Therefore, this is certainly a situation where random experiments are very welcome and the example of Progresa was given early this morning. This is a very brilliant example where you can make the point that some things work and this has influence because it is made through random evaluation.

This is really the context where this set of papers stands, and maybe a first remark for the audience is that it is about whether financial incentives generate effort at studying. Maybe it is worth saying that I'm sure that in this country this is not a policy that would be considered very seriously. It seems to be a strange policy to many people here, so just one remark, in defence of the very question of the paper. Maybe you could think that because people who go for studying, especially in the States where they pay high school fees, had to pay this high cost, they must be very serious about studying, or they are crazy. And then the policy is really worthless.

I think its useful to point out that it is not that clear. One reason is because you pay the cost and then you wonder whether you're going to make the effort or not. And whether you make the effort or not, the cost is paid. This is sunk cost, whether you make the effort or not. And the effort you have to make is now and for sure, whereas the reward—like failing or passing the exam—is later on and is uncertain: maybe you make the effort and you fail, maybe you don't work and you have the exam anyway. So there is certainly an arbitrage and an optimal amount of effort. Thus, there is no reason to think that incentives might not be able to increase this optimum. So I think that it is a very reasonable question and I'm sure with a bit of introspection on yourself you might think that this sort of behavior is not totally absurd. So, this is surely a good question that you want to ask.

Now, if this is such a good question why would you want to estimate the impact of this policy. Well of course there is the matter of elasticities. It is rational to react to incentives, but maybe you react a lot, maybe you react a little, so you still want to measure this. And also there are alternative interventions that can be considered and you want to compare them: this

is the case of one of the two papers, where there is also a remediation intervention. This was just to remind the nature and the interest of the exercise.

What I want to say about this couple of papers, for this meeting today, is rather to illustrate how they illustrate strengths and issues in random intervention. So I think one of the strengths of the second paper in fact is to work over a menu of treatments. Here it's remediation versus incentives, it could be more than that in principle. It is very clear that those two interventions have a very different justification; you want to help if it is lack of skill, you want to provide incentives if individual efforts are inefficient in any sense and maybe the one is the truth, and maybe the second is more important. You don't really know exactly where to act in priority, and maybe there are different costs. It's nice if you consider a whole menu of treatments.

I think we have the same sort of issues in the labor market policies, where you have two big instruments that are incentives, like unemployment benefits, playing on the unemployment benefits, or job search assistance; and it's not under the same hypothesis or situation, etc, that one or the other is efficient and looking at the menu is certainly something useful. And maybe the point here is that with no experimental design it might be very difficult to compare several interventions: you will find a natural experiment for remediation here and another for incentives there and they will not be on the same ground and the same population, with the same output measures and so on. So if you want to have a chance to look at a complex menu of interventions you really need to be in the experimental context. I think this is one of the strengths that is really well illustrated here; the other is that you can have transparent and accurate results. Many of the results were visible right away from simple tables, and here what is striking is that in spite of small samples we have fairly accurate and precise results.

In the college experiment there were not so many people treated; in the Israel experiment the unit which is relevant is schools, and there aren't so many schools either, but this also points to the fact that if you want to have accurate results in this consensus, you have to have really reliable data. If you have measurement errors in the data, then it's over. You will see nothing with rather small samples. And another thing that is illustrated by this paper is that sometimes, what you did not insist on in this presentation but which is very important in the Israel paper, is that it's very simple and transparent but in another way, there are complicated issues with standard errors at some points. That doesn't make everything that transparent all the time. Obviously, in those papers, this has been treated very carefully.

Now a few comments on issues in random evaluation. One, which is not specific to random evaluation but to evaluation in general, is that the effects of an intervention can be heterogeneous over people. And I think that in this paper it is pretty well illustrated by the boys and girls difference. Obviously you have two populations over which the same treatment has no effect or fairly strong effects. Heterogeneity of this treatment is obvious in this paper, this is a very good illustration. The counterpart to this is that, what is the theory for this, how can we understand this? And this is where the results are strong, interesting but also very intriguing. If you go one step further, it also points to the fact that more generally, if there is heterogeneity in the treatment effect, like boys and girls, there is probably also unobserved heterogeneity in the treatment effect and if there is, and if it is important, and because it is

unobserved, well, in the end, we don't really know whom we want to treat preferably. Maybe we want to treat girls, but maybe we want to treat people with other sort of characteristics that we don't really know about. This raises the general question of generalization, which is a general question in evaluation, not specific to random evaluation. But maybe the point here is that random evaluation does not necessarily answer fully this question of generalization when there is treatment heterogeneity.

The second issue is maybe that the sort of results we have here are very reduced forms and maybe it would need additional input to understand more of the process that leads to this sort of results. For instance, in this very situation we could think of looking at interactions between the fact that you offer money to people who work, who study more for school, how this interacts with the fact that they might spend time working for a wage, or how did this interact with the fact that their parents might give them money for their living, and therefore all this sort of interactions. And one question that you might want to ask is, does this result from the fact that people, because you offered them money, they worked less for a wage, and then spent more time working on their studies and this sort of mechanism. And looking at this would require something which you don't seem to have in the data, which is measures of studying effort, measures of wage work, measures of transfer, and so on. One question of course is: would a finer description of the process help us understand something of the boys and girls differences and maybe the general context here is over whether I want to have more theory or a more robust process.

The third issue that is visible in this paper is that there are peer effect problems. There might be peer effects—it is not a problem in itself—especially in education, this is something you might expect. It may be something about work attitudes shaped by social interactions: if my friend works more, I will work more, if he despises working, I will despise working, and if this is the case then just comparing treated and untreated does identify some parameters, but maybe not necessarily the one we are thinking of right away. Here we have a situation wherein the treated might somehow be affected by the Treatment through their interaction with friends, who were randomized as treated. And of course, tackling this sort of problem requires a specific design which we do find sometimes in the literature, but which doesn't seem to be present here and which is not always present.

Maybe a general view of all these questions is to point to the fact that it seems that the literature on randomized evaluations is heading to even more than what is in this paper, more complex designs and multiple randomization levels that will allow us to learn more about the processes at hand and of course those more complex designs are very often more difficult to implement in practice for many reasons. Thank you.

Michel QUERE, Cereq

My question is related to the design of the first experiment that Mr. Angrist presented and is related to the issue of the complex experiment in some sense. I was wondering on the following: you target a population of students as a main target for incentive changes. Why not consider the other side of the coin, which is the population of teachers or even more to consider mixed changes in such of schemes, considering the rewards that increase the reasons

that could affect the population of teachers or the population of students? I understand that it could be a complex issue bearing on the design of mixed changes. But it seems to me very central to mix in some sense of the other side of the coin, which is a population of teacher incentives that could affect them in providing better results with regard to the *bagrut* issue. Thank you.

Joshua ANGRIST
I think it is better to respond to them one by one. Very briefly, I'm very interested in the effects of incentives on teachers as many labor economists and development economists are. This particular study wasn't about that, but there is a lot of interesting research on that, some by my colleague Victor Lavy, and some by the J-PAL group, Esther and Abhijit and their colleagues, including some studies that involve elements of random assignments. Usually we think of that as a distinct question, how teachers should be compensated, no less important but distinct from the question of incentives for students. So we tend to study them separately, but it would be interesting to do an intervention that would combine both, absolutely.

Jonathan PORTES
Jonathan Portes from the Department of Work and Pensions in the UK. One very interesting methodological issue here is the choice between randomization by individuals and randomization by unit of aggregation, whether it is by a school or a village as in some of the development experiments, or, typically in the sort of experiments that we do as a job center plus district. Now we tend to do far more of the latter than the former, because that is much easier. You don't have to toss any coins or press any computers; you just choose some districts and match them, so we've probably done a hundred pilot control experiments by area compared with three or four proper randomized ones. But what interests me is leaving aside the practical issues, what do you think the theoretical considerations in are driving what the choice is. We know that practical issues make it much easier to do the geographical one, typically. But theoretically, when you need a theory of change and you need to know if whether you're interested in person or place effects, what actually the right experimental design is. I'm not sure anyone has really teased out these issues very well.

Joshua ANGRIST
That's a great question Jonathan, something that Victor and I struggled with, in fact we did an early pilot that involved individual random assignment and we ran into a lot of issues that led us to believe that the right way to think about the incentives was at the school level. Partly because that's the way it would probably happen in real life at least in the Israeli context, but not necessarily. And partly because we thought the incentives would be more effective if they were delivered with the knowledge and cooperation of the teachers and the principals. Of course that would have been very difficult if not impossible to do where you have some students who are getting and some students who are not getting. Although, at the university level, we were able to do that because it's a little bit more fragmented, there is a sense in which getting the *bagrut* is a kind of communal activity. The practical number one issue besides conceptually whether you want to randomize at the group level, there is a conceptual side to it. The number one practical issue is power, statistical power. Group randomized trial

you need way more data as Marc mentioned and we grappled with that it can be very difficult to come up with enough clusters to have an adequately powered design.

Marc GURGAND
I think there are three points, one is, it's easier to implement, I think this morning, M. Banerjee pointed out that for one of his experiments there was no way it could be done, else than randomizing areas. This is a power issue, a problem, and then also, there is an advantage of this sort of randomization, it can be that if you are interested in equilibrium effect then you might be able if you have enough zones you might be able to compare, like labor markets, treated and untreated labor markets overall. This might give you information about equilibrium effects that might be there that you cannot have, if you just have a place and have people that are treated and untreated, which would make it more difficult; there is also a conceptual interest in doing that. So it depends very much on the sort of questions that you want to ask.

Thierry MAGNAC, Ecole d'économie de Toulouse
Could you elaborate on the compliance issues, because you said, you can use the randomization as an instrument variable but you changed the parameter into the effect of the compliers. So you changed the way you managed the experiment, so if you want to scale up the experiment you are going to have different measures. In particular, comparing what you get in the second experiment, services and money, yet the compliers are not the same. You need assumptions to compare this to the facts that you got, don't you?

Joshua ANGRIST
That is a great point, Thierry. Indeed, the design has this kind of link where you randomize intention to treat, people comply, and you get the net outcome that's induced by that. There is an alternative design where you get *ex ante* compliance, and this is the more traditional design, and you would find this more in labor training evaluations, where you get all your subjects to agree in advance that they will cooperate with the treatment protocol, whatever it is. But then at the last minute you randomly assign treatment in that population of compliers and then you have hopefully something close to full compliance because you've screened for compliance. But an important point which we understand from the instrumental variables framework, is that because these trials have one sided compliance, one sided non compliance, that is no one in the control group gets treated. Those two designs actually estimate the same things. You're right in general that instrumental variables captures the effect on the compliance of the sub-population, the people who get treated because they are offered treatment. But in the case where no one in the control group is treated, the local average treatment effect is the effect of treatment on the treated. The intention to treat design used in both of these studies is estimating the effect of treatment on the treated. And not fundamentally a different parameter than what a traditional design where there is *ex ante* screening would get. I don't think that's appreciated enough, that fact about experiments, that really, you have a lot of flexibility about the design even if your target parameter is the effect on the treated.

John MARTIN

Merci et merci à Joshua Angrist et à Marc Gurgand.

Retour sommaire

Economie du travail : les expérimentations nord-américaines

Béatrice SEDILLOT, DARES

Bonjour à tous, je vous propose de commencer la session de l'après-midi. Je suis Béatrice Sédillot, chef de service de la DARES. Je suis très heureuse de présider cette séance de l'après-midi, qui sera consacrée à l'analyse assez approfondie de diverses expérimentations mobilisant les méthodes par échantillonnage aléatoire dans le champ des politiques du marché du travail.

Je suis d'autant plus heureuse de présider cette session que, comme nous serons amenés à en discuter notamment demain, la France a commencé à mettre en œuvre ce type de méthode, mais de façon encore très récente et sur un champ assez restreint. Nous sommes tout à fait intéressés à bénéficier de l'expérience des autres pays, qui ont commencé à mobiliser de façon beaucoup plus large ce type de méthodologie, de connaître le bilan qu'ils en tirent tant en termes d'intérêt de ces méthodes que de difficultés et limites parfois opérationnelles qu'ils ont pu rencontrer. Je suis très impatiente de voir l'ensemble des présentations qui seront faites cet après-midi.

Une heure et demie est réservée pour deux interventions et discussions. Les intervenants auront vingt-cinq minutes pour présenter leur exposé. Nous aurons ensuite dix minutes pour les discussions et enfin dix minutes de questions pour la salle.

Avant la pause, une première session sera consacrée à deux expérimentations dans les pays nord-américains. On commence tout de suite par les Etats-Unis, avec Monsieur Jeffrey Kling, de la Brookings Institution, qui va nous parler de l'expérimentation *« Moving to Opportunity »*. La discussion sera introduite par Thierry Magnac, de l'Ecole d'économie de Toulouse.

« Moving to Opportunity »

Jeffrey KLING, The Brookings Institution

Thanks for inviting me. It's delightful to be here. I'm going to talk primarily about a particular project that I've worked on with a large team in the US. And just to give a little bit of context to this, there is a long tradition of having pretty large scale randomized experiments, particularly on US labor market policy and this study comes in that tradition, being a bit larger in scale and more closely affiliated with government policy than some of the things that Abhijit and Joshua Angrist talked about this morning.

Over the last forty years in the US there have been a long series of randomized social experiments, particularly in employment and training, and in welfare to work policy, and there is a nice catalog of all these in the Digest of Social Experiments. There are literally hundreds of trials that have been undertaken, and there's an evaluation industry in the US. There are three major firms and hundreds of millions of dollars of business that they do in evaluating programs for the states and the federal government that have been undertaken. In

the last six years or so, Abhijit was emphasizing the major developments in social experiments in the area of development economics. In the US, that has also been true in the area of education, where the Institute for Education Sciences has sponsored many new randomized studies.

The one I am going to be talking about today is in the housing policy area, in particular about the use of public housing vouchers. The idea is that in the US there are a large number of apartment buildings owned by the government where people can live for very low rent and you have very high concentrations of very low income people all living in these government-owned apartment buildings. So people are very interested in the question of what happens if you deconcentrate poverty and if you offer people housing vouchers that they can use to move out of these government-owned units into private market apartments where there is a lower concentration of poverty around them.

You might think that this would be beneficial for people in families, especially some of the children who would be living in safer neighborhoods; they would tend to have role models who had more education or who were more attached to the labor force. On the other hand you might be worried that they would experience some adverse effects. For instance, if you used to be in the middle of your class then you move to a new neighborhood where you're now at the bottom of the rung of academic achievement then how will that affect kids?

These are some of the questions that we were trying to address by looking at some data. What are the impacts of having families move to different kinds of neighborhoods? The way this experiment worked was that in 1994 to 1998 in five different cities in the US, families in public housing were eligible to participate and there were forty-six hundred families in this demonstration. One thing that could happen to you if you were one of the forty-six hundred families who were interested in participating was that you participated in a lottery and eighteen hundred people received what were called low poverty vouchers. This is something where you could move to a geographical area that had an area-wide average of poverty rate of less than ten percent, and then you could use this voucher to help pay your rent in that area and you also received some counseling to help you move into one of these areas. And so about half of the people that were offered these low poverty vouchers actually used them to move to a new place.

Then there is another group that received a traditional voucher which in the US is known as a Section Eight voucher, closer to two-thirds of the people who received one of these traditional vouchers that you could use to move wherever you wanted—it wasn't geographically restricted to people being in a low poverty area—and a larger fraction of them used this one. Then there was a control group of people who continued living in US public housing who didn't receive any new assistance through this housing voucher lottery. The families who are participating in this demonstration are primarily mothers with children who don't have male adults in the household. Less than a quarter of them were employed at the time when this started, although over time many more people became employed. The results that I'm going to talk about really came from a large amount of data that we collected about five years after people had been offered these vouchers, and so this graph shows you, in the control group,

some of the conditions in the neighborhoods, where you could see what the average poverty rate was, how many people reported being victimized by crime, or other statistics. And then the red bar shows you, for the low poverty voucher group, that these were much lower, so the average poverty rates are lower, the percentage victimized by crime falls from twenty-one to thirteen and so on, so that the local conditions were substantially improved.

For the adults, one of the things they were interested in testing was whether if you moved to a lower poverty area, people would find it easier to obtain employment, perhaps because the labor market was better there, or because they would have more connections to people who were working who would help them find jobs. That turned out not to be true. There is not an appreciable impact on labor market outcomes of the adults. They are a bit more healthy. Particularly in terms of mental health, the adults are doing better. Similarly for youth, people were interested in whether being in potentially different schools or having different peer groups was going to have an impact on educational achievement. That also does not seem to have borne out in that the test scores are not particularly different in math and reading between the control group and the low poverty voucher group.

There were much more substantial differences in some things. The teen girls were much less distressed, much less likely to use marijuana, had fewer behavior problems, and were less likely to have been arrested for crime. The opposite seems to have been the case for the teen boys. The red bars are the control group, and the yellow bars are showing that they are more likely to use marijuana, more likely to have behavior problems and more likely to be arrested for crime. So there was this particularly unexpected gender difference. Joshua Angrist talked a bit how in the educational incentives it seemed that the girls were the most responsive, here not only are the girls more responsive but there actually seems to have been some adverse effects on the boys.

To summarize, there's improved housing, increased safety, and lowered adult depression. There are even lower rates of adult obesity that were accompanied by slightly higher rates of exercise or they were more likely to eat a healthy diet. It seems to have been good on most dimensions for the teen girls, not so good for the teen boys and had little effect on employment or on kids' achievement.

In terms of thinking of the cost benefit analysis for something like this, from the government's point of view, they are paying about the same for having people in public housing units or paying for the vouchers that people used to move to different areas. If you think of the positive effects for teen girls, then the negative effects for teen boys are basically a wash. And since the higher mental health of the adults is a benefit, then you might say this is beneficial overall, although it's hard to make this characterization, because it's hard to say exactly how to weigh the fact that the boys seem to be somewhat worse off.

What we are doing now is to look at much longer term impacts. We are just starting a survey now, where we will be looking ten to twelve years after people have received their housing vouchers, and collecting some data. So we are interested in what the long term effects are, how they evolved over time, and especially, what the impacts on kids were, who were very young when they received a housing voucher. So if you were one or two years old and your

family moved, then we will look at what's like for you to have spent your whole life in one of these areas relative to other kids in families who were not offered a voucher and didn't have this opportunity to live in a different kind of place.

Some of the mechanisms we are trying to look at more are the social ties and whether there was an important difference for boys versus girls, and at their connections to adult role models. We have some preliminary evidence that it was more disruptive to teen boys to have moved away from, say, their uncles or their mother's boyfriends, or other adult male role models; they didn't seem to have these role models after they moved to low poverty areas. We're trying to dig into that some more to understand what was happening to them. We'll be using a variety of different types of data so this gets to some of the innovations of measurement Abhijit talked about this morning. We will do some administrative data, looking at some public assistance receipts, employment and earnings, some survey data where we can craft specific questions so that we can get at things like the adult role model questions I mentioned a second ago. Some things that are more specifically about physical health, that is, height, weight, and waist measurements, blood pressure, trying to look at some of the early precursors in blood samples for cardiovascular disease, having kids do some math and reading achievement tests for us, and looking at what is happening in the neighborhood. So we will look again at the same categories for which I showed you graphs, in terms of education, employment health and risky behavior measures and try to learn what we can in the long term about what the impact of moving to new areas has been. Thanks.

Retour sommaire

Thierry MAGNAC, Ecole d'économie de Toulouse

Thanks for this interesting paper. I'm just going to go over again all the main motivations for the paper, then I will talk about measuring things related to changing neighborhoods and these kind of questions. The main motivation for the paper is to measure the effect of housing policies in the US, so in particular, we have a policy of deconcentration of vouchers which aim to deconcentrate the poverty, get people out from high poverty neighborhoods to low poverty neighborhoods, these are in particular the kinds of questions that are asked in this research project, while the tools that were used are a controlled experiment in which there are two Treatment groups, but in fact the authors are mainly talking about the second treatment, which is an experimental treatment. This is typically the voucher that helps households to move from a high poverty neighborhood to a low poverty neighborhood. So what are the main messages of the paper? In particular, I am talking about the paper published in *Econometrica* four years ago. The first message is that the target rate of the voucher is reasonably high. This is only reasonable because only fifty percent of the second treatment group takes up this sort of vouchers and they don't move out.

That is the first thing. The second thing is that the outcome of these vouchers, of these public housing policies is mixed. There are plenty of outcomes which do not respond to this type of experiment and this is only physical, mental health for girls that seem to really respond to this. In fact the questions that are asked in this paper if we want to enlarge the context in which

these questions are asked, the questions about how to learn about neighborhood effects. The value of effects involved in this kind of measurement, we summarize them usually as selection issues, reflection problem, and this kind of thing so to make my idea clear, I'm going to make a quick example; there will be only one question.

We have an outcome Y for a household living in a certain neighborhood and this outcome is related to the usual controls which are the *xi*, income for example or anything you can think of, socio-demographic composition of the household, and there are variables related to the neighborhood. So in particular you are interested in the effect of the poverty rate in the neighborhood on the outcome. So this is described by the variable Z, and the parameter of interest that you want to measure is β. You have other characteristics of the neighborhood that entered this equation which are the ε and you have unobserved heterogeneity across households which is described by U, so this is a very simple type of model and criticism can be applied. Notice that you have a more generous treatment effect because you are interested in the fact that you move one household from high poverty rate neighborhood to a low poverty rate neighborhood.

The question you ask when you try to think about the estimation of this kind of parameter, well the main question is a question of causality, and the question of causality there is the variation in the Z across households, the Z of the poverty rate in the neighborhood, are these exogenously set and is there exogenous variation of Z and the response is no in general, because households self select into different neighborhoods for reasons we don't know completely but the U could be correlated with the Z in this case, so there is no exogenous variation in that case. There could also be some correlation between the Z and the ε, so if you move Z, you are going to move ε at the same time, so the question is, in survey data in neighborhood studies, it is very difficult to get exogenous variation of Z.

So the response is, we do a social experiment so what would be the ideal experiment that you would like to do in order to get the effect you want to measure. The ideal would be to change the characteristic Z in a characteristic Z' without affecting any other characteristics so it is called an intervention. Obviously, it is quite unfeasible in this context because you need to increase the poverty rate characteristic for one household moving for any household, in the sample from one value to another so it could be done by injecting a lot of money into this neighborhood, by making neighbors rich for example. But it is kind of implausible to do this kind of experiment, so in the controlled experiment that is studied here we do something that mimic the intervention, but it's not exactly the intervention, because what you do is you force people to move from one neighborhood to another neighborhood, and in this other neighborhood the poverty rate is going to be different. This is what is done in this controlled experiment, you randomize and the treatment group is given a voucher to help them to move out of the high poverty neighborhood, to the low poverty neighborhood and this is creating exogenous variation. In conclusion, randomization is an instrumental variable, if you remember econometrics classes, because the causality diagram is very simple.

Randomization affects Z, so R affects the characteristics of the neighborhood Z, which affects the outcome Y. And the only way it affects Y is through the variable of interest Z. But there

are limitations to this, because there are differences between the ideal experiment and the controlled experiment. The first thing is that—and this is a general comment on controlled experiment—is that randomization of this is usually answering a single policy question in the direction of the randomization. You just move from Z to Z', so you don't study anything else related to these housing policies, just this part of the question that is of interest to you, so controlled experiments have usually a quite focused question. The second point is that randomization has the same issue as instrumental variable estimation, when the treatment effect is heterogeneous. What you get is the treatment on the treated and this is not always the case, that you are interested in the treatment on the treated, you might be interested in other parameters, if you want to scale up the program, for example, and this kind of thing. So one issue in the controlled experiment is that the parameter of interest should be defined well in advance of the controlled experiment. The experiment should be tailored to this need and one question that I have is how was the actual process in this experiment in this MTO experiment in order to define these parameters in advance and to tailor the experiment to this need.

On the side of limitations, the first limitation that I see in this experiment is the imperfect control that you have because it could be the case that randomization does not affect only the variable of interest Z, it could affect other variables, and particularly it could affect the ε, so in terms of my model, it means that if I change neighborhoods I am not changing only the rate of poverty in the neighborhood, I'm changing the networks of the person, I'm changing the school where the kids are going, I'm changing everything. I do not control everything that I am changing. So what is the parameter? I agree that I am changing the parameter that is being measured, but what is this parameter?

The other limitation is the imperfect compliance. Only fifty percent of people complying with the voucher used the voucher. In the experiment, apparently there are no selection on observables, but there might be selection on unobservables, and so what you get is the treatment effect on the treated. But it is not clear whether this is exactly the parameter we are interested in in order to scale up or scale down this sort of problem. And there are other things related to controlled experiments, such as general equilibrium effects.

So, on the results, the question is whether on the results or the surprising absence of results that we get can be accommodated by the limitations I just talked about. There is the second question, can we go further, can we do a cost benefit analysis on this, on whether it is worth implementing the policy because that is really the question behind the usual evaluation type of study. The economic question could also be interesting, i.e., why are our households moving? Why do some households move to a low poverty rate, can we use the data produced by the experiment to understand why in this case households are moving. So it is going to explain something about compliance and non compliance and its going to be informative about the way you are going to be able to measure other parameters, than the treatment on the treated and the last question is what are the consequences that you draw for future experiments about the neighbourhood effects?

Béatrice SEDILLOT
I thank you very much. I suggest first you answer some of the questions from Thierry and then we go to questions from the audience.

Jeffrey KLING
I'll address a couple of things. One is about which parameter are we really interested in and how would you scale up this kind of program if you wanted to. There are two questions that the people in the US have been particularly interested in. One is what happens if you introduce a number of new vouchers spread across many cities, and so, if you, say, had five hundred more vouchers in Phoenix and five hundred more vouchers in Las Vegas and five hundred more vouchers in Seattle, what would that do? And then another question is what happens if we blow up the public housing projects and shut them down and make everyone move. This experiment is exactly well suited to the first question of what happens if you have incremental expansions of voucher programs on the order of, say, a billion dollars a year which is roughly what we have been doing in the US. This experiment is very well targeted to that question and less well targeted to the question if we just eradicate public housing. So I think people are interested in both questions and this is well targeted to one and less well targeted to the other.

The other thing I'll comment about is the idea of the bundle of the neighborhood characteristics; so if we think of this demonstration as being about neighborhoods then you could index them by the poverty rate. But I wouldn't want you to think that we were intending to hold everything else constant and just change the poverty rate. Let's say hold the fraction of high school graduates constant or the criminal victimization rate constant like that. The whole bundle of characteristics is changing in the way that you described and so we think that the poverty rate is a way of indexing that bundle to give you a sense of how dramatic the changes have been, but we are consciously intending all of those things change, because, in fact, that happens when people move into very different neighborhoods.

Thierry MAGNAC
But if you propose a voucher instead of indexing the voucher on the poverty rate of the new neighborhood, you could index the voucher on other characteristics in the neighborhood, and you would get a different answer in terms of estimated parameters. Because you targeted the way you constructed the experiments on the low poverty rate so every time you index on some other variables you are going to change the measurement that you do on the impact of this policy. Aren't you?

Jeffrey KLING
I think there are two parts to that, one is the characteristics of the low poverty voucher itself. So the voucher has a specific requirement in it, that you have to move to something that is called US Census Tract, which is a geographic area where four thousand or so people live, and the poverty rate as measured by our 1990 census had to be ten percent or less in order for it to qualify. That is a technical requirement of the voucher itself and you could have had other requirements, you could have based it on the criminal victimization rates in the

neighborhood or some other thing. I think the second part of that is how you think about what the bundle of characteristics is, in particular about whether you think it's a linear poverty rate effect. In some other analysis we've done we looked at different cities. I mentioned that there were five different cities and two different treatments, so you could look at what's the impact in the Boston site for the section eight group, which had relatively small changes in neighborhood characteristics, or the site for the experimental site in Los Angeles that had a very large change in neighborhood characteristics. If you graph that where you're looking at the change in the neighborhoods on the x-axis and the outcomes on the y-axis, those appear to be very linear in poverty rate space. To the extent that you change the measurement of that, that relationship wouldn't necessarily hold, if you use some different characteristics there, that is definitely true. Our main goal there is mainly to illustrate the results in a metric that has some intuitive appeal. Then you suddenly understand what it means to have a poverty rate that has been cut in half; it is a way of communicating the results.

Béatrice SEDILLOT
Thank you. Je cède maintenant la parole à la salle. Y a-t-il des questions ?

Un intervenant
I was a bit surprised to see neither of you relate this to the peer effects literature. It is mostly in education, but I guess my reading of that literature, I'm not expert, is that actually pure effects are pretty difficult to detect. They probably differ between boys and girls which is certainly indicated in literature and thoroughly here as well, but I guess my reaction is to say well, given that literature, it is not surprising that you're not getting terribly big impacts here. Either they don't exist or they are really very difficult to detect through standard statistical methods.

Jeffrey KLING
Monsieur, I think that whether you say they are big or small, or if there are no effects, really depends on what you're interested in. So if you are, say, a labor economist such as yourself, then you might say, oh, there is nothing here, and if you were interested in peer effects on kids' education, you might say there is nothing here, if you're a public health person who's interested in, say, what's the comparison between providing pharmacological interventions where you're giving people medicine, or where you're giving people high doses of counseling, or whether you're giving them a housing voucher to move to a new neighborhood, the magnitudes of those things are approximately the same, which is, in that sense, a big health intervention. I think the magnitude depends on which outcomes you are talking about and what you are interested in.

Béatrice SEDILLOT
Ce qui est très frappant, dans cette expérimentation, c'est le fait que vous vous inscrivez dans un temps long d'évaluation. Vous mesurez l'impact pendant des années, jusqu'à dix à douze ans après l'affectation aléatoire. Certes, cela est important pour arriver à voir des effets de long terme. Mais lorsque l'on se situe douze ans après l'affectation aléatoire, comparer les populations signifie que, pour tous les événements qui vont pouvoir les différencier au fil de

ces douze ans, l'on fait implicitement l'hypothèse qu'ils ne sont liés qu'aux effets de voisinage. Est-ce que cette hypothèse n'est pas forte ? Même si les populations étaient initialement identiques douze ans auparavant, ne peut-on imaginer que les divers événements qui ont pu survenir depuis font que l'on doit les différencier d'une façon ou d'une autre, ce qui pourrait peser sur les résultats que l'on observe après d'aussi longs délais ?

Jeffrey KLING
The ten to twelve year study that I mentioned has not yet been done, but we are collecting that data now. The analysis plan is to address the issues that you are thinking about—looking at the difference between the low poverty voucher group and the control group over time. There we are relying very directly on the power of the random assignment. The average outcome would have been the same except for the fact that one whole group of families received their voucher in the lottery and others didn't—so the only thing that distinguishes them is that they got that different lottery draw. So we are relying directly on the random assignment and the differences between those groups are all going to be traced back to that.

Béatrice SEDILLOT
Nous allons passer à l'intervention suivante avec l'exemple d'une expérimentation menée au Canada sur l'impact des mesures d'incitation financière à la reprise d'emploi.

Philip Robins, de l'université de Miami, va nous présenter les résultats de ces trois expérimentations. La discussion sera ensuite introduite par Denis Fougère, du CNRS.

Retour sommaire

« Self Sufficiency Project »

Philip ROBINS, Université de Miami

Thank you very much. I am Philip Robins from the University of Miami in the USA. I 'd like to thank DARES for inviting me to come and to speak with you about the Self Sufficiency Project (SSP), a randomized experiment for welfare recipients conducted during the 1990s. I was involved with the design and evaluation of SSP and am the author or coauthor of several SSP reports and articles. SSP was actually three labor market experiments that took place in two provinces in Canada.

First, I'd like to point out that this particular labor market experiment should be characterized as an experiment in financial incentives, testing a financial incentive program. As this slide indicates, there has been a renewed interest in recent years in using financial incentives to encourage work effort among low income families. I use the word "renewed interest" because actually the idea of using financial incentives to encourage self-sufficiency has been around for quite a long time. In an expanded set of notes that will be published on the DARES website, I present a little bit about the history of financial incentive programs, which had its beginning, at least in the US, with a series of negative income tax experiments during the 1960s and 1970s.

As you can see in this slide, financial incentives are being used in many countries. I mention here three main programs including the earned income tax credits in the United States, which is the largest financial incentive program. The UK also has a very similar program. While doing research for this particular presentation, I found that France also has financial incentives built into its system of support for low income families. These financial incentives are through an earned income tax credit-like program and also through the type of incentives that exist in the US (that is, earnings disregards for encouraging work among welfare recipients). Financial incentive programs can be categorized as "the carrot approach" to encouraging work. "Making work pay" is a term that has been thrown around over the last decade or so which can be contrasted to "the stick approach". The "stick approach" is also being used in the United States. The stick approach basically requires that certain work obligations, be fulfilled in order to receive benefits from various programs.

Perhaps the most dramatic test of financial incentives for low income families since the NIT experiments, which I mentioned in the expanded version of the slides, is the Self Sufficiency Project or SSP. SSP was an experimental program conducted in the 1990s in the Canadian provinces of British Columbia and New Brunswick. SSP was evaluated jointly by two major research evaluation firms in North America. One was SRDC, which is located in Ottawa, Canada and the other one was MDRC, which is located in New York. Most people are not aware that there were actually three SSP experiments, each one having a slightly different objective. However, the overall design of the experiment was similar among each of the three.

What these experiments did was to basically provide a very generous income supplement for up to three years to welfare recipients who worked thirty or more hours per week. Thirty hours was considered bordering on full time employment. I believe that thirty- five hours per week in France is considered full time work. SSP was a voluntary alternative to welfare. That is, people were not forced to participate in the program. They could stay on welfare if they so chose. They had one year to find full time employment and the supplement formula was actually quite simple. It's given there—half the distance between a target earnings level and what the actual earnings level is. The initial target earnings level was thirty-seven thousand dollars in British Columbia and thirty thousand dollars in New Brunswick.

To give you some idea about the generosity of the incentive, it effectively doubled the hourly wage for most recipients. For example, if a recipient found a job for ten dollars an hour, the supplement essentially gave them a twenty dollar per an hour economic return to working. One of the experiments I will discuss also tested the effects of combining services in addition to the financial incentives. This is something that Josh talked about this morning and like the results that he reported, we also found that the effects were even stronger when you combine services with a financial incentive.

This table basically summarizes the three experiments and I will go over it very briefly with you. The main experiment was called the SSP Recipient Study and it was targeted on long term welfare recipients. In terms of the random assignment that was used in this experiment, potential participants had to be on welfare for a least a year in order to be eligible to be randomly assigned.

The second study, the SSP Applicant Study, looked at a somewhat different question. It looked at people who had just recently applied to welfare, within the last six months. Finally, the SSP Plus Study was similar to the SSP Recipient Study in that it focused on long term welfare recipients. Also, as you can see in the treatment row (the second row there), it provided employment services in addition to the work conditioned earnings subsidy. The sites, as I mentioned earlier in the Recipient Study were British Columbia and New Brunswick. The Applicant Study was conducted in British Columbia and the SSP Plus Study was conducted in New Brunswick. The main outcome of interest in the Recipient Study were the impacts on full time employment and income. The main outcome of interest in the Applicant Study was the size of what are called entry effects. Entry effects essentially are effects that are akin to a moral hazard problem. Any time you have a targeted subsidy like this you have to worry about whether people who are applying for welfare are going to change their behavior in a way to make them eligible for such a program. So one of the things that was of interest was, would people increase their stay on welfare in order to qualify for this very generous supplement for the three year period? In the SSP Plus Study, we were interested in the additional effect of the services. The next to bottom row shows the sample sizes. You can see for the first two studies, they were quite sizable and the statistical power was very strong in these experiments. I don't report statistical significance, but in most of the cases the effects I'll be discussing were highly significant.

In the SSP Plus Study, which was somewhat of an afterthought, the sample size was much smaller. Each of the groups (it was a three way design), had only about three hundred people in them. Small sample sizes can be a problem in social experiments because sometimes it is very difficult to draw inferences. The take-up rate for people that actually received these supplemental payments varied between twenty-seven percent in the Applicant Study to fifty-two percent in the SSP Plus Study. Half or more of the people that were randomly assigned never received the treatment. The impacts that I will be showing are averaged over everybody including the people who didn't choose to receive the treatment. So depending on the assumptions you might want to make if you divide by the proportion that are taking up the program, the impacts that I am going to be showing you in a few minutes are quite large.

The SSP Recipient Study was the main study. Some people consider it to be one of the most successful social experiments ever undertaken. There has been almost a dozen scholarly articles written on these experiments in professional journals. The Recipient Experiment basically doubled the full time employment rate of the treatment group during its peak years. I will show you what that looks like in just a minute. The data that we used to evaluate the experiment came from a combination of household surveys that were conducted at various times during the experiment, along with administrative data from official welfare and program records. Sample attrition—which is a problem in social experiments—and drawing inferences from social experiments—was I think quite modest, as eighty-six percent of the baseline sample completed all the surveys. As I mentioned, in the Recipient Study, about thirty-six percent received a supplement and it was reasonably well targeted, because of those thirty-six percent—sixty percent of those were actually given to persons that actually responded to the financial incentive provided by the supplement payment. The remainder of

those people received the supplement payments, but would have found full-time work anyway. We call those "windfall recipients" —those are the kind of people we would like to minimize because they were given money for something they would have done anyway. On the other hand, you can think of it as an anti-poverty payment that certainly helped those people out as well.

Now, even though the SSP Recipient Study had a large effect on full- time work and poverty during its peak years, the effects gradually disappeared. At the end of the three year period, the impacts basically went away. It's not clear to the evaluators exactly why this occurred, but it could be that the jobs that the people took were unstable jobs that they couldn't hold, or they were jobs that gave them no wage growth. At the end of the period of supplement receipt, when they were not receiving payments anymore, they had similar job prospects as a control group and acted in a similar manner.

This chart right here is an interesting chart because it shows you the full-time employment rates of the two groups (treatment and controls). It also shows the difference between the two, which is the impact. As you can see, during months eleven to fourteen—if you recall, the recipients had a year to find a full time job—the peak effects occurred. After that period of time you can see the employment rate was very low initially among the control group and the treatment group. This reflects the fact that very few people who are long term recipients of welfare work full time. You can see the full-time employment rate accelerated quite dramatically among the treatment group. The control group had a steady increase in full-time employment, reflecting the normal aging of the group, and their children as they eventually left welfare. You can see what happens here is that the financial incentives caused many in the treatment group to get jobs in order to establish eligibility, but then the full-time employment rate was essentially flat from then on and the control group basically caught up to them. At the end of the period, there were no differences in full-time employment between the treatment group and the control group. That doesn't mean the program didn't have an impact, because obviously by speeding up the employment of these people, it did add to their overall lifetime income. I will demonstrate this when I show some results showing a positive effect from a cost-benefit standpoint.

This next graph shows the impact on part-time employment. One would expect many people in part-time employment would be induced to switch to full-time employment. In fact, that occurred as well. The upper line is the part-time employment rate of the control group and the part-time employment rate of the treatment group is below. The part-time employment rate was lower for the treatment group, implying that people did switch from part-time to full-time employment in order to take advantage of the SSP financial incentive. The next graph is the rate of receipt of income assistance by the two groups. You can see that initially it was a hundred percent because everyone was on welfare at the time they were randomly assigned. Over time the receipt rate of income assistance declined faster for the treatment group, but eventually caught up in the end. At the end there was basically no difference in between the treatment and control groups in the receipt rate of income assistance.

One of the most important things to do when one conducts a social experiment is to ask the questions: Did it work? Was it worth it? Should we undertake this program on a larger scale? Should we consider implementing it nationwide or in a particular region? One of the things that can help us make such decisions is a benefit-cost analysis. Benefit-cost analyses are not always done in social experiments, but they should become part of the evaluation routine. In this particular case, you can see that from the benefit-cost perspective, the last column basically shows us the net benefit to society per program group member, was about 2,500 dollars. So there was a positive net benefit to this program, though it did cost the government additional money as represented by the second column. This extra government expenditure occurred because the additional transfer payments paid to the recipients exceeded the welfare payments that they gave up. Those were offset somewhat by the taxes that were paid from the earnings of the additional job holders, but not by enough to make it a wash for the government. The government did end up with higher net payments, as I will show in a little bit. In the Applicant Experiment, net government expenditures were close to zero, so the Applicant Experiment was a wash for the government and had a much larger net positive effect for society.

When SSP was being designed it was recognized that welfare recipients sometimes have very formidable barriers to employment. Therefore, maybe helping them out by offering some services would boost the impact somewhat. Some early results from the SSP Recipient Study suggested this might occur so it was decided to try a smaller scale experiment that combined the supplement payment with these services. This was SSP Plus. Unfortunately, I think the sample size was too small and some of the inferences we were able to draw from this experiment are a little imprecise. I think in this sense, statistically, they are not as reliable as they could have been if the sample size had been larger. So the goal of SSP Plus was to see whether the services could enhance the effects of the financial incentive. You can't find out from this experiment whether the services alone would have an effect. There are numerous experiments now being conducted in several countries that are looking at the impact of services only. In SSP Plus, the objective was to estimate the *incremental* impact of the services.

The three groups that were enrolled in SSP Plus included a regular group which was the same group that was offered a financial supplement in the recipient experiment, then a plus group, which was offered the services in addition to the financial supplement, and a control group. This three-way design enabled the estimation of the incremental or additional impact of the services.

The next table shows the types of services that were made available to people in the SSP Plus group. The expanded sort of slide shows how much of these services were received by the various groups. For brevity purposes I'm going to pass through this right now, but I will say that it turned out that the main effect was on job search assistance while some of these other services didn't have much of an impact. Again, the services were voluntary, so that SSP Plus members were not required to use these services. They were just made available to them, which was more than what the control group could get.

Since I'm getting to the end of my time, I will need to skip a couple of slides, and show you some of the pictures. This next slide is the full-time employment rate for the three groups in SSP Plus. The incremental impact is a bit different from what we saw in the recipient experiment. If you look at this diagram carefully, you'll see that initially there was basically no incremental impact, but there was an impact of the financial incentive alone. That is the difference between the two upper lines and the middle line, but there was no incremental impact. Toward the end of the experiment, an incremental impact began to emerge and it looks like it might have been permanent although it was hard to tell, because the experiment ended.

A lot of theories have been offered as to why this incremental impact emerged at the end of the experiment, but one of the most convincing theories that seemed to be bourne out by the data is that these additional people were harder to employ in the first place and the services might have been able to help them find jobs that they were able to maintain. So it looks like—and again, because the sample sizes were so small, it is difficult to draw firm conclusions—but it looks like combining the services with the financial incentive helped people find more stable jobs.

The final SSP experiment was the Applicant Study, which as I mentioned earlier, was intended to measure an entry effect. There are actually two kinds of entry effects that could occur and this experiment could only estimate one of them. The first kind of entry effect is whether more people are going to apply for welfare in order to be able to eventually qualify for the supplement. The other is whether once they get on welfare, are they going to stay on longer? As you may remember, they had to stay on welfare for one year in order to become eligible for the SSP supplement, so we wanted to answer the question of whether they would extend their stay on welfare? The latter was the entry effect that was measured in this particular study, because only welfare applicants were randomly assigned and one couldn't estimate the entry effect associated with people applying for welfare in the first place. It turns out that there was an entry effect,. People did extend their stay on welfare, but it was a relatively small effect, about three percentage points.

Overall, the impacts in the Applicant Experiment were much larger than they were in the Recipient Experiment, perhaps reflecting the fact that recent applicants to welfare tend to be more job ready than long term welfare recipients. Therefore, the financial incentives was able to push them over the top, if you will. As you can see here, the difference between the full time employment rates seem to persist even at the end of the experiments. So like the SSP Plus Study, there are some permanent long-term effects on full-time employment in the Applicant Study. Similarly, these are also long-term effects on welfare receipt. In the next slide showing the benefits and costs of the Applicant Experiment, you can see it was pretty much a wash for the government and a very large net benefit to society. Thus, the Applicant Experiment was very successful from the standpoint of the benefit-cost analysis. The program basically paid for itself as the government incurred no net additional cost for this program. It yielded large net benefits to society. The reason why it was so much bigger than the Recipient Study, it is thought, is because the participants in this experiment might have been more job

ready to begin with. I apologize for rushing through these latter slides as my time has expired. Thank you very much.

Retour sommaire

Denis FOUGERE, CNRS et CREST-INSEE

It is a pleasure to discuss this communication, but it is also quite challenging because I come after several referees, and probably several very good and competent referees, who have judged and accepted the different papers for publication in high quality reviews. So having something original to say about the papers written by Philip Robbins on this experiment is quite difficult.

First, to go back to the main features of this experiment and then to compare this experiment with the French *Prime pour l'emploi* that has been mentioned by Philip Robins during his talk, it seems that there is something to learn for us in France, first about the way such a program has been targeted and also from the SSP experiment itself. Then I will add some very general questions and remarks, probably "non-original" questions.

Obviously the SSP experiment is the most important and convincing experiment among all in-work benefits programs. Two or three things about that. First the follow-ups which are long in the SSP experiments. It is important to note this point, especially for those working in the French ministries and who want to implement comparable experiments in France. In the SSP Recipient and Applicant Studies, the follow-up covered seven years. It's longer than the interval between two successive presidential elections! Moreover, several outcomes are observed: full-time employment first, but also the number of hours worked, individual wages, and an outcome which has not been commented by Philip during his talk, namely the school achievement of young children, which is very important in my opinion.

Three SSP experiments were conducted, not exactly in parallel, but more or less during the same decade, in order to examine different types of impacts. We should note in particular the experiment designed to observe the entry effects of the program, but also that other experiment implemented to estimate the impact of employment services. There is a big debate in France on this last type of intervention. Philip said that unfortunately these employment services were not so effective, compared to other aspects of the program.

The first thing to retain here is that the target of the SSP supplement is limited to single parents who have been on welfare for at least one year, which means that the target is very narrow. It reduces the incentive for people to enter the welfare system, because it is limited to this subgroup of welfare recipients, those who have been on welfare for more than one year. That's probably one thing to have in mind when designing public policy for low-skilled or difficultly employable people.

In the SSP experiments, the benefits are only offered to people who work more than thirty hours a week. This threshold is intended to limit the drawbacks that have been observed in previous NIT programs, for instance the EITC program. In some sense, we could have the same problem, the same drawback, with the *Prime pour l'emploi* in France. Moreover, the fact that the SSP supplement varied with individual needs, and not with the total household

income, implies that the wage supplement is unaffected by family composition, the earned income and the supplemental payments. Thus, when comparing this supplement policy with the French *Prime pour l'emploi,* the first thing to remark is that in France the eligibility conditions are rather loose.

I don't want to go into further details about that, but every French economist knows that eligibility conditions in the French *Prime pour l'emploi* are essentially defined by a minimum and a maximum threshold on earnings, these two thresholds implying a very large eligibility range. One consequence is that, according to some calculation that has been done by the French fiscal administration in 2001, the year in which the program *Prime pour l'emploi* was launched, more than eight million households in France, which corresponds to one-third of French households—on the whole, there are twenty-five million households in France — benefit from some tax credit through this program. Thus the target is very large! To my best knowledge, there is only one non-experimental evaluation of this policy which has been conducted by Elena Stancanelli (her paper is going to be published in the *Journal of Public Economics*). In her study, the target group is the group of women, and as you can see on this slide, the main result is that we observe once again the same adverse effect, the same disincentive effect, that was observed in the previous NIT programs: many women decrease their employment rate, especially when they are living in married couples. The impact is positive but weakly significant for other women, those who are living in couples but who are not married and lone mothers. So if we have to think about a new design for the French *Prime pour l'emploi*, according to scientific results, we should look at the results of the SSP program.

Another striking result is that entry effects produced by the SSP program were small. There is still a large debate about entry effects, in the academic but also in the political sphere. On this issue the paper by Card and Robbins, which has been published in the *Journal of Econometrics*, is quite convincing. They find that both entry and delayed effects are small. It is an important finding. Second, the SSP program had large positive effects on full-time employment and on income during the first three years, but we don't observe long-term effects after the first three years.

Philip mentioned that there is a reason for that, because, while the accepted jobs are probably low-wage and low-skilled jobs, there is no possibility to accumulate human capital or labour experience in such jobs. Did the people who were in charge of that experiment have any idea about the possibility to complement this policy with some training during the experiment or during the period in which people get the supplement?

And finally, the program has a sizeable positive net benefit for society. This result relies on a cost-benefit analysis which is worthwhile but very rare in this type of studies. That is a very good point. But this analysis is based upon wages, accepted wages, and that is one of my comments. For instance, in the twenty-fifth month, if I read correctly the paper published in the *Journal of Public Economics*, only fifty-five percent of persons in the treatment group were employed. So the cost benefit analysis ignores forty-five percent of the treated persons,

that is a large fraction, those who were non-employed. In principle, this selectivity problem should be taken into account in the cost-benefit analysis.

I know that this problem is difficult to handle. In particular, valid instruments should be used to control for the endogeneity of the employment decision, which may be also affected by the program, but they are difficult to find. A suggestion could be to implement a recent paper by David Lee, who proposes an estimation method for bounding treatment effects on wages, or more generally on any continuous outcome submitted to this type of selectivity bias, in the absence of valid instruments.

I have still two questions about difficulties which are very frequent in experiments. The first is about attrition. Attrition seems to be modest here, but anyway, if attrition corresponds to endogenous dropouts, at the end we are left with comparisons of outcomes for those who don't drop out, and the difference between the outcomes in the two remaining groups (treated and controls) is not, we know that, the average effect of the treatment on the treated. To identify this effect, we have to impose some further conditions. These conditions are recalled on this slide. I don't want to detail these technical points, but you will remark that, under these conditions, the probability of dropping out has to be the same in the control and treated groups.

My last comment is about non-compliance. How do you take into account the non-compliance probability in your estimates? We know that when there is no compliance, the difference between the outcomes does not measure the effect of the treatment on the treated. This difference measures only the effect of the full treatment on the fully treated. So, for identifying the effect of the treatment on the treated, we have to impose some further conditions, like the ones I recalled just before for attrition. I will stop here with my comments. Thank you for your attention.

Béatrice SEDILLOT
All right, thank you, Denis. I think that, because time goes by, there is just time for one or two questions.

Philip ROBINS
Thank you very much for your comments. Let me respond to the last one. I want to look over your comments in more detail because you have much more technical questions, in terms of impact of treatment on the treated. The attrition that I was speaking about was not the non compliers, it was basically attrition for the usual reasons that people drop out of the experiment. In order to derive the effects of the treatment on the treated, you could do that including or excluding the people who "*are treated"* one could do the simplest thing, which is just to divide. All the impacts I reported were calculated for the people who were *offered* the full treatment. It was the effect of being offered the treatment, not the effects of the treatment on the treated. But if you wanted to divide by the take-up rates and assume that the non-takers do not respond, you could get the impacts on the treated, which would have been much larger. We didn't do that because it required more stringent assumptions.

Jonathan PORTES
Thank you. Philip, the SSP was a very large demonstration program, and my question is a rather straightforward one. If it produced such sizeable net positive gains for society, why didn't any of the Canadians in fact expand it or continue it? Why is it that there appears to have been no take-up either among policy makers or politicians in terms of this program that you say was such a great success? There are some other studies that would suggest that it would not have been such a great success if it had been blown up a full scale national program. I'll be interested to hear why you think that this piece of evidence did not convince people.

Philip ROBINS
Excellent question, first of all I'm not Canadian, so I cannot answer that question definitively.

Jonathan PORTES
Hazard a guess.

Philip ROBINS
It was not without trying. I know that when the results first came out, the Canadian people who sponsored the experiment did try to get it on the political agenda, but it wasn't successful and I'm not really sure why, to tell you the truth, but it wasn't. I think I know some of the papers which you are talking about, which try to generalize the findings to the entire country and maybe are not so favorable as perhaps these. But I think there are some methodological problems with some of those papers. Frankly, I don't know. Hopefully, even though they didn't implement an SSP type program, perhaps it will have some impact on the types of financial incentives they'll use through the tax system. I certainly think in the United States the Earned Income Tax Credit, which has been around since 1975, was greatly expanded starting in the early 1990's. I think part of that was a response to some encouraging experimental findings on financial incentive programs.

Béatrice SEDILLOT
You mentioned the very small sample size in the experiment. What would have been, in your mind, the adequate sample size given that take-up rates are not so high, and have you any type of explanation for why this sample size was so small in the beginning?

Philip ROBINS
Several reasons. First the SSP Plus Program was an afterthought. There was a certain amount of resources that the Canadian government committed to this evaluation and I think they simply did not have the money to enroll a larger sample size. I think that was probably the major concern. Also, I have been asked many questions about SSP. Some of the most frequently asked questions are: Why was the eligibility period only a year? Why wasn't it six months, or a year and a half? Why weren't the treatment levels varied? All of these are very legitimate questions that social experiments can answer. But the fact is, the designers wanted definitive answers to a specific question. They felt if this was successful, maybe other social experiments could be conducted that would get at some of these variations. I applaud them for

that, because I think that SSP did yield very definitive, very defensible results apart, from some of the technical problems. Certainly in my experience, over a number of years observing and evaluating social experiments, this was very convincing.

Béatrice SEDILLOT
Merci beaucoup. Je vous suggère de faire une pause pour une dizaine de minutes.

Retour sommaire

Economie du travail : les expérimentations européennes

Béatrice SEDILLOT

Nous allons attaquer la dernière session de la journée. Nous traversons l'Atlantique pour revenir en Europe, avec deux expérimentations européennes. La première va être présentée par Jonathan Portes, du Deparment of Work and Pensions au Royaume-Uni, qui porte sur l'expérimentation *Employment, Retention and Advancement*. Je lui donne la parole pour vingt-cinq minutes. Ensuite, la discussion sera introduite par Etienne Wasmer, de l'OFCE.

« Employment, Retention and Advancement, demonstration for Great Britain »

Jonathan PORTES, Department of Work and Pensions, Royaume-Uni

Thank you very much, and thank you very much to the organizers of this conference, to DARES, and to Esther Duflo for inviting me to make this presentation. The first thing I want to say is that—as I often do at these events—I feel a little bit of a fraud presenting the results of this research, of which I'm at least as much—rather more—of a consumer than a producer. The meat of this presentation (on the results of the ERA project) are very much the product of the teamwork of a huge number of researchers and institutions inside and outside the Department—including, in particular, the Manpower Demonstration and Research organization in New York, who coordinated the evaluation, on the one hand; and, on the other hand, the small ERA team within my department, represented here by Aisha Riaz, sitting there. I point her out, so if anyone really wants to ask some detailed questions, they're much better off asking her than me.

But the fact that I am a consumer, and not a producer, has led me to vary the presentation slightly—to toss in, right at the beginning, something that has nothing to do with ERA, but much, much more to do with what we were discussing in the presentation this morning—the question that John asked, which is "How do you turn these experiments into policy? What do you actually do with them, and how does it work?" In my view—as Abhijit said in his first presentation—the point of this, and the point in spending all this money, and the reason for spending £4 million, as we are, I think, on ERA, or £100 million, as we are, on our "Pathways to Work" pilot), is to do this: it's to give you the "killer" chart. It's to have something that you can stick in front of your ministers or senior officials (or put up to a screen to a non-technical, outside audience) and say, "Look—this is what was happening before we did the experiment. We did the experiment, and this is what happened to the people we did the experiment on. And look—it made a difference. It really did."

You can see here. This is the impact of our "Pathways to Work" pilots for people who are on disability benefits, and it's the impact on their likelihood of being on benefits six months after the intervention. The top, colored lines are what happened to the treated group. I'm not going to go into detail, because this is not what I'm supposed to be talking about. I'm supposed to be talking about ERA, but I put up this chart because, quite simply, this chart secured us about

£300 million a year of funding from the Treasury, and led to national roll-out of the "Pathways to Work" pilot, which means in the future—and indeed, starting on April 1st of this year (in other words, six weeks ago)—we are now extending this program from 10 percent of the country to every single person coming onto the benefit.

Do this right—produce the "killer" chart—and you can get policy if you can get the politics right. I think that's another thing which is particularly relevant: we don't have a research department like DARES in the U.K. anymore within my department (it was abolished shortly after I joined the Department), and that has had a hugely beneficial impact on our ability to influence policy, because one of my job titles is Chief Economist (as you saw in the first slide)—but actually, that's not really a job, it's just an honorific title. My main job is Director for Children and Poverty. In other words, it's a policy job, and people like Aisha, and other economists and social researchers in the Department work primarily in policy divisions, working directly on influencing policy-makers on making policy, and I would highly commend that model of organization to any government department or organization that actually wants to integrate economics and research into policy-making. You need to get in there. You cannot sit in a research department and hope to have the sort of impact on senior officials and politicians and real policy-makers that you really want. That's my preamble, and now to the substance of the talk.

I won't get through all the slides, but they will be on the Web in due course, and anyone who wants to can take a web-link to the 250-page report, so rest assured there's an awful lot more behind this. What I am going to talk about on ERA? Again, I think it's important to look at the policy environment before looking at what we actually did, because that's what really matters. I'm going to start with that and talk through the usual sort of things.

What's "policy environment?" Well, we have a very strong labour market, so we're doing pretty well at getting people into work, and we're doing pretty well at getting people from disadvantaged groups into work. These are the groups that, historically, have significant employment disadvantages in the U.K.; and you can see those gaps are reducing quite considerably, apart from the yellow line, which is people without qualifications. But for lone parents in particular (the main blue line), the employment gap fell from 27 or 28 percent to 17 or 18 percent, so there has been quite a big improvement over the last ten years.

But we do much less well at helping people once they get in work. This is the classic sort of Anglo-American problem of the "low-pay/no-pay cycle," saying, "yeah, you can kick people off Welfare, but they get low-paid jobs and they often drift back into Welfare quite quickly. Even if they stay in the jobs, they stay poor. But we have much less effect. The reason why that happens is that we actually have reasonably good and effective interventions to get people into jobs, we sort of know how to do that, and a lot is based on personal contact; but once they get in jobs, we don't have any levers and we don't have any intervention, so we're looking at what we can actually do to help people, once they've made the transition from Welfare into low-wage jobs. What's going to help them stay in work and move up?" It's what we see as the next step in the Welfare-to-work agenda. We thought this was a big issue because we knew we didn't have anything that worked. We knew we were going to have to

spend quite a bit of money and effort to get something that worked in order to justify that, and then to justify mainstreaming. We need to do it through a pretty robust research methodology and evaluation, which is why we decided to use randomized control trials to look at this.

Target groups—and this is important, because it relates very closely to some of the previous presentations about gender differences: New Deal for Lone Parents (these are lone parents out of work, lone parents working part-time and receiving a Working Tax Credit—in other words, people who were on welfare to start with, but are now in low-wage jobs) and New Deal 25 Plus (these are people who have been unemployed for over 18 months, and in the U.K., this group is predominantly male). You've got three groups the first group being unemployed lone parents, predominantly female and quite disadvantaged, the second group being predominantly female, but rather less disadvantaged, and the third group being quite disadvantaged and predominantly male. We did this in six Jobcentre Plus districts, in sixty offices—that's about 7 or 8 percent of the whole U.K. What was in ERA? I won't go through the details (partly because I'd get it wrong, partly because it's all written down elsewhere), but it is essentially a mixture of advisory services: access to an advisor when in work, financial incentives, and an interesting innovation. Although we don't have any hard evidence, a lot of people to whom I've spoken who were involved with the program say that this was very important - it was actually a combination of the two. The Emergency Discretion Fund was for this discretionary financial incentive (in other words, if you spoke to your advisor and said, "Look, I really need emergency child-care, otherwise I'm going to have to give up my job and drop out," the advisor could give you some money). There's an interesting combination of the advisory services and the financial incentive, and it was extremely popular, both with advisors and clients, although you can't really disaggregate its impact.

That's your standard random assignment chart—how it actually worked and how we selected people. What did we do? Well, as I said, this is pretty big. We randomly assigned more than 16,000 customers, so that's quite a lot of people; you can see that it was a very successful Random Assignment, in the standard sense that they seem to be basically identical, the 8,000 in the Program group and the 8,000 in the Control group were functionally identical.

What did we learn? I think one of the lessons here that's really important—one of the reasons we spent quite so much money on this— was that your process study ("How was it implemented? What actually happened?") is just as important as are the raw impacts—you do actually need to get "under the skin" in these very complicated programs. If you're just offering somebody one financial incentive, then maybe you don't need to worry too much about the process—just as if you were offering them an aspirin and a placebo in a medical test, with maybe no need to worry too much about the precise process—but when you've got a combination of six quite complicated interventions delivered in different districts, each with rather different labour-market environments against a background of a whole host of other things. I think this is another interesting point that comes through when you read the evaluation: in every Job Centre Plus district, there are a whole bunch of other things going on at the same time. Some of these districts were also implementing the "Pathways to Work" pilots that I described earlier for disabled people. It's a different client group, but the same management. All of these districts had what we called "job entry targets." They had hard

targets for which their management was held to account regarding how many of the mainstream clients were getting into work, so, for a lot of the Job Centre Plus managers, this was sort of an "add-on." You've got these people from Head Office coming along, saying, "We've got this really important randomized evaluation," and they would say, "That's fine," and "That's great," and "It's all very interesting, but actually what my manager wants to know is: how am I going to hit my job-outcome target for this month?" Getting "under the skin" of how those things work through is really important for understanding what your results are telling you.

Lots of quantitative data: interestingly, this is something we tend to do in most of our evaluations for big pilots both customer surveys and administrative records. Helpfully, the results of these are quite consistent for this project and, indeed, for "Pathways to Work" as well. But I think that if you want to be really sure of your results, you do tend to need to do both of those. We went through a phase internally, where we thought we were going to be able to stop doing expensive customer surveys, because we had this wonderful administrative data now. Actually, we have not gone down this road for most things just using administrative data. The customer survey shows that it and the administrative data produced very similar results. Where you got significant results in one, you tended to get significant results in the other. We'll probably skip that.

This is about what the implementation told us. This is very unfamiliar terrain, ran across the day-to-day target structure, leading to a lack of management buy-in, so it really needed some intervention from the Centre to keep things on track; and again, I think there's an important lesson here. I remember addressing a conference just as, about six months into the program, a conference of ERA implementation managers from each of the districts was held, and instead of talking about the program, or the design, or how hard it was to do random assignment, they all complained to me, "Well, we were supposed to have this ring-fenced money, and it's not there because my managers have taken it away to use it to hit their targets." My only real contribution to this whole study, to be quite honest, was to go back and ring up the Jobcentre Plus finance director and say, "Look, this is important. You cannot allow this to happen. I want that money to be ring-fenced now, and I want to be sure that this program is being implemented as we designed it." Ensuring that someone has the power and the incentive to do that is an important thing about program design.

This just shows the usual things—that people didn't know this program was going on. Most of them knew that these bonuses were available. The interesting thing—and this is a general labour-market thing, I think— is that when it comes to training and skills acquisition, you can offer people things, but take-up tends to be low; and that was the case here, too. Again, it's very difficult to disaggregate impacts, but I think we can say that the primary root to impact here was not directly through training, and certainly not for the New-Deal-for-Lone-Parent group. For some of the people who were already in work, there was a bit more enthusiasm. What did get really high take-up was the personal-advisor-type support, where take-up for both groups was quite high.

We've already seen impact estimates for two or three programs, and they all sort of look the same so I'm not going to take you through this in great detail. If you are interested in the specifics of this program, we have hundreds and hundreds of different impact estimates in the report. Basically, the impacts are quite strong, and quite significant—particularly on full-time employment. Remember that all these people spent some time in a bit of (often part-time) employment, but their chances of working full-time were significantly bigger if they were assigned to the program, and correspondingly therefore, there was a pretty significant impact on earnings. You can see that New-Deal-for-Lone-Parents groups (and a few others) made a substantial impact on total earnings over the two years following random assignment. It's notable that the impacts in the second year look to be just as big as the impacts in the first year—certainly not significantly smaller—so we are seeing something that looks to be reasonably persistent.

The evaluation will continue until about 2011, I think, so we've got several years more of impact estimates, and we'll see whether any of these impacts begin to fall away; but so far we are reasonably optimistic that this isn't just a purely transitory impact. In the world of labour-market programs, where effects do tend to decay quite a bit over time, generally, that's something to be quite pleased about. That's just a counterpart of the chart, looking at Benefit Receipt as opposed to Earnings.

What do we conclude? We conclude that we can do this sort of program—that even a big, inflexible organization like Job Centre Plus can deliver both random assignment and, on the basis of random assignment, work support that makes a difference. The gender differences are quite striking. The impacts for the New Deal 25 Plus group show the men's group, the male long-term unemployed workers, are much smaller and generally not significant. This is not a "boys versus girls" but a "women versus men" distinction, but it is very similar to the gender differences that we saw earlier on. I don't know whether anybody's going to come up with a general theory on "Gender Differences for Social Policy Interventions," but there's clearly an interesting empirical regularity there, which seems to apply in quite a wide variety of fields, so there's certainly a paper there for somebody.

Where we go from here is to look at the longer-term impacts. We haven't done the cost/benefit analysis that you saw in the previous talk yet, but we expect to; and we want to continue to do—on the basis of the data we've collected—non-experimental analyses to try and "un-pick" some of the different impacts, and what was it that actually worked, what made a difference.

Going back again to the question about how you turn this into policy somewhat occasionally, against the better instincts of my research staff, we have already effectively decided to roll out some of the elements of ERA nationally. On the basis of these results, even though evaluation will continue for some time. For example, the in-work advice and support—the access to a Job Centre Plus advisor for lone parents (once they move into work)—we are moving to national roll-out effectively this year. In accordance with the last talk, we are getting reasonably good at taking the results of these experiments and translating them into policy in real time. We won't wait until 2011 to turn this into mainstream policy. That may mean we

make some mistakes, because if you don't have full evaluation it's not perfect. We may roll it out nationally and it will be taken in the wrong way, or will have taken some of things that aren't quite right; but still, that's better than waiting for 2011, by which time the current set of ministers and the current set of officials will have moved on and will no longer have so much invested in this particular program.

That goes to my last slide. You can do this, and you can use it to provide convincing evidence, but you have to be careful. You have to think about how—if you get anywhere from this—how you are going to translate it into policy? What is both the internal and external political economy of turning an experiment into a policy that uses this in the way it's supposed to be used, which is to come up with things that are genuinely timely and generalizable to the wider social policy context in which you're operating? I'd be very interested to hear—and perhaps Esther will tell us tomorrow—how that works in a developing country context, where there are these same issues of political economy and governance. So, I think that will be enough. Thank you.

Retour sommaire

Etienne WASMER, OFCE

First, thanks a lot for the invitation. I think this is a very important to have such a conference here in France, where we are not so much used to evaluating policies. The paper I will discuss is another example of a Work-fair program, which is called ERA, which has been applied in a country with a 5.2% unemployment rate and had a 75% employment rate, which is quite impressive over the period of application of the program 2004 to 2006. Mostly, I will discuss two of the three targets—the lone parents especially (mostly women), who are living on income support; and part-timers, who'd like to access full-time jobs.

It's a fairly common workfare program as far as I can judge, which consists of three payments of £400 a year for those in continuing full-time jobs, training tuition subsidies (the combination of the two is quite interesting), and emergency income. First, you can see that the British don't have much to envy to the French in terms of acronyms for schemes (we have or had SMIC, RMI, PPE, and so on). There is however a big difference with France, which is that in France, we don't evaluate our acronyms, and it is sometimes even worst than that: sometimes we don't get, as researchers, access to the data that would allow us to evaluate those schemes. I just want to tell an anecdote, quite relevant in this context.

At some point, the *Conseil d'analyse économique*, which is the institution advising the Prime Minister, was discussing the *politique de la ville*; in one of the reports we were surveying, we found a long discussion on the fact that this program was fine because we had spent billions of French francs in deprived urban areas. And right after the discussion, there was a footnote saying that—due to a lack of credit—it had not been possible to evaluate these policies, unfortunately. I think it's really important for us to think about why we don't want to evaluate policies in France and try from now on to organize proper evaluations with treatment and control groups before generalizing them.

Now, on the ERA scheme—as I tried to say from the first slide—I will have questions about why we want to help lone parents (and lone mothers in particular) to work more, in a country that has already a pretty efficient labor market. There is already a 75% employment rate, and additional employment has an opportunity cost. But, despite the fact that it could be difficult to have lone parents back to work, it's a rather efficient scheme. It appears from the report that the differences in earnings and subsidies to individuals between the treated group and the control group are relatively significant—the treated receive higher earnings, and the welfare subsidies are smaller for them.

As to the employment effect, I am more—not skeptical—but circumspect: it was a much less clear effect, at least from the charts you showed (which was on the variable "worked ever" during the 24 months of the program). It seemed to be significant, but when you look into the tables—which are in part 5 or 6 of the report I have to look at—it happens that it's only on two regions: Northeast England and Northwest England. There is an issue, obviously, of the heterogeneity of the effect—why it has been more important in some regions, and apparently insignificant in other regions. And for the other targets, I found it less spectacular, but we can discuss that too.

An interesting and novel contribution of this program compared to previous labor-market policies is that it focuses on new issues such as attempting to reduce the turnover rate of employed workers. We know that the typically eligible worker has a high turnover— they're stuck in low-pay/low-skill jobs, so there has been some effort in ERA to try to improve on those dimensions. That was probably an adaptive dimension of the program, and so certainly that kind of training and advancement and to try also to keep people at work.

In the report, there are very few methodological details on the econometrics. The statistical techniques employed here are rather conventional (these are standard difference-in-difference methods) and there is little worry about it (if anything, there are academic discussions only on the computation of standard errors on estimated coefficients).

I'm going to discuss more about the possibly undesirable effects of ERA, which are not really discussed in this report, but that could be relevant.

First, one thing that could be discussed is why those programs might fail in some geographical areas, and the second is why those programs might have other undesirable effects on the lives of people onto which they are applied. The last point is very simple—we are targeting lone mothers with kids, but is that such a good idea, beyond moral reasons (that is, the idea that those who receive benefits should contribute to society)? We can imagine why we want people back to work, but at some point there will be a cost to them: a private cost such as sometimes two hours of transportation, a social cost such as less time devoted by the lone mothers to kids' education and control. I don't know if this issue is addressed, and there was very little on that in the report.

There was something on the wellbeing of children, but it was not so easy to understand what was going on—basically, there was no effect. There could be many other dimensions affected here—it could be that women who are working leave the kids alone, so they can do all sort of

things, or they could also have more incentives to have additional kids. These are dimensions that, of course, are difficult to estimate because they are more long-term ones, but an exhaustive evaluation of the ERA program should provide a view of this kind of effects and mechanisms.

Regarding the econometrics, it is important to consider the heterogeneity of the effects of the program. Indeed, group variation is always very useful for us economists to know why the effect of the program changes from groups to groups. Of course, the average effect is extremely important, because it gives us an evaluation of the return to public funds, so these are obviously important dimensions in the policy debate. But if we want to know about the replicability of those schemes, we would like to know whether it works everywhere and for everyone, or if it fails in some places or for some groups—we actually learn a lot from investigating the heterogeneity of the effects. For instance, it could well be the case that, for local areas with few or no vacancies, the low employment rates are due, not to constraints on labor supply, but to constraints on labor demand; and if there are no jobs, incentives provide by ERA would therefore not work very well. This economic cycle could also affect the efficiency of the program.

So, just to think about these things in a more formalized way, just think of these programs as some fine shift in the labor supply curve. Here you have the labor supply curve. You shift it to the right, because you give incentives for people to work; but then the labor demand that you have from the local labor market might be more or less elastic, that is for instance, it could have an important slope so in a case where, in this dashed line, you have a relatively inelastic labor demand, the effect of the policy might not be that large. In addition, you can also see that there is a possible decline in aggregate wage in the whole labor market. That brings me back on wages.

An issue which I didn't see discussed, but which is always important to these kinds of programs is whether the employer knows that the recipient received those benefits, in which case he might have incentives to cut down wages. This is another possible effect.

Now, a better model to think about these issues is to have a better representation of the dynamics at work. Here we have job creation and job destructions. What we want here is to help people in going back to work. To the extent that we have job vacancies, this is going to shift the job-creation curve, but it could also be that there are no job vacancies in some places, and so the scheme is basically useless. It would be better, from a policy viewpoint, to help people move to another region area, instead of trying to provide incentives to find a job when there is none. You also saw some retention effects in terms of job destruction costs, so overall, you might have a positive effect, but it really depends on the number of vacancies available.

Now just to finish on this, let's look at the two regions for which the effect was the strongest in terms of access to work during the 24-month period: Northeast and Northwest. These two regions happen to be fairly similar to the other regions in the U.K. if you except London, they have relatively low GDP per capita compared to the average. But—given the fact that London is so rich—they are essentially comparable to other regions in the UK.

In terms of unemployment and dynamics, they are relatively different. Northeast is a region that had a high unemployment rate with respect to the U.K. average. Northwest was relatively closer in the beginning of the period, and the dynamic was very divergent, too. For instance, Northeast went from 6.5 to 5.8, which was relatively closer to the U.K. average over this period, while Northwest actually had an unemployment rate that increased. So, it's not that one of the regions was booming—the two regions were booming—and that helped people in the program to get jobs. It's actually more general than that—there were some things in those regions which were different, and despite that, the two succeeded in getting people back to work. In the other regions, ERA did not succeed.

One thing on the Canadian S.P.S.S. (Self sufficiency program): it's a relatively similar program to ERA. It is not exactly the same in the details, but it implements the same kind of idea, typically a quite large wage subsidy for lone mothers. You have here the employment rate of the treatment group (which is a solid line), of the control group (for which employment has been increasing continuously), and an impact (which is very strong and positive in the beginning, and then declines slowly over time). One of the things I always have in mind when I see this chart is that the reason for which the Control group has had such a success despite the fact that it was not receiving the benefits is that there was probably a correlation with the economic cycle; and it could very well be that—similar to the point I was making before—in a region which is booming (Canada was booming at that time) we still have a stock of unemployed people which is relatively large. We have a large pool of non-employed lone-mothers, but the best take jobs first, so the initial impact of the program is very large; and once the best have picked up the jobs, we are left with people in the "bottom of the skill distribution". That's why it's very important to control for economic conditions and for local conditions, and to understand the effect of those programs and their effect to economic conditions.

This is the last slide to say the same thing. On the ERA impact by groups (if you look at it, it's very well detailed in the report): it worked well for minorities, it worked well for the "more-skilled-among-the-low-skilled" (that is, the people with no qualifications didn't benefit from the program), and it worked better for people with better health. And that's the usual question—it works fine for some of the people you want to help among the more-skilled, but not for others who presumably need even more help.

In conclusion, it's great news for the UK if politicians have in mind the kind of ideas behind an evaluation of a workfare program with a control group and a treatment group. In France, I suspect most of the time they have no clue at all. This is why I actually try to teach the principles of these methods in my introductory to economics course in first year undergrad in *Sciences po,* but it will take some time before it comes into the minds of politicians in France. One of the things to conclude about this ERA program evaluation is that the opportunity cost of the program is not discussed: in this case, the time spent at work for the mothers could have been efficiently used elsewhere (and could be used for education, for instance), so let's at least provide support and question those things. Financial support could have been used for alternative policies—especially in these high-employment-rate countries—for child care, with

potentially higher results and then the interaction between the micro and regional context. Nevertheless, this is an interesting experiment. Thanks.

Béatrice SEDILLOT
Thank you very much. We have one or two minutes to react to some of the questions.

Jonathan PORTES
There's an awful lot there. I don't want to take up too much time from questions, so I'll just concentrate on one point and try not to be too controversial or confrontational (though I tend to be on this subject). There is very little to no evidence in the U.K. that the labor demand curve is downward-sloping. The labor demand curve is as close to flat as you can get it. In other words, it doesn't matter where you live, it doesn't matter whether the economy is booming or not—your probability of getting a job in the U.K. is overwhelmingly determined by your individual characteristics. There is virtually no place in the U.K. where there aren't vacancies available for people who are employable and looking for work.

The longer I've been in this job (and I've been in it for about six years), the more of a dogmatic, "supply-sider" I've become. It's the supply side of the labour market—the individuals and how they behave and what their incentives, motivations, skills, and employability—that matters overwhelmingly. There are umpteen anecdotes that demonstrate that, the most obvious being that the lowest employment rate in the U.K. is found in the borough of "Tower Hamlets" (which is also where Canary Wharf, the city of London, and the largest concentration of financial jobs—and associated low-skilled jobs, actually—it's not just finance jobs are located). There are jobs coming out your ears in most of the areas. Most of the areas where unemployment—particularly unemployment among disadvantaged groups, like the groups we're talking about here—is high, are inner-city areas, where there is an abundance of relatively low-skilled jobs located relatively closely, so that's one point.

The second point on that (and this is the other area in which I have actually done my own original research) is the impact of immigration. We've had a huge labour-supply shock to the U.K. over the last three or four years, in terms of low-skilled workers or migrants from the new member states coming into the U.K.; and they have had, as far as we can tell, absolutely no impact whatsoever on the job prospects of natives, either positively or negatively. Your chances of getting a job are determined, not by demand or economic conditions, but overwhelmingly by the individual, which is why active labour market policy programs like this one are so important. There are lots of other things I could also point to.

Etienne WASMER
I want to believe you, but in the six areas, it works in only two of them; and that's why I think the point is relatively interesting to make, at least.

Jeffrey KLING
Is there any evidence on the other questions that he was asking about—the spill-overs onto the kids or things of that sort?

Jonathan PORTES
There's a table at the back of the report with some figures which show that, when you asked the mothers about the kids, you get virtually no significant differences in anything that you really care about very much. There's like twenty outcome variables a year, one of which—swimming, gym, or dance lessons—there appears to be a significant negative impact. And, slightly bizarrely, there's a rather large and hugely significant impact on drug use, which is that, in the Treatment group, drug use is very small (it's only about one percent), but it goes from one-and-a-half percent in the Control group to zero—literally zero—in the Treatment group. I find this almost impossible to believe, that the mothers going out to work stopped the relatively small number of kids who were using drugs from using drugs, but that's what it says.

Anyway, leaving that aside, the answer is: this experiment doesn't tell us very much. I think that there's quite a lot of other research evidence in the U.K. (and, to some extent, in the U.S.) that, in my view, certainly doesn't show any negative impact on child outcomes from lone mothers going out to work—at least after the kids are older than one or two. But I think there's an important interaction here with the availability and quality of child care. I'm not saying the outside environment doesn't matter, I'm saying labour demand doesn't matter. The outside environment does matter. One thing that's important to remember about this experiment is that it was contemporaneous with an absolutely massive expansion of state child care provisions in the U.K. (through the Sure Start and Children's Centre programs), particularly in disadvantaged areas (including all of these areas), where we were literally trying to go from American levels of provision of child care to French levels of provision of child care in ten to fifteen years—a huge ramping up of investment. And there's clearly a major interaction between and that wasn't a sort of randomized controlled trial, but there's clearly a major interaction between this program and that - we wouldn't have done this program if we hadn't been ramping up the child care provisions at the same time. And again, there are no simple answers there, but it's a very important policy context for this program.

Béatrice SEDILLOT
Thank you. I am interested in coming back just a bit more on the process of implementing such experiments. We heard how important it was to have a governmental collaboration with research teams in order to be able to implement such experiments. Could you say a bit more about how the job center staff react regarding the Random Assignment process? It's a concern for us in France, you know. The job center staff may find it hard to admit that it's not necessarily unfair to make random assignment for a limited period of time. So were they positive to this? And my second question is: how were the participants informed on the random selection process? How do you deal with that?

Jonathan PORTES
I think the first question is a very interesting one. I think the idea here is—and we are very grateful to M.D.R.C. in particular for their advice, help, and support—you can't just drop this on people from the top and expect it to happen. You actually have to get people out there—talking to staff; explaining. You have to create ambassadors, and I think one of the interesting

things was that—in each of these districts—some mid-level management staff who had no economics or research background at all—once you explained to them why you were doing this, why it mattered, why it was important, how it would work, etc.—some of them became very enthusiastic, and became ambassadors for us, and were quite happy to sell that. You can do that. It was important to have messages from the center. It was important to the politics of it. You have to do the work on the ground—you can't just flick a switch at a head office and expect it to happen.

In terms of informing the participants, I'm not actually familiar with the details of how that was done, but I think people were safe in their work offering this program, and this has become quite culturally embedded now in Jobcentre Plus that there are a lot of pilots going on. Every district has three or four pilots of one sort. Not many of them have Random Assignment pilots, because we don't do that very often; but the idea that there are pilots—that things are happening in different parts of the country, that things are happening in some places but not in others—is quite well embedded, and even our customers understand that that does happen some of the time, and that there are good reasons for it. It's become a part of the culture to some extent.

Béatrice SEDILLOT
Merci. Nous allons maintenant passer à la deuxième partie.

Nous partons en Norvège, pour une expérimentation sur un champ très différent : les troubles musculo-squelletiques. Je laisse tout de suite la parole à Astrid Grasdal, de l'Université de Bergen, qui va nous présenter l'expérimentation qui a été menée en Norvège. Ensuite, Philippe Askenasy, du CEPREMAP, introduira la discussion.

Retour sommaire

« Bergen Experiment »

Astrid GRASDAL, Université de Bergen

Thank you. Let's move to Norway—to the west coast of Norway. Bergen is the second largest city in Norway, but it's a small city in the European context. I'm going to tell you about two experiments that took place in Norway in the 90's. The second experiment is a follow-up of the first one. I'm going to tell you a little bit about the background for these experiments—the assignment data, and treatment effects; and, finally, what I find to be some of the most important lessons from these experiments.

First of all, thank you very much for inviting me to this conference. This is really a pleasant surprise, to be invited here, because I haven't been working on this for a few years now. It was in my drawer, so it was nice to see that someone still found it in the literature.

Some background information first. Sick-leave rates are high in Norway, so we are now looking at a program designed to bring sick-listed workers back to work. During the 90's (and still), sick-leave is very high. It's really hard to tell exactly how many days workers, on average, are on sick-leave in Norway per year, because not everything is registered, but a

"guesstimate" is about twenty days on average per year, which are in addition to five weeks on holiday. Muscular problems (neck pain, back pain, generalized muscle pain) account for a substantial part of this. About 43 to 44 percent of all sick-leave days are due to muscular or skeletal problems, and about one-third of all new entrants into disability/pension also have musculoskeletal problems of some kind.

We have a very generous sickness benefit system in Norway. Most workers receive 100% compensation from day one, and they receive it for a whole year. Employers pay for the first sixteen days, and the rest is paid by the national social insurance. In addition, a worker in Norway does not risk being fired as long as he is on sick-leave, so the employment protection here is strong. For economists, it's quite obvious what has to be done here to motivate people to go back to work; but cutting compensation rates or allowing employers to fire people while on sick leave is not a political issue in Norway, so we have to look for a good Treatment instead.

And that is what they did then, in the early 90's. The question that was asked was "Is there a treatment that can reduce the amount of sick-leave due to musculoskeletal pain? Can we induce workers on sick-leave to return more quickly to work, and can we prevent them from becoming disabled?" The first experiment was designed to evaluate the "four weeks" treatment program for workers on sick-leave. This was inspired by a program that was conducted in Sweden. There, they have a similar program, working on the same target group, and they reported that 70 percent of those who'd been through the program returned to work. But they didn't have a Control group in Sweden, so they couldn't really tell how many had returned to work even without the program. In Norway, we wanted to have a Control group.

In order to do this experiment, a clinic was established in Bergen with a team consisting of a neurologist, a psychologist, physical therapists, and nurses (and secretaries). The Treatment here: four weeks, with a combination of a cognitive—as well as a physical—treatment. Participants came in groups, and they were treated in groups and individually. The Treatment was really about reducing anxiety about their problem, convincing them that taking up a normal daily behavior/activity was not dangerous to them, and that this was the best thing to do to cope with their problems.

In order to be included in this program, workers had to be on sick-leave a minimum of eight weeks, they had to live in the Bergen area, and they had to hold a permanent job; so that all of these are people who have a job to return to, full-time or part-time. They were invited by the local social insurance office by mail, so as soon as a person had been on sick-leave for eight weeks, he or she was invited by the local insurance office; and in the invitation letter, the experiment was explained. It was explained that they could end up in the Control group, and that then they would receive "treatment as usual," which would mean follow-up by a general practitioner and some physiotherapy during the period when they were on sick-leave.

Before treatment, they were tested outside the clinic (in the college for physiotherapy in Bergen) and they were also there on the follow-up twelve months after inclusion in the program. In addition to the data that was collected in this pre- and post-test, we also followed-up participants (and non-participants) with labor market outcomes from national

administrative records. Participants were included over a period of one-and-a-half years. In total, 1,648 were invited to participate. Of these, 560 accepted the invitation, and about two-thirds did not accept (or did not respond to) the invitation. Along with the invitation letter, there was a letter that they could return to the local insurance office, where they could say if they accepted or if they did not accept, so those that did not want to participate could also return the letter, making a cross indicating that they did not want to participate. They could give a reason as well, and the alternatives were: "I'm on my way to go back to work very soon," or "I don't think that this is a program that will fit me."

Of those who accepted the invitation, 358 were sent to Treatment and 202 to the Control group, so there was a Random Assignment in a 2-to-1 order, in order to insure that Treatment groups were always filled up and no one had to wait very long for treatment. There was an evaluation based on the follow-up data—a comparison of pre- and post-Treatment data—and this evaluation showed that the Treatment group scored better on some measures related to pain, functionability, and life satisfaction. However, these data are hampered by attrition—only 60 percent of the Controls showed up at the follow-up examination. So we don't know here really or we can suspect that this effect is affected by a selection bias.

When it comes to the evaluation of labor-market outcomes (which was based on registered data), then we don't have the problem with attrition, and we have, as you see here data on both those who decided that they would participate and on those who did not participate. In fact, we also have data on the non-responders, which was quite unusual; and I think, if this should be done again, we wouldn't have that data. In the invitation letter it was said that, if they did not do anything, they would be followed-up with registered data; so they would have had to actively say that they did not want to be followed-up through the register in order to avoid this. That is really on the edge of what is ethically right to do, and I don't think it would be possible to do that again. But it's very useful to have data on non-responders, and on those who do not want to participate—because we can see: "How are things going for them? Who are they?" And what we can see from these figures is that they are, on average, healthier than those who decide to participate—at least in the beginning.

If you look at those who participate, you see that there was no Treatment effect in this program, so although 70 percent returned to work in Sweden, here only 50 percent returned, and we see that there is no difference, really, between the Treatment group and the Control group. This came as a great disappointment to those who initiated this experiment and to those in the Treatment team and it really came as a surprise as well. Before the evaluation was done, it was already decided to run a second experiment, and they had this second experiment funded as well. I think it was good that they had that before the results came here, because otherwise, I'm not sure if the second experiment would have been done. But useful things were learned from the second experiment.

So, okay—no Treatment effect here, but the Treatment team had, during the period when they were treating people here this feeling that not everyone should have this kind of treatment, so although they had well defined inclusion and exclusion criteria from the beginning, they saw

that the group that they had into the clinic was quite heterogeneous, and that maybe not all could benefit from treatment.

In the second experiment, they decided to evaluate two different treatment programs: the four-week program, which was the "Extensive" treatment, against a one-day program (the "Light" treatment), and the Control group, who received treatment as usual in the primary healthcare sector. They used the same inclusion criteria and recruitment process as in the first experiment, so when we compare background characteristics for those in the second experiment with those in the first, we can confirm that they are really the same group—the same kind of participants.

The idea now is to use a slightly more sophisticated design. So, before randomization, a systematic and standardized screening was done in order to group participants according to prognoses for return-to-work. This standardized screening battery consisted of some physiotherapy tests, a questionnaire regarding a motivation to return to work, and beliefs on what has to be done in order to be able to return to work (what the person can do himself, or if the person needs someone else to do something for this person to go back to work). According to this screening battery, they were given a prognosis for return-to-work, and after the screening (and independent of the screening result), they were randomly assigned to Extensive treatment, the Light treatment, or the Control group.

The hypothesis now is: when comparing to "treatment-as-usual," it's assumed then that sick-listed workers with poor prognoses for return-to-work should benefit from the Extensive treatment, those with medium prognoses should benefit from Light treatment, while those with good prognoses for return-to-work should not benefit from anything in addition to treatment at the clinic (compared to treatment in the primary healthcare sector).

Based on the randomization, participants were distributed like this. We see that, in total, there are 628 participants, and if you look at the diagonal, we have those who received the "assumed right" treatment. We also see that, in each Treatment group, you have some with good, some with poor, and some with medium prognoses for return-to-work. If you just ignore the screening for a while, and look at how things are going, this figure can be compared with the figure I already showed you for return-to-work in the first experiment, and we see here that there is a Treatment effect of the Extensive program that lasts for a while. I think the reason why we see the difference here that we didn't see from the first experiment has to do with the treatment in the clinic having been improved. They have learned what to say to their patients, how to approach them, and so they have learned from what they experienced from the first experiment. Also, what they actually said was that maybe they had put too little pressure on patients in the first experiments, because they said, "Okay, take your time—take the time you need before you go back to work," and now they have been more eager to say "you should focus on returning to work." And I guess that is what we can see here.

But this is not really what they wanted to look at. What they wanted to know was if there were differences for the different screening categories, and if you look at the participants with a good prognosis for return-to-work, and look at return-to-work rates here, we see that there is

not really a systematic difference between the groups out here, which confirms the hypothesis, I think, that there's no point in giving the Extensive treatment—or the Light treatment program—to those with a good prognosis for returning to work. They manage well after all. For those with medium prognoses for return-to-work, the hypothesis was that they should benefit from the Light treatment program, and we see that they do, for a while. They also benefit from the Extensive treatment, but we see that there is no additional gain from the Extensive treatment compared with the one-day treatment. And for those with poor prognoses for return-to-work, we also see that there is a clear effect of the treatment here, and it lasts for quite a while. We have a lot of data for subsequent years, and they show that the effect is still there. It fades out after about three years.

Some lessons from the experiment? I think most of it has been said already, more or less explicitly, but I will sum it up anyway. Attrition from post-program follow-up may very well hamper the randomization, especially because Controls have a tendency not to show up for follow-ups; so if possible, it's very nice to have registered data, because then you are not that much affected by the attrition problem. And if you don't have registered data, it's really worthwhile putting a lot of effort into collecting data on follow-ups. It's also important to know what the Controls actually receive. One of the explanations from the first experiment (for why there were no Treatment effects) was that maybe the primary healthcare sector, which was aware of the fact that there was an experiment going on, were so concerned with that, that they really put additional effort into what they were doing, and tried to copy some of the activities that were done at the clinic, and maybe that was something that improved return-to-work for the Controls. That was one of the explanations for why there was no effect. But we had data on those who came to the follow-up examination—on how much physiotherapy they received (and what they had received from other kinds of treatment); and they had some more physiotherapy than the Treatment group, but then the Treatment group had some physiotherapy treatments while they were at the clinic, so I don't think that could really explain the difference.

If possible, collect information about those who fulfill the inclusion criteria but opt out of the experiment, so that we know who they are and what happens to them. We are also seeing today that it's interesting to know how long the effects last. In some programs—for some kinds of treatments and different kinds of target groups—we see that the effect lasts for a while, then disappears; and in others, it seems to be a permanent effect. It's really nice to have a long follow-up period. From the Bergen Two experiment, we also see that there is a heterogeneity in Treatment effects, and sometimes this heterogeneity is linked to unobserved characteristics. So if there is reason beforehand to think there is a heterogeneity in Treatment effects, it's worthwhile thinking a little about what this heterogeneity is, and if there's something one can do to collect data that makes it possible to identify such Treatment effects. Thank you very much.

Retour sommaire

Philippe ASKENAZY, CEPREMAP

Thank you for your very stimulating presentation. I think that such experiments are really original from a French point of view. In fact, in France, we still believe that improving transition to employment from non-employment status (for example, non-employment, sick-leave, and so on) that the policies should be mainly, on the one hand, classic training, and on the other hand, financial incentives. For example, the *Prime pour L'Emploi*, presented by Thierry Magnac this afternoon, or simply to cut compensations.

Today, we have a general strike in France against the pension reforms, and the argument of unions is really strong. But just say that previous reforms were unable to really increase the employment rate of aged persons in France; and when we try to understand the failure of these policies, we find that the subjective or objective health status of the aged workers in France were actively bad, and so I think we have an important room for alternative programs—maybe programs that could be both an efficient labor policy and also an efficient public policy.

In that perspective, focusing on musculoskeletal disorders is particularly relevant. Why? Norway is not an exception. In most European countries, in North American countries, etc., musculoskeletal disorders are the main occupational illnesses. For example, in France, they account for about two-thirds of the total occupational illnesses; and we know that musculoskeletal disorders induce large human and economic costs, so we have diverse values and evaluations—that's about one or two percent of GDP. The problem of musculoskeletal disorders has also induced extensive and intensive new disciplinary clinical research in order to determine the causes of musculoskeletal disorders to try to build preventions and also to improve return-to-work.

We know that the main determinants of musculoskeletal disorders are individual factors: behavior, gender (the risk is higher for women), and also collective factors. In fact, the workplace organizations— there is some paper by economists specifically on this topic—were combinations of mental strength and physical strength; and this also combined with a lack of specific training, specific to the workstation. We know quite well why we have so much musculoskeletal disorders, and the Bergen experiments also help, then, to understand how to improve return-to-work for people who suffer from these disorders. Now I would turn directly to the paper. This is a very interesting experiment, because there are various experiments, designed to improve the return-to-work of persons who suffer musculoskeletal disorders, but generally samples are small—about 100 workers.

Here, we have a quite large sample—about 500 persons. We have both light and in-depth programs, reminiscent of two phases and also quite a medium-run study in general. The study is just on short run, so this is very, very interesting; and I think the results are quite positive. They prove that we may have significant positive results for participants with poor prognoses, and I think it's really important. This shows that we could have a double outcome—both for return-to-work and to have less pain for these workers.

These results are globally consistent with previous experiments—again, based on smaller samples—in fact, there are dozens of experiments mainly conducted in the U.S. or in Canada. Recently, the European OSHA has collected all these studies and they conclude (this is the example for back-pain disorders—musculoskeletal disorders—neck, shoulders, and back pain: the three main types of musculoskeletal disorders) that we know that we could have a real positive effect when you have a multi-disciplinary approach, including cognitive behavior. The Bergen experiments are perfectly consistent with this result only for chronic low-back pain—in one sense, that is close to your category of "Poor Prognoses". This is not really surprising. The efficient programs are programs which modified work, and this is, in one sense, natural, because we know that the main cause of the chronicity of musculoskeletal disorders are individual factors, and mainly the behavior of the workers, while the acute musculoskeletal disorders are due to the organization of work inside the firm. So the good policy is to improve the organization of work inside the firm—a sort of collective policy to change the workplace organizations—and the interventions only at the individual level are efficient for chronic situations. You can see there is a sort of scientific consensus around what we have to do in order to improve return-to-work for these workers.

I have some important questions: why do we have this consensus, and why have these programs never been recognized? They seem efficient, yet they have never been generalized. There may be two reasons. The first one is that the scientific consensus may be not-so-strong; the second one may be that also the samples are still to small to conclude. Can we trust weak results from small samples?

The second issue is about the cost benefit of these programs. We have generally poor information on the cost of these programs, and also on their benefits. We don't know if they are more efficient than standard tools (incentives, and so on). So I think that maybe the next frontier of the research is to merge both economic experiments and health experiments with, for example, three groups: one control group, one group involved in a health-program, and one group involved in an efficient incentive program. With such experiments, we could be able to disentangle the incentive phase and the rehabilitations, and improve behavioral phases. Thank you.

Béatrice SEDILLOT
Do you want to add some comments on Philippe's?

Astrid GRASDAL
Maybe I can respond to your last comment about cost-effectiveness. These experiments, or the second one—for there was no point in doing a cost/benefit analysis on the first one—but on the second one, the cost/benefit analysis clearly shows that it pays off to do this, although the Treatment Effect fades out after three years. It's been published in medical literature but it's true—there are not many publications or analyses where you also find the cost/benefit analysis. But I think, in general, if you have a program that is able to bring people back to work (where you really have a Treatment Effect), then the value of having people working rather than not being working—the value of increased production for the society—is so high that the Treatment program has to be really costly for this not to be valuable to the society. As

long as you have the Treatment Effect, you can almost always conclude that it's beneficial for the society.

Béatrice SEDILLOT
Y a-t-il des questions dans la salle ?

Un intervenant
The question is, do you take into account any of the variables that some people could cheat or fraud the system and could have influenced your evaluation? In the Volvo car company, for example, the TMS's were very expendable, so even some groups were excluded from recruitment because of that. I wondered if you took that, in your panel, as a question, because it's well known in Sweden that they have experienced such trouble with TMS. Do you understand what I mean? I mean that people who don't want to work just stop working. The first time, in Sweden, they were paid, and nobody cared; but the second time, they introduced a law not to pay for the first three days, so they tried to change the thing.

Astrid GRASDAL
Okay. Here we cannot really know if they are not able to work, or if they don't want to work, because a quite substantial part of this is subjective health complaints—they cannot really be observed by objective measures, so we cannot really tell. But for the Treatment effects, you shouldn't expect that. Those who do not want to work are equally distributed among the groups, so it should not affect Treatment effects, I think. But I do believe that, as long as we have 100% compensation from the first day, people do not have incentives to hurry up and go back to work, and that obviously affects the duration of sick-leave periods.

Béatrice SEDILLOT
I have just one last question. Do you think the experiment had any sort of consequences for policy-making in Norway regarding TMS and, if not, why?

Astrid GRASDAL
While the clinic is still there and they are treating sick-listed workers according to the screening battery, they do not treat those with good prognoses for return-to-work. They concentrate on those with a poor prognosis for a return-to-work. That means particularly women with generalized muscle pain. They are the majority in that group. It's always difficult to build this out to a national scale because much of the treatment factor is about experienced personnel, and it takes time to obtain that experience. Still, in Norway, the treatment effect from the first experiment is the one that people remember, because that came as a big surprise, as I mentioned. It was really a problem, in the next round, to convince politicians and official bureaucrats that there were Treatment effects. Very often, when I meet someone, they say, "Oh, that was the experiment where there were no Treatment effects," so it has been a problem to convince the public that there was a Treatment effect, after all.

Béatrice SEDILLOT
Thank you very much.

Peut-être deux, trois mots pour conclure cette première journée, tout d'abord en remerciant l'ensemble des participants, notamment les intervenants et les discutants pour leurs présentations très stimulantes. Un ou deux commentaires, peut-être un peu personnels sur cette première journée de colloque. On a vu un ensemble de travaux évaluatifs sur des sujets assez variés mais qui ont tous un certain nombre de caractéristiques communes. La plupart des travaux aujourd'hui présentés ont montré qu'ils s'inscrivaient dans la durée, qu'ils se donnaient le temps pour l'évaluation, qu'ils menaient aussi des analyses sur un ensemble assez riche de variables d'intérêt – pour les politiques du marché du travail, sur le retour à l'emploi mais pas seulement – qu'ils essayaient fréquemment de combiner des analyses en matière d'efficacité, mais aussi d'efficience, avec des analyses coûts-bénéfices dès que cela est possible. Pour cela, ils s'appuient sur des systèmes d'information – on n'en a pas énormément parlé, mais on en parlera peut-être plus demain – assez riches et développés, qui mêlent des données administratives, des enquêtes de type statistique, mais également des enquêtes qualitatives. Ce sont souvent des processus conçus avec le temps nécessaire à ces expérimentations, suffisamment en amont, en associant si possible décideurs et chercheurs. Tout cela présente des enseignements très riches dans le contexte français puisque, comme l'a rappelé le Haut Commissaire ce matin, les pouvoirs publics souhaitent développer de façon plus systématique l'évaluation des politiques publiques. Toutes ces expériences qui ont été présentées sont pour nous très riches d'enseignements.

Ces expérimentations sont plus ou moins simples à construire, cela dépend des contextes. Elles supposent une forte adhésion, ce qui pose parfois certaines difficultés. Nous aurons l'occasion, demain, de discuter plus précisément de ces sujets, en revenant sur la partie pratique de mise en œuvre de ce type d'évaluation utilisant les méthodes d'échantillonnage aléatoire.

Je vous donne rendez-vous demain pour la suite de nos travaux, à 9 heures 15. A demain.

Retour sommaire

Aspects opérationnels des expérimentations

Ouverture

Antoine MAGNIER, directeur de la DARES

Hier, après l'intervention d'ouverture du Haut Commissaire, nous avons eu une série d'exposés très éclairants sur les expérimentations menées à l'étranger : des exposés généraux sur l'apport des expérimentations, notamment dans le domaine de l'économie du développement, également des exposés sur un certain nombre d'expérimentations importantes menées à l'étranger, dans le domaine de l'éducation et de l'emploi, des exposés sur les objectifs de ces expérimentations, les méthodes déployées et les résultats que l'on peut en tirer. Ces exposés nous ont permis de toucher du doigt les potentialités qu'apporte cette démarche expérimentale, potentialités sur le plan analytique d'une part et puis, nous l'avons vu, potentialités pour ce qui est de la capacité de ces expérimentations à influencer la prise de décision publique dans ces domaines.

Ces exposés nous ont aussi permis de toucher du doigt les exigences associées à ce type de démarche et les difficultés, parfois, de mise en œuvre opérationnelle de ce type d'expérimentation. Nous avons aussi pu voir certaines limites inhérentes à ce type d'approche.

Je souhaiterais une nouvelle fois remercier les intervenants, étrangers notamment, qui nous ont particulièrement éclairés, hier, sur leur expérience des expérimentations qu'ils ont pu conduire ou étudier à l'étranger.

Aujourd'hui, nous allons revenir sur ces questions de nature méthodologique, opérationnelle, juridique et aussi éthique, dans une certaine mesure, de manière un peu plus systématique, ce matin, sous la présidence de Yannick Moreau, présidente de la section sociale du Conseil d'Etat, qui nous fait l'honneur de nous accompagner et de nous guider sur les aspects opérationnels et juridiques associés aux expérimentations, notamment dans un contexte français.

Cet après-midi, nous aborderons certaines des expérimentations qui sont aujourd'hui menées en France dans le domaine de l'emploi, ainsi que les expérimentations qui vont se mettre en place, notamment dans le domaine de l'insertion sociale, sous l'impulsion du Haut Commissaire.

François Bourguignon, de l'Ecole d'économie de Paris, pour finir, nous dira quels enseignements il tire aujourd'hui de ce type d'approche, déployée dans le domaine de l'éducation, de l'économie du développement, de l'emploi. Esther Duflo nous présentera également le bureau Europe du laboratoire d'action contre la pauvreté, qui vient de se mettre en place. Pour terminer, Thomas Fatome, directeur de cabinet de Laurent Wauquiez et de Christine Lagarde, nous apportera leur voix pour conclure ce colloque.

Je voudrais terminer par deux points concrets et pratiques. Le premier, pour dire que les supports d'intervention des intervenants seront disponibles dans les jours qui viennent sur

notre site Internet. Les actes du colloque seront également disponibles dans quelques semaines. Le deuxième, je demanderai aux différents intervenants, aujourd'hui, de ne pas parler trop vite, afin que la traduction puisse se faire convenablement, de sorte que nos partenaires étrangers puissent suivre nos échanges et nos débats dans le détail. Je vous remercie et je passe la parole à Yannick Moreau, qui va présider la première séance.

Retour sommaire

Yannick MOREAU, Conseil d'Etat

Bonjour à tous, merci à Antoine Magnier et à l'équipe de préparation du colloque d'avoir fait appel à moi. C'est une manière d'assurer ma formation permanente, sans doute.

Avant de donner la parole aux intervenants, parce que ce sont vraiment eux qui ont des choses importantes à vous dire, je souhaite vous dire quelques mots.

Pour la juriste que je suis, l'expérimentation n'est pas un gadget. C'est peut-être l'un des moyens d'éviter que l'on ne fabrique du droit inutile, voire pire du droit nuisible. Car le droit instable est du droit nuisible. Cela n'est pas un propos strictement personnel. Le Conseil d'Etat, dans un rapport annuel d'il y a deux ans, a vraiment insisté sur la difficulté que nous avions, en France, avec l'instabilité du droit. Ceci est parfaitement vrai dans le droit de l'emploi et notamment des aides à l'emploi. Nous n'avons souvent, la plupart du temps, pas d'étude d'impact suffisante. Des expérimentations seraient très utiles pour constituer des études qui mériteraient ce nom.

Je ne mésestime absolument pas la difficulté d'expérimenter. C'est parce que cela est difficile que vous vous réunissez, j'imagine, pour voir comment non seulement expérimenter, mais expérimenter bien, en sachant en tirer les vraies leçons, sans extrapoler au-delà.

Nous allons tout de suite passer au vif du sujet et à l'exposé que va nous présenter Esther Duflo, du MIT, sur la boîte à outils, la définition, les objectifs, la mise en œuvre des expérimentations. Il est très intéressant de voir concrètement ce que l'on peut faire.

Définition, objectifs et mise en œuvre des expérimentations : la boîte à outils

Esther DUFLO, MIT

Merci beaucoup. Je voudrais remercier Dominique Goux, Antoine Magnier, Julie Moschion d'avoir organisé cette conférence. C'est vraiment une occasion formidable pour présenter nos travaux.

Hier, nous nous sommes essentiellement concentrés sur un certain nombre de résultats de projets, sans vraiment s'attarder sur la méthode ou sur les aspects pratiques. En quarante minutes, je n'aurai certainement pas l'occasion de dérouler l'ensemble des aspects pratiques permettant de mettre en place ces évaluations. Je voudrais attirer l'attention sur un certain nombre de points, pour peut-être rendre les choses un petit peu plus complètes sur la manière selon laquelle cela se passe sur le terrain : est-ce que cela est possible, est-ce si difficile, comment introduit-on la randomisation d'une manière à la fois faisable, compatible avec les exigences de l'implémentation du programme et éthique ?

Comment introduire la randomisation dans un programme ? Comment trouver les partenaires ? Quels sont les partenaires ? Hier, on a essentiellement parlé de partenariats avec des gouvernements ; quels sont les autres partenaires possibles ?

Quand on introduit la randomisation dans un projet, on veut le faire de manière à maintenir la rigueur de l'expérimentation, tout en restant pratique. On fait souvent une analogie avec les méthodes de laboratoire, l'industrie pharmaceutique, mais il est clair que l'on parle en fait de quelque chose d'assez différent. On n'est pas dans un laboratoire. On contrôle assez mal ce que les gens font. Comment réussir à marier ces deux exigences ?

Ensuite, je parlerai rapidement de quelques problèmes auxquels on a un petit peu touché dans les discussions hier, mais de manière un peu accidentelle. Je voudrais mettre l'accent sur les problèmes auxquels on fait face quand on analyse les données d'expériences qui ne sont pas des expériences de laboratoire, mais qui sont des expériences de terrain, où on a des problèmes constants. Comment éviter ces problèmes et comment les résoudre ? Puis, je parlerai quelques minutes sur certains choix, en relation avec une question que Jonathan Portes a posée hier : faut-il se mettre au niveau de la communauté ou au niveau de l'individu pour introduire la randomisation ?

A propos des partenaires, hier, le Haut Commissaire a fait référence à l'appel à projets que le Haut Commissariat avait lancé. Si l'on regarde la liste de ces projets, on peut voir que tout un ensemble de partenaires sont prêts à s'engager dans des expérimentations. Il est bon de garder ces partenaires en tête, parce que, même si l'on pense à expérimenter quelque chose qui pourrait *in fine* se traduire dans la loi et donc comme un projet applicable à tous et qui serait appliqué par le gouvernement, la phase expérimentale, la phase pilote, on peut tout à fait la concevoir comme étant mise en œuvre, soit directement par le gouvernement, mais aussi par les collectivités locales – on a parlé de Tulle hier et on en reparlera cet après-midi –, les organisations non gouvernementales – qui sont souvent des réservoirs d'idées extrêmement riches dont il faut s'inspirer et qui ont le grand avantage de ne pas avoir vocation à toucher tout le monde par leur programme. La question juridique à laquelle le Haut Commissaire a fait allusion hier, qui apparaît en partie réglée, ne s'est en fait jamais appliquée dans le cas des ONG qui, de toute façon, n'ont pas besoin de toucher tout le monde, puisque telle n'est pas leur vocation. Elles sont aussi dans des situations où elles ont en général des budgets extrêmement limités. Donc, de toute façon, elles ne toucheront pas tout le monde et de toute façon elles sont dans une situation où il faut qu'elles trouvent un moyen de gérer la pénurie. Il est beaucoup plus facile d'introduire un élément aléatoire de sélection des bénéficiaires quand il y a une pénurie derrière. Il faut donc penser aux ONG, même si l'on pense aux programmes des ONG comme potentiellement des pilotes pour des politiques qui, après, peuvent être mises en œuvre de manière plus générale. On peut aussi penser aux compagnies privées, qui peuvent avoir des programmes à expérimenter. En particulier, on travaille avec Veolia, par exemple, au Maroc, sur des questions de raccordement aux adductions d'eau. Tout cela, pour rappeler qu'il existe tout un tas de partenaires possibles et qu'il faut les garder en tête.

Deuxième question que je voudrais aborder : comment introduire la randomisation dans le projet ? Quelles sont les contraintes ? La première, c'est qu'il faut que cela soit opérationnellement faisable, c'est-à-dire que cela soit compatible avec les objectifs du programme et que cela puisse être implémenté de manière simple par les gens sur le terrain. Sinon, cela ne va pas fonctionner. Il faut que cela soit simple et que cela ne demande pas beaucoup de travail à la personne finale. Sinon, cela ne se fera pas ou ne se fera pas bien. Il

faut que cela soit éthique et équitable. Il faut aussi que cela soit perçu comme équitable, ce qui est un peu différent. Il existe des processus que l'on peut considérer comme équitables en tant que chercheur ou administrateur mais que les bénéficiaires ou les administrateurs locaux ne percevront pas ainsi. Ensuite, il faut conserver la capacité de comparer deux groupes, où, à un moment ou un autre, on a une assignation aléatoire.

Je vais passer en revue assez rapidement différents types de solutions, adaptées à certains cas et inadaptées à d'autres, qui peuvent maintenir toutes ces contraintes ensemble. La première et la plus simple correspond au modèle pilote aléatoire simple dont on a vu des exemples hier. Presque tous, hier, y correspondaient : le projet *Moving to Opportunity*, le projet de Joshua Angrist, les expériences de Bergen. Tout est explicitement conçu comme un pilote. La randomisation se fait au niveau individuel. Manifestement, c'est un ensemble de cas où cela a paru possible pour tout le monde. Quand il en est ainsi, c'est bien. C'est le modèle le plus similaire à un modèle médical et le plus simple à analyser. La seule différence avec le modèle médical, dont on a parlé un petit peu hier, c'est qu'ici, en général on ne force pas les gens à suivre le traitement qu'on leur propose. Ce n'est qu'une proposition. C'est le modèle de base, celui dont on a le plus parlé hier. Mais ce n'est pas le seul modèle.

Un modèle similaire, mais qui s'applique dans des cas différents, est à nouveau un modèle de pilote complètement aléatoire mais où la randomisation est au niveau du groupe. Ce ne sont plus des individus qui sont désignés comme étant des individus contrôles ou des individus traitement, mais des communautés ou des écoles. Tous ceux qui sont éligibles pour le programme dans cette communauté y ont accès. Un exemple d'hier : les expériences d'incitation en Israël, le programme PROGESA au Mexique, les expériences sur la vaccination présentées par Abhijit Banerjee.

Il existe deux manières de concevoir ces projets aléatoires où la randomisation se fait au niveau de la communauté. Soit le programme est défini au niveau de la communauté, par exemple au niveau de l'école, alors la randomisation se fait à ce niveau. Soit c'est un programme qui s'applique à des individus ou à des familles, comme dans le cas de PROGESA, mais on pense qu'il ne serait pas faisable, pratiquement, d'avoir dans une même communauté des gens concernés par le programme et d'autres qui ne le sont pas ; même quand on pense à un programme individuel, par exemple un programme d'accompagnement des chômeurs, on peut concevoir un choix entre choisir comme niveau de randomisation, le niveau individuel ou choisir le niveau de mission locale de l'emploi. En principe, le programme s'adapterait aux deux, mais après il faut voir, d'un point de vue pratique, ce qu'il sera le plus facile à faire accepter et à mettre en place de manière rigoureuse. Ce sont les modèles les plus simples, mais il existe d'autres manières de procéder. Quelque chose dont Abhijit a parlé hier : l'introduction progressive aléatoire. La différence avec un pilote, c'est que tout le monde, dans l'étude, sait qu'à un moment ou un autre, il va être bénéficiaire du programme. Et, en général, tout le monde le sait à l'avance. Le programme va être introduit progressivement sur un certain nombre d'individus ou sur un certain nombre de communautés si la randomisation se fait au niveau de la communauté et l'on choisit au hasard l'ordre d'entrée. C'est quelque chose que, dans nos travaux, dans l'économie du développement, l'on emploie énormément. Parce que, en particulier quand on travaille avec des ONG, c'est une

contrainte qui apparaît complètement naturelle, à la fois aux ONG et aux bénéficiaires. Beaucoup de programmes progressent de cette manière. D'après mes discussions avec des ONG ici en France, il en va de même : on ne va pas partout au même endroit. On a des limites de capacité, un certain nombre de membres dans l'équipe, un certain budget, etc. Un plan d'expansion est prévu sur une durée précise. Ce plan est prévu pour se mettre en place sur toutes ces communautés, alors, plutôt que de choisir l'ordre en commençant par le plus près de Paris, puis le plus loin ensuite, choisissons l'ordre de manière aléatoire. C'est quelque chose qui est souvent facile à faire accepter, à faire comprendre. C'est donc quelque chose à garder en tête.

Par exemple, Abhijit Banerjee a parlé hier du programme de traitement des vers intestinaux au Kenya. C'était un projet qui a été conçu de cette manière : l'ONG voulait traiter 75 écoles, 25 la première année, 50 la deuxième année et 75 la troisième. Les chercheurs ont convaincu l'ONG de choisir les 75 d'emblée et au hasard l'ordre dans lequel les différents groupes arrivaient. Le problème à garder en tête avec cette méthode, c'est que l'introduction du programme doit être suffisamment lente pour laisser le temps au temps et laisser aux effets le temps d'arriver. Une fois que le groupe de contrôle a rattrapé le groupe de traitement, c'est fini, on n'a plus de groupe de contrôle. Donc, par exemple, si l'on pense à un projet de microcrédit et que l'on dit que l'on va faire un *randomize phasing* de cette manière et qu'au bout de six mois, on va passer dans les communautés contrôle, cela ne va pas marcher. Parce que cela prend beaucoup plus de six mois, pour les effets du microcrédit, de se faire sentir.

Un autre type de solution à garder en tête, ce sont les loteries, voire potentiellement les loteries publiques, dans le cas de contrainte de ressources. Il y a des programmes – souvent au début, mais parfois de manière permanente – qui n'ont pas assez de ressources pour toucher tous les gens qui voudraient en bénéficier et où l'assignation aléatoire est vue par les personnes qui implémentent le programme comme la manière équitable de le faire, indépendamment de l'évaluation. Et là, pour l'équipe d'évaluation, le seul travail, c'est d'être là au bon moment pour collecter les données. Un exemple sur lequel Bruno Crépon et Marc Gurgand travaillent est celui de la comparaison ANPE/Unédic. Il était prévu de faire bénéficier du nouveau plan de l'Unédic un certain nombre d'individus de toute façon. Donc, une fois que ce nombre est fixé, la manière dont on l'alloue, finalement, a assez peu d'importance. Le nombre de bénéficiaires potentiels est suffisamment élevé pour qu'une loterie soit aperçue comme équitable.

Un autre exemple, en Colombie, où une ville voulait donner aux étudiants d'écoles secondaires des bons pour payer l'école privée. Le budget de ce programme permettait de distribuer à peu près deux mille de ces bons et pas plus. Ils ont demandé aux individus de poser leur candidature pour ce programme. Ils ont reçu beaucoup plus que deux mille candidatures et ils ont donc fait une loterie. Ils n'avaient pas du tout l'intention de faire une évaluation, mais quelqu'un passait par là, a entendu parler de cette loterie et a dit : « Très bien, on va collecter des données ». Un autre exemple plus récent : vous avez sans doute entendu dire, si vous suivez la campagne électorale américaine, qu'un des problèmes majeurs est la question de l'assurance santé. On va peut-être être capables de faire des progrès là-dessus, puisque l'Etat de l'Oregon a décidé d'introduire une nouvelle assurance santé

complémentaire pour ses résidents. Mais ils ont suffisamment d'argent pour dix mille personnes. Ils ont demandé aux gens de poser leur candidature au programme et ont reçu cinquante mille demandes. Ils ont donc choisi dix mille personnes au hasard, parce qu'ils pensaient que c'était ainsi qu'il fallait faire. A nouveau John Gruber du MIT, qui s'intéresse beaucoup aux questions d'assurance santé, a entendu parler de cette situation et s'est immédiatement mis en quête d'une équipe pour collecter les données.

D'autres exemples : souvent, les écoles que l'on appelle *Magnet Schools*, à savoir les écoles d'élite aux Etats-Unis, ont trop de candidats et les choisissent au hasard, parmi les candidats qualifiés.

C'est souvent quelque chose à garder en tête, jusqu'à ce que cela se produise. Du point de vue de l'évaluation, il s'agit d'être là quand cela se produit.

Une autre manière de produire, est ce que l'on appelle la "randomisation dans la bulle", avec deux exemples récents : un projet que l'on a commencé avec Philippe Zamora, sur l'école de la deuxième chance à Paris et un autre terminé sur un programme de microcrédit en Afrique du Sud. Souvent, les personnes qui implémentent le programme sont issues de trois catégories. Quand on pense à des programmes individuels, comme le microcrédit, ou comme l'école de la deuxième chance, qui requalifie les enfants qui sont sortis du système scolaire sans aucune qualification, il existe trois types d'individus. D'abord, il y a ceux qui sont importants pour le programme ; la banque pense que ce sont de bons clients et veut absolument leur prêter ; ou l'école de la deuxième chance pense que ce sont des enfants très prometteurs, auxquels il faut donner cette deuxième chance. Il y a des enfants avec lesquels ils ne veulent rien avoir à voir : ils pensent que ces enfants ne vont jamais y arriver, ont trop de problèmes, etc. ou des clients de microcrédits qui ne vont jamais rembourser leur emprunt. Et puis, il existe une troisième catégorie, que l'on appelle "dans la bulle", c'est-à-dire qu'ils sont un petit peu différents. S'ils avaient plus de ressources, s'ils prêtaient plus, s'ils avaient plus de budget, on serait content de travailler avec eux, mais dans l'état actuel des choses, ils ne le font pas. Ce sont deux exemples réels. Dans les deux cas, c'est le cas. On a eu ces conversations-là.

Donc, quand on parle avec ces personnes, elles n'ont certainement pas du tout envie d'exclure les personnes de la première catégorie, pas du tout envie d'inclure les personnes de la deuxième catégorie, mais sont tout à fait disposées à choisir au hasard dans la troisième catégorie. Parce que la troisième catégorie est suffisamment élastique pour que cela n'attaque pas la nature de leur programme. Donc, en général, quand on a une première conversation avec des promoteurs d'un programme qui disent : « Ah non, on ne peut absolument pas choisir les gens au hasard, parce qu'on a ce système de sélection extrêmement précis, auquel on réfléchit vraiment très fort ». Souvent, il est possible d'aller un pas plus loin et dire : « Oui, mais votre système de sélection extrêmement précis et auquel vous réfléchissez très fort ne lèse-t-il pas un certain nombre de gens dont le cas est plus ou moins indéterminé ? » Si c'est le cas, on peut se concentrer sur ces personnes. Alors, évidemment, quand on se concentre sur ces personnes, on n'obtient pas l'effet du programme sur la population dans son ensemble, puisqu'on n'a pas randomisé sur la population dans son ensemble. Il faut garder cela en tête.

Si les effets sont hétérogènes, on obtient l'effet du programme sur la population du milieu. Or, il se trouve qu'en général c'est exactement la population à laquelle on s'intéresse, d'un point de vue politique, pas forcément d'un point de vue économique, puisqu'il s'agit de la population qui serait touchée par une expansion du programme. Par exemple, si l'on parle du microcrédit, si la question que l'on se pose est, faudrait-il étendre le microcrédit d'une manière plus large, à des gens qui, aujourd'hui, sont exclus par les banques parce que les banques sont un peu frileuses, la question que l'on se pose est exactement : il faut que j'aille regarder dans cette bulle. Si le microcrédit était étendu, ce serait à ces personnes. De la même manière, si l'école de la deuxième chance se demande s'il faut que l'on étende nos opérations pour toucher plus de jeunes, les jeunes qu'ils toucheront, mais qu'ils ne touchent pas aujourd'hui, sont ceux qui sont aujourd'hui dans la bulle, pas ceux qui sont déjà traités, pas ceux qui sont vraiment intraitables. Cela est potentiellement très prometteur.

Un autre type de solution dont on a parlé hier également, c'est l'encouragement, qui peut être utile quand on essaie d'évaluer une politique ou un programme auquel, en principe, tout le monde a droit. Par exemple, une politique nationale ou quelque chose qui existe déjà, comme par exemple l'effet d'un vaccin pour la grippe. Comment faire ? On ne va pas dire aux gens : « Non, non, vous n'avez pas le droit de vous vacciner contre la grippe, parce que l'on veut en tester l'effet ». Ni : « Non, vous n'avez pas le droit d'aller prendre un emprunt dans cette banque, parce que l'on veut en voir l'effet ». Ce que l'on peut utiliser, c'est un certain nombre de politiques ou d'actions que tout le monde ne suit pas. Ainsi, tout le monde n'obtient pas un vaccin contre la grippe. Tout chômeur ne prend pas avantage de toutes les possibilités de formation qu'il peut obtenir, même s'il y a en principe droit. Parce qu'il existe toute une panoplie de programmes et tout le monde ne fait pas tout. Le principe de l'encouragement consiste à prendre un groupe de personnes, par exemple un groupe de médecins et de leur envoyer une lettre en leur disant : « Pensez à faire vacciner vos patients contre la grippe ». Si cette lettre est efficace, les patients des médecins du groupe de traitement auront plus de chances d'être vaccinés contre la grippe. On peut donc regarder après s'ils sont plus en forme pendant l'hiver.

Autre exemple plus pratique, sur lequel on travaille avec l'Adie dans un des projets du Haut Commissariat : l'Adie est une agence de microcrédit, en France, qui a un programme de soutien et d'accompagnement – financier et formation – pour des jeunes des banlieues, pour commencer un petit business. Ils n'ont pas envie de dire : « Non, vous n'avez pas le droit d'avoir le programme, vous avez le droit d'avoir le programme ». Mais, par contre, ils sont disposés à travailler avec les missions locales, pour choisir un certain nombre de jeunes, pour leur faire une publicité particulièrement agressive du programme de l'Adie. Cela ne signifie pas que si quelqu'un entend parler de l'Adie par un autre canal, qu'il vient au bureau de l'Adie et qu'il dit : « Moi aussi, je veux participer au programme », il n'aura pas le droit de participer au programme. Mais, il y a moins de chance qu'il le fasse, parce qu'il n'aura pas reçu une propagande agressive comme le groupe de traitement.

Dans ce cas, quels sont les avantages et les désavantages de cette méthode ? En général, elle est très populaire, parce qu'elle n'implique pas de changer quoi que ce soit dans le programme lui-même. C'est quelque chose que l'on fait en amont, pour encourager le

programme. Par ailleurs, cela fait de la publicité pour le programme, ce qui est bien. Désavantage : cela ne marche pas si le mécanisme d'encouragement n'est pas extrêmement puissant. On doit aussi faire attention à ce que le mécanisme d'encouragement n'ait pas d'effet direct. J'ai parlé de l'exemple des vaccins à dessein, parce que c'est en fait un exemple où cette méthode n'a pas fonctionné, parce que les médecins, quand ils ont reçu leur lettre disant qu'il fallait vacciner leurs patients contre la grippe, ont été réveillés : « Oui, c'est l'hiver, il faut que j'encourage mes clients à prendre tout un tas de mesures préventives ». Et, en plus de les vacciner contre la grippe, ils leur ont aussi proposé de prendre la vitamine C, de manger des oranges, etc. On a pu ainsi voir chez les patients de ces médecins une augmentation de toutes les mesures préventives, pas seulement une chance plus grande d'avoir un vaccin contre la grippe. De ce fait, on estime ici l'effet de la lettre, mais pas l'effet du vaccin contre la grippe. Parce que cette lettre lance tout un tas d'actions différentes.

J'ai passé en revue de manière un peu brève et cavalière un certain nombre d'options pour introduire la randomisation en pratique. La raison pour laquelle je l'ai fait, c'est pour montrer qu'il existe un certain nombre d'options et que l'on ne doit pas forcément penser à la méthode aléatoire comme : « Il faut d'abord que j'ai tout mon échantillon, il faut que je choisisse les personnes au hasard pour participer au programme, pour ne pas y participer, cela ne va jamais marcher ». En fait, cela peut marcher si l'on est un peu créatif, si l'on a un dialogue avec les personnes qui implémentent le programme. Il y a souvent un moyen de faire les choses de manière à satisfaire tout le monde.

Je vais brièvement passer en revue quelques questions d'analyse. Je vais juste montrer les problèmes et pas vraiment les solutions.

Le problème que l'on rencontre dans l'analyse des évaluations aléatoires, est l'attrition, c'est-à-dire la perte de données. On a une *baseline*, un avant-programme, le programme se passe, une enquête et on a perdu des personnes. Cela est à éviter. Après coup, il est extrêmement difficile de résoudre ce problème. Cela est particulièrement difficile si l'attrition n'est pas la même dans le groupe de traitement et dans le groupe de contrôle. Il faut donc penser à avoir des méthodes de collecte des données similaires dans le traitement et dans le contrôle. Si l'on se dit : « Je n'ai pas de problème avec mes traitements parce que, de toute façon, ils sont suivis par le programme, je vais tous les recevoir, mais mes contrôles, il va falloir que j'aille leur téléphoner pour leur poser la question », cela va être une catastrophe. Parce que non seulement, il y aura eu de l'attrition chez les contrôles, mais une attrition différentielle des traitements et donc les échantillons ne sont plus comparables *ex post*. Deux manières de l'éviter : les données administratives liées – je parle ici à des personnes qui ont la capacité de mettre cela en place, donc gardez cela en tête – et si ce n'est pas possible, mettre en place dès le début des méthodes pour être sûr de retrouver les personnes, quoi qu'il ait pu leur arriver entre-temps, donc leur donner des minutes de téléphone, un téléphone portable si nécessaire, trouver cinquante téléphones de leurs parents, amis, voisins, etc., leurs adresses, etc.

Autre problème d'analyse dont on a déjà un peu parlé hier, c'est celui du manque de *compliance*, c'est-à-dire les personnes qui dans le groupe de traitement ne suivent pas leur traitement ou les personnes qui dans le groupe de contrôle trouvent un moyen d'être traités.

Cela n'est pas si grave. On a vu beaucoup d'exemples, hier, par exemple MTO où à peu près 50 % des personnes ont bénéficié du *voucher* ; c'est un problème qui fait partie intégrante de l'encouragement dont on a parlé : quand on envoie une lettre aux médecins pour leur proposer de faire vacciner leurs clients contre la grippe, seulement un certain nombre va décider de le faire. Cela est moins grave.

Un point essentiel à garder en tête, c'est qu'il faut toujours comparer les personnes qui ont été initialement assignées au groupe de contrôle, à celles qui ont été initialement assignées au groupe de traitement. Ce que l'on n'a pas le droit de faire du tout, c'est soit de jeter le groupe de traitement qui finalement n'a pas été traité à la poubelle, soit, pire encore, de les mettre dans le groupe de contrôle. Il faudra toujours comparer ceux que l'on avait l'intention de traiter au début, avec ceux que l'on avait l'intention de traiter à la fin. Deux choses à garder en tête : le manque de *compliance* aura un effet sur la puissance de l'expérience, à savoir sur la capacité de détecter des effets, qui vient non de l'effet du traitement, mais de l'effet sur ce que l'on appelle la forme réduite, c'est-à-dire l'intention traitée. Comme à la fin, on va devoir comparer tous les traitements avec tous les contrôles, l'effet va être dilué par le manque de *compliance*. Donc, si seuls 20 % des individus prennent le traitement, la différence entre le groupe de traitement initial et le groupe de contrôle initial sera de 20 % multiplié par l'effet du traitement. Et donc, la puissance va être la puissance pour un effet de 20 % multiplié par l'effet du traitement. Il n'y a pas de problème conceptuel, mais, du point de vue de la taille de l'échantillon, cela peut être extrêmement décevant, s'il se trouve que le *take-up* est faible.

On peut passer de cette intention de traiter, quand on compare tout le groupe de contrôle à tout le groupe de traitement, à l'effet du traitement sur les gens qui ont été de fait traités. On en a parlé un petit peu hier, avec la question de Thierry Magnac. Parfois, c'est un paramètre de politique qui nous intéresse, parfois non, donc il faut y réfléchir.

Autre question d'analyse, la question des externalités : est-ce que les individus qui sont dans le groupe de contrôle sont affectés par le fait qu'ils sont proches d'individus qui sont dans le groupe de traitement ? Quelles sont les possibilités ? On a parlé des vers intestinaux, ce qui est extrêmement contagieux. Donc, si on avait fait une randomisation au niveau individuel, les enfants du groupe de contrôle sont eux-mêmes touchés par le fait qu'il y a des enfants traités dans leur école, puisqu'il y a moins de vers dans l'école. Par ailleurs, les enfants du groupe de contrôle réinfectent les enfants du groupe de traitement. On trouvera donc des effets trop faibles. Autre possibilité : les effets de groupe, en particulier par exemple en éducation. Si l'on se dit qu'un certain nombre d'enfants bénéficient d'un programme de soutien scolaire, peut-être que si ces enfants font des progrès, ils aideront leurs camarades ; il y a à nouveau une contamination.

Autre chose dont on a parlé hier, ce sont les effets d'équilibre général. C'est une chose à laquelle on s'est confrontés avec un projet sur lequel on travaille avec entre autres Bruno Crépon et Philippe Zamora, sur un projet qui vise à aider les jeunes chômeurs à trouver un emploi. Si on aide un certain nombre de personnes, dans une localité, à trouver un emploi, il est possible que ces personnes trouvent un emploi beaucoup plus facilement que les personnes du groupe de contrôle, mais qu'ils leur prennent leur emploi. Ils sont en première ligne de la

file et il n'y a donc pas d'effet net. Si on ne prend pas en compte les externalités, on peut soit sous-estimer l'effet du traitement, par exemple dans le cas des vers, soit surestimer, quand on essaie de passer de l'intention de traiter à l'effet du traitement. Reprenons l'effet des vers, si l'on avait fait une randomisation au niveau de l'école et que l'on dise : « Seuls 20 % des enfants ont été traités, donc je vais multiplier mon effet par cinq, parce que l'effet est dû à ces 20 % », on surestime l'effet du traitement, parce qu'en fait, les enfants du groupe de contrôle, qui n'ont pas été traités, sont aussi traités.

Je vais m'arrêter là pour ne pas dépasser mon temps. Je voulais juste passer en revue les questions de *design*. Il faut garder en tête la puissance d'une expérience, ce que j'appelle la probabilité d'être déçu, c'est-à-dire qu'il y ait l'effet d'une taille donnée, mais que l'on n'arrive pas à le calculer. Cela dépend de la taille de l'effet, de la taille de l'échantillon, du niveau de *compliance* dont on a parlé mais aussi du fait que l'on ait une randomisation au niveau individuel ou au niveau du groupe. Ma recommandation est : « Ne faites pas d'expériences trop petites, qui n'ont pas assez de pouvoir, pas assez de puissance ; tout le monde sera déçu ». Et c'est vraiment une tentation de dire : « On veut vraiment le faire, donc on veut vraiment convaincre, on y va » et on n'a pas assez de personnes dans l'expérience. On fait toute l'expérience et à la fin, il n'y a pas assez de puissance et ce n'est vraiment pas bien. Cela ne vaut pas la peine. Il vaut mieux ne pas faire d'expérience que faire une expérience qui n'aura pas assez de puissance. Cela veut dire aussi que quand on fait nos calculs de puissance, on a besoin de savoir quelle est la taille de l'effet que l'on pense avoir. Et une tentation que j'ai observée chez beaucoup d'ONG, c'est de dire que l'on aura un effet formidable. On va multiplier par deux les revenus de nos jeunes après un an. Très bien. Le problème est que si on fait nos calculs de puissance en pensant que l'on aura un effet formidable, on n'a pas besoin d'un échantillon très grand, puisque l'effet est formidable, donc on va pouvoir être capables de détecter l'effet facilement. S'il se trouve que l'effet est respectable, mais pas formidable, on a fait nos calculs de puissance en pensant qu'on aurait cet effet formidable, on sera déçus. Il ne faut pas être trop optimiste au début. Il faut faire les calculs de puissance en partant d'effets raisonnables.

Dernière question, celle que l'on a posée hier : celle de la randomisation par groupes. Est-ce que l'on randomise au niveau individuel ou au niveau de la communauté ? Pourquoi randomiser plutôt au niveau de la communauté ? Soit pour minimiser ou enlever la contamination, par exemple dans le cas du *deworming* ; ou parce que ce n'est pas faisable autrement, par exemple dans le cas du programme PROGESA, il y aurait eu des émeutes dans les villages si certaines personnes étaient éligibles et d'autres non ; où parce que c'est la seule manière de procéder, par exemple si on pense à un programme de formation des enseignants, tous les enfants d'une classe sont concernés. La randomisation par groupes a un impact important : ce qui arrive à tous les individus qui appartiennent à un même groupe, par exemple tous les enfants d'une même école, leurs résultats sont corrélés, parce qu'ils partagent un enseignant, un programme, ils sont dans la même ville, etc.

Je reviens à mes questions de calcul de puissance : quand on fait nos calculs, il faut prendre cette corrélation en compte. Ce n'est pas tout à fait exact, mais presque, il faut penser que la puissance vient du nombre de groupes, pas du nombre d'individus.

Je vais vous laisser sur cette dernière phrase, si vous prenez deux départements et que vous en choisissez un au hasard, cela peut être intéressant, mais ce n'est plus une expérience aléatoire, même si cela est choisi au hasard. Parce qu'avec deux départements et un au hasard, votre puissance vient de deux groupes. N'importe quoi d'autre aurait pu être différent dans ces deux départements à part le programme. Ce n'est pas un message très sympathique, parce qu'on aime souvent se dire : on va faire deux départements, en choisir un au hasard, formidable. Cela peut être utile. Cela peut être un bon pilote. Mais on sort du domaine de l'expérience aléatoire, puisqu'on sort du domaine où on a suffisamment d'échantillons traités pour dire que, à part le programme, on est sûr que tous les gens du groupe de traitement et tous les gens du groupe de contrôle auraient été similaires.

Je vous remercie.

Yannick MOREAU
Merci beaucoup de cet exposé particulièrement clair. Il montre bien que nous avons maintenant un capital d'expérience tout à fait important, qui invite à la fois à "doucher" les enthousiasmes qui seraient trop faciles et aussi à répondre à un certain scepticisme : « Mais, est-ce que ces résultats ont une quelconque portée ? Il y a des effets aléatoires, etc. ». La capacité de les prendre en compte, de savoir quand ils faussent ou non est évidemment quelque chose de très important, notamment pour rendre compte des expériences. Parce que ces expériences sont utiles si elles sont prises au sérieux par les décideurs et le public. Je suis frappée du capital d'expérience qui se réunit progressivement et qui est probablement très supérieur à ce que nous attendons, en France, ce que spontanément nos administrations et nos décideurs politiques sont prêts à mobiliser. Cet exposé me rend personnellement tout à fait optimiste.

Avez-vous des questions ?

Stéphane JUGNOT, INSEE
J'avais deux questions. La première était sur l'aval, que vous n'avez pas trop abordé. Il me semble que dans les expérimentations, notamment en matière de programme de politique publique, un des aspects importants est la transparence, la pédagogie et l'aspect contradictoire. J'ai un contre-exemple en tête, avec ce qu'a fait l'Unédic pour mettre en place et généraliser le profil des demandeurs d'emploi, où il y a eu juste une publication de quatre pages qui donnait peu d'informations. Dans les expériences que vous connaissez, en pratique, quels types d'informations sont mis à la disposition du grand public et des spécialistes ? En particulier, y a-t-il un accès aux bases de données utilisées pour tous les chercheurs qui le souhaiteraient ?

En matière de santé, quand l'on voit que les expérimentations fonctionnent ou non, on peut les interrompre ou les généraliser. Avez-vous déjà vu faire ceci dans les expérimentations qui ne concernent pas le domaine de la santé ?

Esther DUFLO
Pour la première question, je suis extrêmement partisane de partager les données au maximum. Dans toutes les expériences dans lesquelles je suis impliquée, les données sont systématiquement mises en ligne. On met même les données de *baseline* avant que l'expérience ne soit terminée, quand elles sont prêtes. Ensuite, on met les données en ligne quand l'expérience est terminée. Je me suis rendu compte qu'elles sont téléchargées extrêmement fréquemment. Quand je les ai mises en ligne, je me demande toujours à qui cela peut servir. En fait, manifestement, cela sert. Je suis donc extrêmement favorable à la mise de données en ligne.

Certes, j'imagine qu'il peut y avoir des contraintes, dans le cas de politiques publiques en France, dont je ne suis pas au courant, donc je ne m'étendrai pas dessus.

Du point de vue de la transparence, cela dépend un peu des expériences. Avant, je travaillais surtout avec de petites organisations qui font leurs expériences dans leur coin. Les personnes impliquées dans l'étude sont toutes au courant, mais, en général, personne d'autre, parce que cela n'intéresse personne. Mais dès que l'on met en place quelque chose de plus important, la communication est extrêmement importante.

En général, quand on interrompt des essais cliniques, on a déjà un certain nombre de résultats. Ainsi, par exemple, on avait pensé faire un suivi pendant deux ans et l'on se rend compte, au bout d'un an, d'effets formidables ou du contraire. Dans les projets pour lesquels j'étais impliquée, je n'ai jamais vu qu'il y ait de résultats intermédiaires qui permettent d'arrêter ou de généraliser. Par contre, une fois que les expériences sont terminées, il m'est souvent arrivé de voir que l'on va dans une direction ou dans l'autre, en fonction des résultats obtenus. Par exemple, sur le projet de *deworming*, au début, l'ONG pensait atteindre 75 écoles dans l'expérience. Au bout de deux ans, les résultats étaient extrêmement positifs. Alors, au lieu de faire 75 écoles, ils en ont fait beaucoup plus. Et inversement, sur des projets qui n'ont pas fonctionné.

Yannick MOREAU
Autre question ? A ce stade, vous avez répondu à tout. Cela rebondira peut-être un peu plus tard. Non, une autre question, là.

Nicole ROTH, DREES
Ma question porte sur l'aspect généralisation. Quand on fait des expériences, on le fait sur de petites tailles, de petits échantillons, des zones témoins, expérimentales… Il peut y avoir une question sur la généralisation, une fois que l'on passe à l'ensemble du territoire, notamment en matière sociale, d'avoir des effets qui ne seraient pas les mêmes à cause d'un effet de taille et justement, d'un effet du général, par rapport à un effet plus particulier. Avez-vous des recommandations sur ce point ?

Yannick MOREAU
Je me permets de compléter la question. N'y a-t-il pas un effet de motivation qui, notamment dans des expériences sociales, fait que pour une expérience, tout le monde est motivé, mais

après, quand on va généraliser et quand on va être dans la durée, est-ce que cela va continuer ?

Esther DUFLO
C'est une très bonne question et une très bonne extension à la question. C'est peut-être le problème le plus difficile auquel nous sommes confrontés. Je n'ai pas une réponse magique et qui résoudrait tous les problèmes, mais plutôt un certain nombre de petites remarques à ce sujet.

La première, c'est que l'effet de motivation peut être là et en particulier quand on veut évaluer un programme, en général, on veut faire en sorte que le programme ait au moins lieu. On essaie de ne pas évaluer ce que l'on appelle les programmes "plaqué or". Cela ne sert à rien si cela ne peut être généralisable. On essaie d'avoir des protocoles dans ce que l'on évalue, qui soient simples et réplicables. Si ce n'est pas le cas, ce n'est pas très utile. En revanche, ce sont des programmes qui, en général, ont lieu. On les vérifie, il y a une évaluation de processus qui suit le programme, pour être sûr que, de fait, quelque chose se produise. Sinon, il n'y a pas vraiment de leçon à tirer d'un programme qui n'a pas eu lieu. Après, au moment de la généralisation, une question va se poser : est-ce que cela aura lieu aussi bien ? Dans une évaluation, on évalue plus l'aspect *proof of concept*, à savoir quelque chose qui, s'il est bien fait, induira l'effet que l'on peut en attendre. Si cela est bien fait, à grande échelle, cela peut être un grand SI.

Mais, il y a d'autres problèmes de généralisation. Une population donnée peut ne pas être représentative d'une population plus générale. Dans ce cas, que peut-on dire ? D'une part, on a touché hier, au sujet de l'exposé d'Abhijit Banerjee, le fait que les évaluations ne sont pas un substitut à une théorie. Il y a toujours une théorie qui sous-tend la décision de mettre en place ce programme plutôt que tel autre. Et c'est cette théorie qui nous guide sur la possibilité que le programme s'étende. Par exemple, je prends l'école de la deuxième chance. Je sais que si j'étends ce programme à des gens qui sortent de maîtrise, il n'aura pas d'effet. Car, la théorie c'est : l'école de la deuxième chance, cela vise des personnes qui sont sorties du système avec telle ou telle qualification, donc peut-être que cela s'étendra à ces personnes. La théorie nous guide sur l'extension que je peux raisonnablement attendre. Par ailleurs, elle peut aussi nous guider sur le fait de se demander s'il serait nécessaire de refaire une expérience dans un autre endroit. Nous essayons donc de faire les mêmes expériences en même temps dans différents sites, ce qui permet de toucher potentiellement la population concernée si l'on étendait le programme.

Troisième possibilité : des effets d'équilibre général. Par exemple, si je fais un tout petit programme d'accompagnement des chômeurs, cela va aider mes cent chômeurs ; si je mettais en place pour tout le monde, dans tout le marché du travail de la région, cela ne fonctionnerait pas. J'en ai un peu parlé au niveau des externalités. Si l'on pense que ce sont des effets importants, il faut mettre en place l'expérience de manière à pouvoir répondre à ces questions.

C'est pour cela que dans le projet d'accompagnement des jeunes, on a eu une randomisation à deux niveaux : au niveau du marché du travail, c'est-à-dire de la mission locale, et au niveau

individuel. Il y a des missions locales où tout le monde est traité ; d'autres où personne n'est traité ; d'autres où une partie des gens sont traités. Cela permet de dire : est-ce que l'on va avoir un effet dans les endroits où un certain nombre de personnes sont traitées quand on les compare aux autres qui ne le sont pas dans les mêmes missions locales ? Mais si l'on n'a pas d'effet là où tout le monde est traité *vs* personne n'est traité, ceci nous dira : c'est un effet local, mais qui ne se généralisera pas si cela devient une politique nationale.

Ce sont des expériences qui sont plus difficiles à mettre en œuvre, puisqu'il faut le faire au niveau d'un marché. Mais il est extrêmement important, sur de telles questions, de garder ces effets en tête. D'un autre côté, s'il n'y a pas d'effet local, il n'y aura probablement pas d'effet d'équilibre général non plus. Donc, un résultat négatif est quand même utilisable. C'est un résultat positif qui a des problèmes d'interprétation, dans ce cas-là.

Yannick MOREAU
On va peut-être s'arrêter là, parce qu'il est important que chaque intervenant ait le temps de présenter son intervention. J'imagine qu'un président de séance doit normalement dire que nous reprenons impérativement dans dix minutes. Dire moins de dix minutes, cela n'a pas de sens, donc impérativement dans dix minutes. Merci beaucoup.

Retour sommaire

Histoire et cadre juridique des expérimentations

Yannick MOREAU

Nous allons écouter les deux exposés, puis vous pourrez poser les questions aux deux intervenantes. Je vais donner la parole à Madame Calvès, de l'université de Cergy-Pontoise, pour nous parler de l'histoire et du cadre juridique des expérimentations, sujets qui ont bien entendu leur importance.

« Histoire et perspectives du cadre juridique des expérimentations législatives et réglementaires en France »

Gwénaële CALVES, Université de Cergy-Pontoise

L'expérimentation dont je vais parler – l'expérimentation législative et réglementaire, dite parfois « expérimentation normative » - ne s'apparente que d'assez loin à la démarche expérimentale dont il a été question jusqu'ici.

En effet, de manière très prosaïque, l'expérimentation normative consiste à tester, pendant une période donnée, une nouvelle règle de droit : le législateur ou le pouvoir réglementaire statuent à titre provisoire, *sous bénéfice d'inventaire*. Au vu des résultats produits, la règle sera pérennisée dans les mêmes termes, ou modifiée, ou abandonnée. En ce sens, de nombreux auteurs qualifient d' « expérimentale » la loi Veil de 1975, dont vous vous souvenez qu'elle suspendait pour une durée de cinq ans l'application des dispositions du Code pénal applicables aux femmes qui recourraient à l'avortement. La loi était ainsi dotée d'une « clause de rendez-vous », par laquelle le législateur s'imposait à lui-même de réexaminer la question à l'expiration du délai de cinq ans. En l'occurrence, vous le savez, les dispositions suspendues à titre provisoire ont été définitivement abrogées.

Pour caractériser l'expérimentation normative, il me semble toutefois qu'on ne peut pas s'en tenir au critère de la limitation dans le temps. Il faut ajouter que la « mise à l'essai » de la norme législative ou réglementaire s'opère *sur une échelle réduite*. Le champ d'application de la norme testée est circonscrit, soit géographiquement (à un département, par exemple), soit matériellement (elle n'est applicable qu'à une catégorie de personnes). Pour apprécier la pertinence d'une mesure de déconcentration administrative, on a par exemple décidé, en 1962, de ne l'appliquer qu'à cinq départements et deux régions, avant de généraliser la réforme en 1964. Quant à la limitation matérielle du champ d'application d'une norme nouvelle, on peut citer l'expérimentation qui se déroule en ce moment même en matière de notation des fonctionnaires : il est prévu que l'entretien individuel ne sera pratiqué que dans certains corps ou certains ministères, afin de comparer, au terme de l'expérimentation, les résultats obtenus dans le « groupe de traitement » d'une part, dans « le groupe de contrôle » d'autre part.

Vous le voyez, l'expérimentation normative relève simplement d'une approche pragmatique de l'art de légiférer (ou de l'art de gouverner). Il s'agit, comme le rappelait Yannick Moreau dans son propos liminaire, d'éviter de légiférer inutilement, voire à mauvais escient.

Une telle démarche peut sembler de simple bon sens. Mais il faut comprendre – et ce sera le cœur de mon propos – qu'elle soulève des problèmes juridiques assez considérables. Un commentateur pourtant favorable à la démarche expérimentale a même pu dire, récemment encore, qu'il existe « une confrontation systématique entre les principes de l'expérimentation et l'ordre juridique [français] » [1].

Le problème est d'abord d'ordre culturel. La démarche expérimentale, même entendue au sens lâche que j'ai indiqué, ne va pas de soi dans un pays comme la France, où la loi, même si elle a perdu son ancienne majesté, est toujours plus ou moins perçue comme devant revêtir un caractère de permanence, de généralité et d'incontestabilité. Cette vision classique de la loi est certes démentie par la pratique législative contemporaine, qui donne plutôt à voir une loi jetable, bavarde, mal ficelée, contestée de toutes parts. Il n'en reste pas moins que la figure d'un législateur qui tâtonne, qui avance pas à pas, qui se reconnaît explicitement le droit à l'erreur, s'intègre mal dans notre culture juridique. Cette figure, en tout cas, est plus difficile à acclimater en France que dans d'autres pays qui n'ont pas entretenu avec la même ferveur le mythe de la loi.

Au-delà de ce problème culturel, l'expérimentation normative se heurte à de réelles difficultés d'intégration dans notre droit positif. La principale tension (ou contradiction) tient au fait que l'expérimentation contrevient nécessairement au principe d'égalité, puisque la norme qu'elle teste, par définition, n'est pas valable pour tous. On peut même dire que l'expérimentation *organise délibérément une inégalité devant la règle de droit*, puisqu'elle vise précisément à comparer la situation dans laquelle vont se trouver, au terme de la phase de test, le groupe qui a été soumis à la règle mise à l'essai et celui qui est resté assujetti à la norme ancienne. Juridiquement, l'expérimentation se définit donc comme une période de coexistence entre deux normes : la norme provisoire dont on cherche à évaluer les effets, et la norme antérieure, qui continue à s'appliquer là où ne s'applique pas la norme nouvelle. Or ce qui rend le problème particulièrement épineux, c'est que ces deux ensembles de normes sont applicables *à des situations par ailleurs identiques*. Si on cherche à évaluer l'intérêt de l'entretien individuel pour la notation des fonctionnaires, il faudra bien pouvoir comparer ce qui est comparable – par exemple la situation d'un secrétaire administratif dans un ministère qui entre dans le périmètre de l'expérimentation, et celle d'un secrétaire administratif en poste dans un ministère extérieur au champ de l'expérimentation. La démarche expérimentale consiste toujours à traiter différemment des situations sinon identiques, du moins similaires.

Du point de vue de l'application du principe d'égalité, il est alors manifeste que l'expérimentation crée le trouble. Or, ce principe joue un rôle absolument structurant dans notre droit. Plusieurs de ses déclinaisons me semblent directement pertinentes pour le propos de ce colloque : l'égalité dans la jouissance des droits, notamment sociaux ; l'égalité devant le service public, notamment de l'enseignement ou de l'emploi ; l'égalité, enfin, de tous les citoyens en tant qu'ils sont détenteurs d'une parcelle de la souveraineté nationale : vous avez reconnu le principe d'indivisibilité de la République, dont la conciliation avec l'expérimentation dite « locale » est très loin d'aller de soi.

---

[1] B. Faure, note sur CE 18 décembre 2002, *Conseil national des professions de l'automobile*, AJDA 2003 p.749.

C'est pour permettre une conciliation entre l'expérimentation normative et le principe d'égalité dans ses différentes composantes que la Constitution a été révisée le 28 mars 2003. Je me propose de vous présenter le cadre général issu de cette révision, avant d'examiner les conditions d'admissibilité, aujourd'hui, d'une loi ou d'un règlement expérimental.

Concernant les bases constitutionnelles de l'expérimentation normative, deux dispositions nouvelles ont été introduites dans la Constitution par la révision du 28 mars 2003. Elles gouvernent deux types d'expérimentation dont le régime juridique est assez différent : la première est faiblement encadrée, la seconde l'est très strictement.

Le premier type d'expérimentation recouvre ce qu'on appelle parfois (dans la nomenclature de Légifrance par exemple) « l'expérimentation d'État » ; l'autre, que je serai amenée à présenter un peu plus longuement, est dite « expérimentation locale » ou « expérimentation-dérogation ». L'expérimentation d'Etat est introduite à l'article 37-1 de la Constitution, sous une forme quelque peu elliptique. Le nouvel article 37-1 prévoit en effet que « *la loi ou le règlement peuvent comporter, pour un objet et une durée limités, des dispositions à caractère expérimental* ». Le problème est… qu'on le savait déjà ! Un certain nombre d'expérimentations législatives ou réglementaires ayant été menées à bien tout au long de la V^{e} République, on savait bien que l'expérimentation normative était possible, d'autant plus que les juges – juge administratif pour le règlement, juge constitutionnel pour la loi - avaient admis, dans son principe, la démarche expérimentale[2]. Il était donc acquis, avant la révision de 2003, que « la loi ou le règlement peuvent comporter, pour un objet et une durée limités, des dispositions à caractère expérimental ». Dans ces conditions, quelle est la portée de la nouvelle disposition constitutionnelle ? Force est d'admettre qu'on ne le sait pas encore précisément.

Deux hypothèses peuvent être avancées. Première hypothèse : la formule retenue se borne à entériner la jurisprudence du Conseil constitutionnel et du Conseil d'État. La révision constitutionnelle ne changerait alors rien à l'état du droit. Elle aurait été adoptée dans le seul souci de créer une symétrie formelle entre l'expérimentation d'État et l'expérimentation locale visée par le nouvel article 72 al. 4. Le garde des Sceaux, lors des débats constituants, a parfois affirmé que c'est ainsi qu'il fallait comprendre le nouvel article 37-1.

Mais le même garde des Sceaux, au cours des mêmes débats constituants, s'est également prononcé dans le sens d'une seconde hypothèse. Cette seconde hypothèse permet de lire dans l'article 37-1 non pas une codification de l'existant, mais une réelle avancée : « ce nouvel article 37-1 », avait affirmé Dominique Perben, « donne à l'État une capacité d'expérimentation plus importante que précédemment ». C'est cette seconde option que le Conseil d'État avait proposé d'adopter en toute franchise. Dans un avis non publié du 11 octobre 2002, il avait indiqué qu'il ne servait à rien d'adopter une position qui codifiait simplement l'état du droit antérieur et avait proposé, pour pouvoir aller plus loin dans la voie

[2] Le Conseil constitutionnel a clairement indiqué qu'il est loisible au législateur de « prévoir la possibilité d'expériences […] de nature à permettre d'adopter par la suite, au vu des résultats de celles-ci, des règles nouvelles appropriées » (décision n°93-322 DC du 28 juillet 1993, *Loi relative aux établissements publics à caractère scientifique, culturel et professionnel*).

de l'expérimentation normative, la formulation suivante : « La loi et le règlement peuvent comporter des dispositions à caractère expérimental, *sans que puisse y faire obstacle l'application du principe d'égalité* ». Une telle formulation, si elle avait été retenue, aurait permis un véritable retournement de situation : la porte de l'expérimentation normative se serait ouverte à deux battants. Mais le gouvernement n'a pas retenu cette proposition. Il s'est borné à indiquer, par la bouche du garde des Sceaux, que « le fait d'inscrire le principe d'expérimentation dans la Constitution modifie l'équilibre entre le principe d'égalité et principe d'expérimentation » [3].

Dans quelle mesure modifie-t-elle cet équilibre ? La question reste ouverte. J'y reviendrai lorsque je présenterai les conditions d'admissibilité de la norme expérimentale.

Le deuxième type d'expérimentation, expérimentation locale ou « expérimentation-dérogation », rendue possible par la révision constitutionnelle représente, à l'inverse, une rupture certaine avec l'état du droit antérieur. Le nouvel article 72 alinéa 4 de la Constitution prévoit en effet que « *dans les conditions prévues par la loi organique et sauf lorsque sont en cause les conditions essentielles d'exercice des libertés publiques ou d'un droit constitutionnellement garanti, les collectivités locales ou leurs groupements peuvent, lorsque selon le cas, la loi ou le règlement l'a prévu, déroger à titre expérimental et pour un objet à une durée limités, aux dispositions législatives ou réglementaires qui régissent l'exercice de leurs compétences*. »

Cet article a été adopté pour surmonter l'obstacle que représentait, pour l'expérimentation locale, le principe d'indivisibilité de la République, qui interdit de reconnaître aux collectivités locales un pouvoir normatif autonome – *a fortiori* s'il leur permet de poser des règles dérogatoires au droit national. Cette solution est constante en droit français, mais le Conseil constitutionnel l'avait réaffirmée, en 2002, pour fermer toute possibilité d'expérimentation normative par les collectivités[4]. Il s'agissait en l'occurrence de l'Assemblée de Corse, à qui la loi déférée voulait permettre de demander au gouvernement que le législateur l'autorise à déroger aux règles en vigueur pour procéder à des expérimentations dans des matières relevant du domaine de la loi. Or le Conseil a estimé que ce dessaisissement du Parlement, « fût-ce à titre expérimental, dérogatoire et limité dans le temps », était contraire au principe selon lequel la souveraineté nationale appartient au peuple, qui l'exerce par ses représentants et par la voie du référendum (art. 3 de la Constitution). L'enjeu est bien *l'unicité du pouvoir normatif*. Les principes d'indivisibilité de la République et d'unicité du titulaire de la souveraineté permettent au Parlement d'organiser une expérimentation qui se déroulera *dans* une collectivité territoriale, mais non de déléguer *à* la collectivité la faculté de décider du contenu des normes nouvelles à expérimenter.

C'est cette solution qu'a voulu modifier la révision du 28 mars 2003. Désormais, les collectivités locales peuvent expérimenter des règles nouvelles, et des règles qui dérogent à la législation ou la réglementation nationales. Elles peuvent élaborer elles-mêmes la norme, en

---

[3] JO Débats AN, séance du 21 novembre 2002, p. 5522.
[4] Décision n°2001-454 DC du 17 janvier 2002, *Loi relative à la Corse*, cons. 20 et 21.

lieu et place des autorités nationales ; la norme « locale » qu'elles élaborent peut s'affranchir du respect des règles générales. Cette nouvelle faculté reconnue aux collectivités territoriales s'avère, naturellement, très étroitement encadrée. La loi organique du 1er août 2003 érige à cet égard toute une série de garde-fous (codifiés dans le Code général des collectivités territoriales, aux articles LO 1113-1 à 1113-7).

Ces garde-fous sont de trois ordres :

Premièrement, c'est l'État qui décide de l'expérimentation. La faculté d'expérimenter localement des normes nouvelles est subordonnée à l'adoption d'un texte d'habilitation (législatif ou réglementaire, selon la matière.) Ce texte fixe l'objet et la durée de l'expérimentation (cinq ans au plus, prorogeables le cas échéant pour une durée de trois ans). Il fixe également les conditions à remplir par les collectivités candidates, étant entendu que toutes les collectivités qui remplissent les conditions fixées par le texte d'habilitation seront admises à participer à l'expérimentation, ce qui est une différence de taille avec l'expérimentation d'État, où les pouvoirs publics peuvent librement écarter certaines candidatures. Enfin, le texte d'habilitation précise explicitement les dispositions du droit national auquel les collectivités sont autorisées à déroger.

L'exemple du RSA – qui est du reste le seul dont nous disposons à ce jour – illustre parfaitement la démarche : le texte d'habilitation est l'article 142 de la loi de finances pour 2007 (ultérieurement modifié par la loi TEPA) ; il précise l'objet de l'expérimentation : « améliorer les conditions financières de retour à l'emploi et simplifier l'accès aux contrats aidés » ; il en fixe la durée : trois ans à compter de la publication du décret d'application de la loi ; il précise les dispositions du Code de l'action sociale auxquelles les départements expérimentateurs seront admis à déroger. Si vous consultez le Code de l'action sociale en ligne sur Legifrance, vous verrez que l'article L.262-11 est assorti d'une note précisant qu' « il a été dérogé aux dispositions de cet article par une délibération de conseil général de l'Eure » datée du tant, ou par « une délibération de conseil général du Loir-et-Cher » datée du tant, etc. La dérogation est explicite. Elle est portée à la connaissance des sujets de droit.

Deuxièmement, les actes dérogatoires pris par les collectivités sont soumis à un régime particulier. Ils sont en quelque sorte placés sous surveillance. Leur régime de publication est tout à fait exceptionnel, puisqu'ils sont publiés au Journal officiel. Si vous vous reportez par exemple au JO du 1er mars 2008, vous trouverez les délibérations des conseils généraux expérimentant le RSA : ces textes sont longs, détaillés, assez divers au demeurant dans leur contenu. Tous mentionnent, comme l'exige la loi organique, la durée de validité des dispositions prises. Ces actes, d'autre part, font l'objet d'un contrôle de légalité renforcée[5] : le représentant de l'État – le préfet dans le cas du RSA – peut les déférer au tribunal administratif, et ce recours a des effets suspensifs.

[5] Ce dispositif, comme l'observe F. Crouzatier-Durand, « dissipe toute ambiguïté sur la nature des actes pris dans le cadre d'une expérimentation : bien que dérogeant à la loi, ce sont des actes administratifs soumis en tant que tels au contrôle du juge administratif » («L'expérimentation locale », *RFDA*, janv.-fév . 2004, p. 25).

Troisièmement, c'est l'État qui décide des conditions de sortie de l'expérimentation. L'article LO 1113-6 du Code général des collectivités territoriales prévoit que la loi détermine trois modalités de sortie de la phase expérimentale : la prolongation ou la modification de l'expérimentation, pour une durée qui ne peut excéder trois ans ; le maintien et la généralisation des mesures prises à titre expérimental ; l'abandon de l'expérimentation.

L'objectif de l'expérimentation locale est bien la mise au point d'une norme *nationale*. Le Conseil constitutionnel, dans sa décision relative à la loi organique du 1er août 2003, a bien souligné que le nouvel article 72 al. 4 de la Constitution permet simplement au Parlement, « dans certains cas », « d'autoriser temporairement, dans un but expérimental, les collectivités territoriales à mettre en œuvre, dans leur ressort, des mesures dérogeant à des dispositions législatives *et susceptibles d'être ultérieurement généralisées* » [6].

Pour conclure cette brève présentation du nouveau cadre constitutionnel, je voudrais souligner que la distinction entre l'expérimentation locale et l'expérimentation d'État n'est pas, dans la pratique, d'une totale netteté. Deux facteurs favorisent au contraire une certaine ambiguïté. Premier facteur : l'expérimentation d'État vise souvent à expérimenter des transferts de compétence qu'on envisage de pérenniser ultérieurement. L'expérimentation, de ce point de vue, est un auxiliaire de la politique de décentralisation[7]. Avec le RSA, par exemple, on transfère aux départements la gestion de la prime de retour à l'emploi. Or la différence entre l'expérimentation d'un transfert de compétence et l'expérimentation-dérogation est finalement assez ténue, dès lors que la décentralisation de la gestion d'une matière comporte toujours le risque de créer une rupture d'égalité entre les citoyens. Deuxième facteur de confusion entre les deux procédures : elles peuvent se combiner. C'est le cas du RSA, puisque d'une part on expérimente le transfert aux départements volontaires d'une compétence nouvelle et que, d'autre part, on habilite les départements volontaires à déroger aux dispositions de code l'action sociale (en changeant, par exemple, la périodicité du versement de la prime de retour à l'emploi versée aux bénéficiaires du RMI).

La détermination de la base constitutionnelle de telle ou telle expérimentation peut donc donner lieu à des querelles de qualification dont l'enjeu n'est pas simplement doctrinal. En effet, les conditions d'admissibilité de la norme expérimentale ne sont pas les mêmes selon qu'on se situe dans un cadre d'une expérimentation d'État ou dans celui d'une expérimentation locale.

S'agissant des conditions d'admissibilité de la norme expérimentale, pour trouver grâce aux yeux du juge, elle doit satisfaire à quatre séries de conditions. Elles se ramènent toutes - sauf peut-être la dernière - à une volonté de limiter l'atteinte portée au principe d'égalité. Première condition : une durée limitée. Cette condition, qui figure expressément dans le texte constitutionnel tant à l'article 37-1 qu'à l'article 72 alinéa 4, est inhérente à la logique de la

---

[6] Décision n°2003-478 DC du 30 juillet 2003, *Loi organique relative à l'expérimentation par les collectivités territoriales.* (souligné par nous).

[7] Elle permet notamment de mesurer la capacité des collectivités à exercer les compétences qui leur sont transférées. Sur ce point, v. l'intéressant rapport d'information AN n°3199 du 28 juin 2006 relatif à la mise en application de la loi n°2004-809 du 13 août 2004 relative aux libertés et responsabilités locales.

norme expérimentale entendue comme norme « à l'essai », ou comme dérogation temporaire instaurée en vue d'une généralisation ultérieure. Elle a été imposée, dès l'origine, par les juges administratif et constitutionnel. La limitation dans le temps de l'expérimentation, tous les juges y insistent, est une condition nécessaire pour éviter qu'elle ne se transforme, selon la formule souvent citée du Conseil d'État, en « subterfuge destiné à instaurer une réglementation à géométrie variable, ayant pour effet de suspendre *sine die* l'égalité des citoyens devant la loi »[8]. Les juges vérifient d'une part que la « date butoir » de l'expérimentation est explicitement précisée ; ils s'assurent, d'autre part, que la durée de l'expérimentation n'est pas « excessive ».

C'est le souci de l'égalité qui commande la fixation explicite d'une condition de durée. Pour s'en convaincre, on peut se reporter à un arrêt du Conseil d'État relatif à une expérimentation d'ampleur assez modeste, qui consistait simplement à tester, dans certains départements, une nouvelle modalité de notification des résultats de l'épreuve pratique du permis de conduire[9]. Il s'agissait d'évaluer la pertinence d'une procédure d'annonce différée des résultats, transmis par voie postale au lieu d'être communiqués de vive voix. L'expérimentation organise ici une différence de traitement entre des candidats dont la situation est identique à tous points *sauf un* : le département dans lequel il se présentent à l'examen. Selon que ce département entre ou non dans le périmètre de l'expérimentation, les candidats devront ou non attendre quelques jours pour savoir s'ils ont réussi l'épreuve.

Or l'arrêté ministériel qui instituait l'expérimentation ne contenait aucune date butoir. Il semblait bien, dans ces conditions, devoir être jugé illégal. Mais le Conseil d'État a opiné en sens inverse. Dans un arrêt du 18 décembre 2002, il a estimé que l'administration, ici, n'était pas tenue de fixer un terme à l'expérimentation, *dans la mesure* où la différence de traitement entre les usagers ne s'analysait pas comme une rupture d'égalité (ses effets sur la situation juridique des uns et des autres n'étant pas suffisamment graves). Lu à l'envers, cet arrêt met bien l'accent sur le lien entre la présence d'une rupture d'égalité et la nécessité de limiter explicitement la durée de l'expérimentation.

Le deuxième point est plus délicat, car l'appréciation par le juge du caractère « excessif » de la durée d'une expérimentation peut faire intervenir une part de subjectivité. Le juge administratif pratique ici un contrôle minimum, c'est-à-dire qu'il se borne à censurer le caractère « manifestement excessif » de la durée fixée par l'auteur de l'expérimentation. Ce contrôle, pour autant, n'est pas toujours indolore. Il est minimum, mais il n'est pas de pure forme. On peut penser, en l'occurrence, à l'exemple bien connu de l'expérimentation lancée par l'IEP de Paris : la création, à titre expérimental, d'une «filière ZEP » pour accéder à Sciences Po a été censurée, en 2003, par la Cour administrative de Paris qui a estimé qu'une durée de cinq ans renouvelable une fois par tacite reconduction, c'était quand même bien long pour une expérimentation[10]... Une telle durée est manifestement excessive au regard de la

[8] Conseil d'État, Rapport public 1996, *Sur le principe d'égalité*, EDCE n°48, La Documentation française, p. 52.
[9] CE, 18 décembre 2002, *Conseil national des professions de l'automobile*, AJDA 2003, p. 748, note B. Faure.
[10] CAA Paris, 6 novembre 2003, *Union nationale interuniversitaire*, AJDA 2004, p. 343, note A. Legrand.

rupture d'égalité créée par l'expérimentation (entre bacheliers issus d'un lycée situé dans le périmètre de l'expérimentation et bacheliers issus d'un lycée exclu du périmètre de l'expérimentation).

Le juge constitutionnel se montre également vigilant sur ce point. S'il ne s'est jamais placé sur ce terrain pour censurer une disposition législative expérimentale, il s'est réservé la possibilité de le faire dans une décision du 6 novembre 1996 relative à l'instauration « à l'essai », pour une durée de trois ans, d'une procédure de négociation collective dérogatoire du droit commun[11].

Pour conclure sur la question de la durée de l'expérimentation, je rappellerai qu'il est évidemment question ici d'une durée *maximale*. Dans les faits, un très grand nombre d'expérimentations ont été interrompues, ou court-circuitées, ou généralisées avant terme, pour des raisons qui relèvent essentiellement de considérations politiques.

Deuxième condition : un objet limité. À l'instar de la limitation dans le temps qui doit être explicitement fixée, la limitation matérielle doit être clairement précisée. La nature et la portée de l'expérimentation doivent être clairement précisées. La loi organique relative à l'expérimentation locale formule explicitement cette exigence de précision, qui vaut aussi pour l'expérimentation d'État. Le texte de l'article 37-1 reprend, sur ce point, une condition fixée par le Conseil constitutionnel, qui avait précisé dès 1993 qu' «il incombe au législateur de définir précisément la nature et la portée de l'expérimentation et les cas dans lesquels celle-ci peut être entreprise ». Ces précisions doivent venir du législateur lui-même, qui ne peut s'en remettre sur ce point au pouvoir réglementaire[12].

L'exigence de précision vaut à tous les niveaux de la hiérarchie des normes : dans le cas par exemple de l'expérimentation d'une « voie ZEP » d'accès à Sciences Po, le Conseil constitutionnel, saisi en 2001 d'une disposition législative dont la direction de l'IEP avait obtenu le vote pour sécuriser son projet, a tenu à souligner que les modalités de l'expérimentation (choix des établissements partenaires, mode de sélection des élèves…) devraient être fixées, « sous le contrôle du juge de la légalité », en fonction de « critères objectifs » « de nature à garantir le respect de l'exigence constitutionnelle d'égal accès à l'instruction » [13].

De même, l'objet doit être matériellement limité. Cela signifie que certaines matières sont exclues du champ de l'expérimentation. Pour l'expérimentation locale, le domaine exclu est clairement précisé par le nouvel article 72 al. 4 de la Constitution : il s'agit des matières touchant aux « conditions essentielles d'exercice d'une liberté publique ou d'un droit constitutionnellement garanti ». « Conditions essentielles » : l'expression ouvre un certain espace à l'interprétation. Le principe d'égalité devant la Justice, par exemple, n'a pas interdit d'expérimenter, dans certains ressorts judiciaires, une peine de substitution comme le travail

---

[11] Décision n°96-383 DC du 6 novembre 1996, *Développement de la négociation collective*.
[12] Décision n°93-322 DC du 28 juillet 1993, *Loi relative aux établissements publics à caractère scientifique, culturel et professionnel*, cons. 12.
[13] [13] Décision n°2001-450 DC du 11 juillet 2001, *DDOS*.

d'intérêt général (réforme testée pendant quelques années avant d'être généralisée par une loi de 1983). La différence de traitement - à situation identique - entre les condamnés, ne portait pas atteinte aux « conditions essentielles d'exercice d'un droit constitutionnellement garanti ». Il en irait à l'évidence différemment pour une expérimentation (pourtant bien intéressante pour les criminologues…) qui constituerait à faire varier le *quantum* d'une peine d'un ressort judiciaire à l'autre. Entre ces deux extrêmes, il existe toute une gamme de situations dont seul le juge pourra dire si elles tombent du bon ou du mauvais côté de la prohibition constitutionnelle.

Est-ce que cette limitation matérielle vaut aussi pour l'expérimentation entreprise sur le fondement de l'article 37-1 ? C'est une question à laquelle il n'est pas possible de répondre de manière certaine. Avant la révision constitutionnelle, il est tout à fait clair que les conditions essentielles d'une liberté publique devaient être identiques sur l'ensemble du territoire national, et qu'il était donc exclu d'expérimenter en cette matière[14]. La liberté de l'enseignement, par exemple, doit s'exercer dans des conditions identiques sur l'ensemble du territoire et ne saurait « dépendre des décisions des collectivités territoriales » [15]. Lorsque le législateur ouvre une marge de liberté aux collectivités, il doit prévoir des garanties suffisantes pour éviter les ruptures d'égalité « caractérisées ». À propos par exemple de la prestation spécifique dépendance instituée en 1997, le juge constitutionnel a admis que les départements puissent assurer la gestion de cette prestation, mais uniquement parce que la loi déterminait avec précision les conditions d'éligibilité (âge, degré de dépendance, conditions de ressources), que la décision d'octroi était soumise au contrôle du juge administratif, et que le montant minimum de la prestation fixé par décret[16].

On pourrait multiplier les exemples (un des plus célèbres étant la tentative de révision de la loi Falloux par François Bayrou ministre de l'Éducation nationale). Mais le point important pour nous est de savoir si ces solutions jurisprudentielles *seraient réaffirmées aujourd'hui dans les mêmes termes*. Ségolène Royal, lors des débats constituants, avait posé la question en s'appuyant, justement, sur l'exemple de la loi Bayrou[17]. La question est restée sans réponse…

Troisième condition : le respect des « autres exigences de valeur constitutionnelle »[18]. J'ai déjà évoqué la question des droits constitutionnellement garantis, dont les conditions d'exercice ne doivent pas être mis en péril par l'expérimentation. La même solution vaut pour l'ensemble des *principes* constitutionnels. On ne pourrait pas, par exemple, conduire une expérimentation qui consisterait à suspendre, dans certaines parties du territoire ou pour certaines catégories de population, l'application du principe de laïcité, ou du principe en vertu duquel la langue de la République est le français.

Mais, c'est surtout sur le principe d'égalité que je voudrais m'arrêter, car c'est là que se concentrent les difficultés juridiques les plus aiguës. Avec l'expérimentation, c'est toute la

---

[14] A. Heymann-Doat et G. Calvès, *Libertés publiques et droits de l'homme*, 9e éd., 2008, p. 125 sq.
[15] Conseil constitutionnel, 85-185 DC du 18 janvier 1985, *Loi Chevènement.*
[16] Décision n°96-387 DC du 21 janvier 1997, *Prestation spécifique dépendance.*
[17] JO Débats AN, séance du 21 novembre 2002, p. 5517.
[18] Décision n°2004-503 DC du 12 août 2004, *Loi relative aux libertés et responsabilités locales*, cons. 14.

technique du contrôle opéré par le juge de l'égalité qui *entre en crise.* La formule de ce contrôle, mis en œuvre tant par le juge administratif que le juge constitutionnel, est la suivante : « le principe d'égalité ne s'oppose ni à ce que le législateur règle de façon différente des situations différentes, ni à ce qu'il déroge à l'égalité pour des raisons d'intérêt général, pourvu que, dans l'un et l'autre cas, la différence de traitement qui en résulte soit en rapport avec l'objet de la loi qui l'établit » [19].

Cette formule, vous le voyez, se présente sous la forme d'une alternative (étant entendu que l'intensité du contrôle varie en fonction du critère de la différenciation opérée et/ou de la matière où celle-ci intervient). Première branche de l'alternative (la plus fréquente) : la différence de traitement entend tenir compte d'une différence de situation, en application de l'adage « à situations identiques, traitement identique ; à situations différentes, traitement différent ». Le juge, ici, vérifie la réalité de la différence de situation alléguée. Il recherche ensuite si cette différence de situation est pertinente au regard de l'objet de la loi. Il s'assure, enfin, que la différence de traitement n'est pas excessive au regard de l'objectif poursuivi. Cette méthode de contrôle en trois temps s'avère évidemment totalement inopérante pour examiner une norme expérimentale qui s'applique, par définition, à des situations identiques : si l'on veut pouvoir tirer des enseignements du test réalisé « grandeur nature », il faut bien que les caractéristiques du groupe de traitement d'une part, du groupe de contrôle d'autre part, soient identiques ou, au moins, globalement similaires.

Deuxième branche de l'alternative : la différence de traitement n'est pas justifiée par une différence de situation, mais par un intérêt général suffisamment puissant, qui permet de justifier que soient traitées différemment des situations globalement similaires (sous réserve que la différence de traitement ne soit excessive). La réduction de la pollution atmosphérique, par exemple, constitue un objectif d'intérêt général qui permet de moduler les tarifs d'accès à un pont en fonction du nombre de passagers présents dans une voiture. La différence de traitement ne repose pas sur une différence de situation pertinente au regard de l'objet de la loi, puisqu'une voiture pollue autant quel que soit le nombre de passagers transportés. Mais la discrimination tarifaire sera considérée comme une dérogation justifiée au principe d'égalité (sous réserve qu'elle ne varie pas de un à cent !), en raison de l'intérêt général poursuivi : inciter au covoiturage dans la perspective de réduire la pollution atmosphérique.

En ce qui nous concerne, la question posée revient donc à savoir si la volonté d'expérimenter un nouveau dispositif (de formation, d'accès à l'emploi…) forme un intérêt général suffisamment fort pour justifier une différence de traitement, c'est-à-dire tenir en échec l'application stricte du principe d'égalité. Cette question reste pendante. Lors des débats constituants, sa solution a été expressément remise entre les mains des juges. Le garde des Sceaux a formé le vœu que le nouvel article 37-1 « modifie l'équilibre entre le principe d'égalité et le principe d'expérimentation ». Modifie l'équilibre, évidemment, au détriment du principe d'égalité…. Jusqu'à quel point ? L'avenir le dira.

---

[19] Conseil constitutionnel, 87-232 DC du 7 janvier 1988, *Mutualisation de la CNCA* (solution constante).

Quatrième condition : elle prévoit une procédure d'évaluation et de retour au droit commun. Concernant la sortie du dispositif, l'horizon de l'expérimentation, c'est le retour au droit commun. Sans revenir sur les options ouvertes par la loi organique en matière d'expérimentation locale (prolongation, généralisation ou abandon), je voudrais simplement souligner que le juge constitutionnel avait déjà tracé des perspectives du même ordre pour l'expérimentation d'État. On peut penser qu'elles sont toujours en vigueur, puisqu'elles sont en quelque sorte inhérentes à la démarche expérimentale. Le choix d'une modalité de sortie du dispositif doit être prise, normalement, au vu des résultats atteints au terme de la phase d'expérimentation.

Par ailleurs, l'existence d'une procédure d'évaluation des résultats est une condition *sine qua non* de la légalité ou de la constitutionnalité de l'expérimentation normative. Pour l'expérimentation locale, la loi organique organise cette procédure avec un grand luxe de détails. Ce sont bien sûr les pouvoirs publics nationaux qui sont chargés de la mener à bien. Pour chaque expérimentation, le gouvernement doit transmettre au Parlement un rapport d'évaluation assorti des observations des collectivités qui ont participé à l'expérimentation. Ce rapport présente les effets constatés au terme de l'expérience, « notamment en ce qui concerne le coût et la qualité du service rendu aux usagers, l'organisation des collectivités territoriales et des services de l'État, ainsi que leurs incidences financières et fiscales » (art. LO113-5).

Pour l'expérimentation d'État, le Conseil constitutionnel a précisé que le bilan de fin d'expérimentation était une obligation de valeur constitutionnelle. Le législateur doit prévoir *ab initio* « les conditions et les procédures selon lesquelles les expérimentations doivent faire l'objet d'une évaluation » [20]. À défaut, la loi sera censurée. C'est ce qui s'est produit pour une loi de 1993 qui permettait aux universités d'expérimenter de nouvelles règles de fonctionnement[21].

Rapportées à la réalité de l'expérimentation normative telle qu'elle se pratique en France depuis le début de la V^{e} République, ces exigences peuvent laisser perplexe... Plus généralement, on ne peut pas nier qu'entre le cadre juridique que j'ai présenté à grands traits – relativement contraignant même s'il est encore en construction sur certains points – et les expérimentations souvent « sauvages » organisées par l'administration ou le législateur, il existe un écart parfois béant. Il est permis de penser que cette situation, qui n'est pas très saine, n'augure pas d'un avenir radieux pour l'expérimentation normative….ou pour les principes constitutionnels adoptés en 2003 !

Yannick MOREAU

Merci beaucoup. Il ressort clairement de l'exposé que, pour le législateur et pour le juge, la différence entre évaluation et expérimentation au sens de ce colloque n'a pas encore pénétré dans les esprits. Ce qui est demandé, c'est simplement une évaluation. On voit d'ailleurs,

---

[20] Décision n°93-322 DC du 28 juillet 1993, *Loi relative aux établissements publics à caractère scientifique, culturel et professionnel*, cons. 9.
[21] *Ibid.*

semble-t-il, avec le RSA, qu'une véritable expérimentation aurait probablement été difficile. Finalement, à condition de respecter les termes de la loi, une certaine souplesse existe quand on ne touche pas des règles juridiques absolument fondamentales. Bien sûr, si on veut faire une expérimentation juridique et que l'on ne présente pas une procédure d'évaluation, cela ne sera pas accepté. Mais, c'est quand même le minimum que l'on puisse exiger, quand cela coûte de l'argent. On ne sait pas bien faire l'évaluation, mais ce n'est pas en n'en faisant aucune que l'on saura mieux en faire. Je défends donc un peu la loi – ce qui n'est pas toujours le cas – et je donne la parole à l'intervenante suivante.

Retour sommaire

« L'examen expérimental des politiques sociales aux Etats-Unis »

Judith GUERON, ex-présidente de Manpower Demonstration Research Corporation

Bonjour, Mesdames et Messieurs, et merci à la DARES pour cette invitation.

My remarks, and this conference, focus on a research method: experimentation; but in opening, I want to remind us that what we are really talking about is a particular vision about how to make government more rational and effective.

In all countries—France, the United States, et cetera—we face complex social problems, and we have options about how to address them. Decision-making is hard; choices create winners and losers; political actors have to balance different interests; but are there ways to improve the process? I and other proponents of experiments believe that a key to making smarter choices is getting reliable evidence on whether particular policies do, or do not, work. While having such evidence will not assure that it is used, without it we are often groping in the dark.

Experimentation makes sense if you view government decisions, at least in part, as emerging from a continuous process of testing and refining alternatives, rather than emerging full-blown from theory or the discovery of some eternal truth. It also only makes sense if you can get reliable evidence. The good news is that, over the past forty years, researchers—primarily in the United States—have found a way to produce such evidence; and today what I want to do is briefly share with you the story of how this occurred in one policy area: the effort to get people to move from Welfare to work. As far as I know, welfare reform is unique in having forty years of uninterrupted experiments that are widely viewed as having had an important influence on laws and actions.

This is a U.S. story, but even in the U.S., research is only one factor that influences policy. I leave it to you to judge—particularly after this presentation—whether the vision of improving policy through a process of continuous testing is relevant in France. After all, since its founding, the United States has often been referred to as "an experiment in democracy." I have never heard people refer to France as "an experiment."

Today I am going to address five questions about the history of welfare reform. How did this happen? What does it teach us about research methods? What does it tell us about what works? Did these studies affect policy? And what are some of the big lessons?

Before turning to my subject, I want to give you two warnings and a few words on methodology. The first warning is really a reminder: experiments inform the process of decision-making—they do not set the goals of policy. That seems trivial, but from many perspectives, France has social policies that are superior to those in the United States, so while we may have something to teach you about experiments, you have much to teach us about social justice. The second warning is that I am not an impartial observer. I spent thirty years helping to build and then leading a non-profit organization—MDRC—that was one of the pioneers in demonstrating the feasibility and the persuasiveness of using experiments to assess social policy.

The methodological issue relates to the crux of the evaluation challenge, and I will summarize this briefly—basically just to assure that we have a common vocabulary for my remarks. In assessing any social program, you need to address a range of questions. Was the program well implemented? Did it achieve its goals? Are costs reasonable in relation to achievements? Do the answers to these questions vary for different groups of people, in different contexts, and in different approaches? What are the lessons?

Today I, and basically this conference, have focused mainly on the second question. Ordinary people (and some unordinary ones) find it hard to understand why it is not easy to answer this question. Why can't you simply follow people after they enter a program and find out if their behavior changes? To understand this, it's critical to recognize the difference between what researchers call "outcomes" and "impacts." A program's outcomes show the status of people at a particular time. Outcomes are such things as "how many people, on leaving a particular program, get a job or move out of poverty?" Impacts are the difference between the outcomes that did occur and the outcomes that would have occurred had the people not been in the program.

The key methodological challenge in evaluating any kind of reform is getting a reliable estimate of what people would have done, on their own, without the program being tested (this is what researchers call "the counterfactual") in order to determine what the program really accomplished. The problem is that you can never see, or directly measure, the counterfactual. Politicians and administrators who promote or run programs tend to attribute all the successful outcomes to their work. They will say with pride (and how many times have I heard this?), "I have put ten thousand people into jobs. My program has moved five thousand people out of poverty." But we all know that people don't stand still. Many get, and lose, jobs all the time. The counterfactual is a moving target. So, is placing ten thousand people in jobs a number to be proud of?

How can we determine whether a reform actually causes a change? How can we avoid repeating the rooster Chantecler's false reasoning that his crowing made the sun rise? It is now accepted by many researchers in the United States that the most reliable method to do this—to see what difference a program makes—is to use random assignment: to put

individuals into two or more groups: the group(s) being tested, and a control group, which gives you a measure of the counterfactual. If the study is well designed and well implemented, as Esther said earlier, then the difference between the behavior of people in these two groups will give you an estimate of the program's impacts. Such studies are the foundation of evidence-based medicine, and they are increasingly used to assess social programs; but do not think they're that widely used in the United States either.

Language matters, so I want to make my terminology very clear. In casual conversation, people use the word *experiment* to mean "trying out something new." Today, when I use the term *experiment*, I am only referring to a random assignment study. Now, I'm not going to use my time today to promote the value of experiments, but I do want to encourage you by saying that my personal experience, with over thirty major experiments involving over three hundred thousand people (or maybe four hundred thousand), has convinced me that random assignment lives up to its reputation. With experiments, you can know something with much greater certainty, and as a result, more confidently separate fact from hype.

I am going to start my story of welfare research with a few words about three aspects of the context in the United States that are relevant to judging its applicability to France. The first is a strong desire for change. In the U.S., the word *welfare* usually refers to the program of cash assistance for poor, lone parents (primarily mothers) who are not working. The design of this program reflects the shifting priorities placed on different goals: reducing poverty, increasing work, and limiting costs. In the years I am discussing, many factors made welfare very unpopular, including the rapid increase in program costs and the increase in labor force participation by women. As women (including single mothers with young children) flooded into jobs, public support basically evaporated for a program that came out of the New Deal in 1935, but then paid one group of women to stay at home while others—in not very different circumstances—were working (often not by choice). The public clearly favored reforms that would get people to leave welfare and go to work.

The second contextual factor is that the U.S. is a highly decentralized country with multiple sources of innovation and money. Seventy-five years ago, the Fifty States—they weren't fifty at the time, but they were famously called "laboratories for policy experimentation" —and this has certainly been the case in welfare, where benefits and rules are partly determined (and have always been partly determined) at the state level. As a result, the country is very used to variation; and ambitious governors—from Ronald Reagan, when he was governor of California, to Bill Clinton in Arkansas—tried to make their reputations by being called "successful welfare reformers." It was politically very powerful. Another example is that the entrepreneurs in the story I'm going to tell were often not in Washington, but in private foundations or research organizations—a factor that turned out to be critical in sustaining experiments when governments changed.

The final contextual factor concerns the demand for proof. My story begins in the "Dark Ages" of the early 1970's, when the United States had no answers to the most basic questions about programs moving people from welfare to work. Did they have any effect, positive or negative; and if positive, what was the magnitude? What was the cost? Do impacts vary for

different groups of people and types of programs? Is there a trade-off among program goals (such as increasing work and reducing poverty)? Is the story all about variation, or can impacts be replicable in different environments? Can the answers to such questions be obtained in a way that will be widely believed? In particular, can you do experiments? Are such studies feasible at large scale in operating programs? Will high quality information—even if you have it—make any difference? Can evidence rise above politics or academic squabbles?

The push for answers for these questions came from two sources. First, as the number of social programs in the United States grew, and measured poverty did not decline dramatically, government officials and the public increasingly demanded that programs prove their effectiveness in order to get new or continued funding. Also, advocates for poor people argued that they deserved to know whether programs promising to improve their well being actually delivered anything. As a result of a sustained program of experimental studies, we now had some answers to all seven of these questions; and in describing how this happened, I am going to focus on several examples that reflect the evolution of research and policy.

My first example is the National Supported Work Demonstration that started in 1974, and was the first random assignment study of a multi-site employment program. Supported Work offered up to eighteen months of paid employment to four groups of unemployed people: long-term welfare recipients, prisoners, people leaving prison, former drug addicts, and young school drop-outs. The goal was to produce a sustained increase in employment and a reduction of behaviors ranging from criminal activities to welfare receipt. How did this experiment come about?

By the mid-1970's, the U.S. government had tested many different approaches to helping people find jobs; but the process usually ended, to some extent, in a stalemate. Because of weak research designs, and despite spending lots of time, money, and effort, at the end of the study academics would sit around a table debating methodology and whether to believe the results. In the U.S., that is the "kiss of death" for having an impact on policy. The entrepreneur behind Supported Work was not someone in government. He was a vice president at the Ford Foundation, who basically wanted to find out the potential of a program he had funded in New York City. Since the program was costly, the idea was to learn whether it would work and what it would cost at small scales, spending tens of millions of dollars before passing a national law and spending billions of dollars. To get answers, the Foundation recruited federal partners and set up MDRC, where I was Research Director originally.

In designing the study, we explicitly sought to avoid the legacy of the 1960's, where social policies were often designed on a hunch and discredited on an anecdote, without building a reliable record of what worked and did not work. We argued that without such a record, knowledge could not accumulate, and that there was a risk that the same strategies would be trotted out every few years, good ideas would be ignored, cynicism would increase, and the country would fail to make progress. Our solution was to test Supported Work as an experiment.

In doing this, we did not naively believe that research would (or should) drive policy, but we did believe that if you could improve the quality of evidence, you would have a chance at achieving several desirable results: improving the lives of low-income people, increasing public support for social programs, and getting a higher return on scarce public investments. To assure the reliability of the study, we proposed not only random assignment, but also a large sample, multiple sites, adequate follow-up, and high-quality data.

What did Supported Work tell us about research methods? Most importantly, it showed the feasibility of using an experiment to evaluate an employment program. With more than thirty years of successful experiments, it's easy to get blasé; but at the time, this was a revolutionary idea, as it may be now in France. Most people thought it would simply be impossible to persuade local program staff to use a lottery to determine who would be served, since the suspicion was that they would react like a doctor being told to deny a patient a known benefit, and thus would reject the concept as being either unethical or illegal.

Ultimately, we sold the idea by convincing people that the whole reason for the experiment was that although Supported Work sounded like a program that could not fail, we didn't really know whether it would help people, and we had money to enroll only a small number of those likely to volunteer. We argued that, under those conditions, a lottery was the fairest way to allocate scarce opportunities (something Esther mentioned this morning). We also paid close attention to meeting (and setting) prevailing ethical and legal standards, including getting the informed consent of research subjects and paying great attention to protecting the confidentiality of very sensitive data (including criminal activity and drug use). Finally, the staff promoting the experiment in the field were not researchers speaking in academic jargon, but people who spoke the language of the people they were trying to convince.

At the same time, at the national level, we began what turned out to be an ongoing dialogue about how experiments are not more costly than other alternative, high-quality longitudinal research methods, but are instead, in a sense, more cost-effective, in that they produce results that can actually be trusted. The second method's finding was the feasibility of a survey for tracking thousands of very disadvantaged people for years, collecting high-quality data, and minimizing attrition (again, as Esther mentioned this morning). Third, we were fortunate, in that we could measure impacts that were highly transparent and relevant to our audience (for example, the percentage of people working, and what they earned).

What did this tell us about what works? The project produced many lessons, and I'll mention only a few because, in some sense, they were counterintuitive. We had expected that Supported Work would have its smallest impact for welfare recipients, since poor women have a harder time finding jobs than men, and when they do, they often have lower work incentives because they are paid lower wages, and they had welfare as an alternative to fall back on. Instead, we found long-term positive impacts for welfare mothers, and not for the other three target groups. Thus, Supported Work provided the strongest evidence to date that an employment program can work, but also a caution: that good ideas, which seem like obvious winners, may not pan out in practice, and, in some cases, can actually do harm.

We also learned a critical lesson to all the subsequent studies that may, in part, explain the discussion we had yesterday about differences between impacts for men and women. People with high outcomes can have low impacts. Thus, the men in Supported Work were more likely than the women were to find jobs. They had higher outcomes. But the men in the control group got jobs just as frequently, which was not the case for the women.

These findings led to what have become major themes in all of the subsequent studies. "Keep your eye on the control group." Programs can have unintended consequences. The story is often in the subgroups. It's difficult to explain why programs succeed, and without random assignment—that is, if we'd only looked at outcomes—we would have reached the wrong conclusion.

Did Supported Work affect policy or practice? Supported Work showed the multiple ways that you can affect policy. The negative—or, more accurately—the "null findings" for three of the groups had an immediate effect. The federal government avoided spending large sums of money on ineffective employment programs. This is what the planners had hoped when they espoused testing the program before passing a law.

In contrast, the positive findings for welfare recipients did not lead to an immediate expansion. In my view this is, at least in part, because it was a "stealth project" designed in Washington and New York. While the low profile was useful in keeping the very controversial random assignment process "below the radar screen," it meant that the state and local political actors, who play a key role in the United States, didn't have a stake in the results. But even that was not the end of the story. Because the findings were from an experiment, they were used in subsequent syntheses of research that did have a major impact on policy. In other words, knowledge did cumulate.

Most of the lessons for other fields that I drew at the time included that—it seemed to me—it was related to the value of other experiments. It seemed to me that our success in implementing that experiment was tied to our control of the funding and design of the new program. We are not trying to convince already-funded programs to tack on random assignment, which is always a very tough sell, but instead consist on it as a condition of getting substantial funding. As a result, we could require some level of standardization, and assure a large difference in treatment between people in the program and the control group. A final lesson came from the re-analysis of the Supported Work data by a number of researchers, showing that alternative, non-experimental research designs would have yielded incorrect conclusions.

My second example begins with the election of President Reagan in 1980. This marked a major turning point in the story, because it led to dramatic changes in three areas. Welfare policy became much more conservative, the states were given much greater freedom to test mandatory programs requiring people to go to work, and the federal government basically stopped funding most policy research. As a result, the prospects for experiments looked very bleak. The surprising outcome is that despite this, experiments not only flowered, but within five years shaped a new decentralized paradigm that flourished in the next fifteen years and

had a greater impact on policy than the experiments carefully nurtured in the more controlled conditions of the 1970's. How did this happen?

When the federal government decided not to study what states did in response to the 1981 law, MDRC got a grant from the Ford Foundation to address the three most important questions. Would states run tough mandatory programs? Would they reduce welfare, increase work, and what would be the impact on poverty? And would such programs cost, or save, money?

Because requiring lone parents to work was highly controversial at that time in the United States, we knew we needed rigorous evidence to defend any findings—positive or negative—and thus, proposed random assignment. Because we anticipated modest impacts, and had to assess each state as a separate experiment (because each state ran a different program), we needed very large samples—ultimately involving over thirty-five thousand people and, if you count the succeeding experiments in those states, you usually get up to sixty-five or seventy-five thousand people. As a result, we couldn't afford to rely primarily on surveys, but tracked people using existing state administrative records.

In this field, experiments had never been done before at this scale, in operating welfare offices independent of the central government in Washington, and without offering states any special operating funds. We sought to recruit states that met a number of conditions. They had to be planning a large new program, they had to have usable data, and collectively the group of states had to be representative of the national response to the new flexibility of the law passed in 1981, and of the conditions we thought were likely to affect program impacts. Further, state governors had to be willing to accept two risks. First, the potential for backlash, or negative publicity, from introducing random assignment into the high-stakes, regular welfare intake process, in offices throughout the city of Chicago, or the county of San Diego. And they also had to accept the possibility that we might produce negative findings for a very high-profile gubernatorial initiative. Collectively, these features defied the conclusions I mentioned a few minutes ago—those I and others had reached from Supported Work—about the importance of money and control to the successful implementation of an experiment; so one could reasonably ask, "Why did states participate in this experiment? What was in it for them?"

The answer is that we very consciously designed and marketed the study as an opportunity for states to answer the questions they cared about, to receive valuable assistance on program design, and to get visibility—a big plus—by participating in "*The* most important national study." I don't want to exaggerate, but when President Clinton was running for office, the successful findings from this study (about what he did as governor in Arkansas) made a difference. But candor requires me to say that it was a very tough sell, recruiting states. Implementing experiments in the U.S. has almost always, until very recently, been a real fight. There was enormous pressure to use a weaker, less intrusive research design. The absolute worst moments I recall are when MDRC's staff in the field were called "Nazis experimenting on individuals," or when I was accused of repeating the most infamous medical travesty in the United States: the Tuskegee Syphilis Study.

One reason marketing random assignment was hard is that, at that time, there was very little support in the universities for this type of work. Quite the contrary—there was widespread, and very vocal, opposition. Some of it was about statistics, theory, and the legitimacy of different disciplines and visions of truth; but part of it was also inter- and intra-disciplinary (and organizational) competition about who got money and had influence. The rare endorsements at that time (from scholars and prestigious panels at the National Academy and elsewhere) were absolutely vital to defending those experiments as ethical and uniquely reliable.

In terms of methods, the studies provided two key lessons. They showed that it was feasible to conduct experiments in regular Welfare offices—without disrupting the normal intake process (again, as Esther mentioned, making this very unobtrusive, as it took one minute to do the whole process)—and to use administrative data to produce reasonably accurate estimates of impacts. In terms of policy, the studies produced numerous lessons. Programs that required—mandated—lone parents to work (or to look for work) generally proved successful in increasing earnings and reducing welfare. They did not appear to hurt children in those families. However, average impacts were small-to-modest, many people remained on welfare, and there was usually no impact on poverty—which reflects, to a large extent, the way the welfare system in the U.S. is designed. Finally, in some states, every dollar invested in operating the program produced more than a dollar in budget savings. Those findings absolutely transformed the debate. Suddenly, reforms could be defined, described not just as "do-good" social programs, but as "investments," with measurable returns.

Did these studies affect policy or practice? Members of Congress, and other people writing about this period, concluded that they had an unusual impact on attitudes toward welfare and welfare recipients, on the design of state programs, and on federal legislation; and they generally attributed this to six factors. The first two—the technical strength of random assignment and the replication of similar findings under very different conditions—gave the studies unusual credibility. Basically, nobody questioned the findings. The third was timing and relevance—the findings came out in time to affect the debates in Congress and the states. Fourth, the programs operated at a scale that was very convincing. This was not "testing small," it was "testing large." Fifth, the researchers paid great attention to marketing and communication, and to sharing the good news as well as the bad. And finally, the political context in the U.S. at that time was less partisan and divisive than it is now, and modest impacts were enough to sell Congress on the value of Change.

The next ten years saw a flowering of experiments and the number of researchers involved in such studies. A key reason for this was the government's insistence on budget neutrality. During these years, state governors proposed increasingly radical ideas. They wanted flexibility—they weren't in the least bit interested in research; but for the first time, federal officials declared that state reforms could not increase federal budgets. Moreover—and this is key—they insisted that the yardstick for assessing budget neutrality would be a random assignment study. As a result, agreement to an experiment by a state had to agree as the *quid pro quo* for state flexibility.

I have time to share only a few findings from this period. The first is evidence of the risk of using outcomes to judge a program's success. While maximizing outcomes—how many people get a job, how many people get out of poverty—can be a useful tool to motivate the managers, it can also prompt them to make inefficient decisions. High outcomes may reflect a program's success, but they can also reflect a strong local labor market, or the enrollment of more motivated participants. An experiment can distinguish this, but the emphasis on outcomes alone can prompt managers to change whom they enroll rather than to improve what they do.

The second finding was that no single approach did best on all the goals people cared about. For example, as I said, requiring people to get jobs quickly, increased employment, saved money, and did not hurt children; but it did not get people out of poverty. In contrast, supplementing the earnings of welfare recipients—a policy with echoes in the *Revenu de Solidarités actives* discussion in France—increased work, reduced poverty, and had a positive effect on the school performance of children (we heard about this in the S.S.P. experiment yesterday), but this approach cost more money.

The third lesson was about methodology. A series of studies, using data from the welfare experiments, showed the failure of alternative research methods to replicate the results. It was those studies, combined with the evidence of the feasibility and persuasiveness of random assignment, which, over those forty years, made reluctant converts of many of us to a strong belief in the unique virtues of experiments.

At the beginning of my remarks, I described experiments as a means to make government policy more effective. In the United States, the welfare story is used as a model of how research can inform decisions; and in conclusion, I want to offer twelve lessons from that experience for others who seek to use experiments to improve policy. The first is to address important issues. To be successful, the study should address issues that matter, that will still be of interest when the results come out, and about which there are important unanswered questions.

Second: have a reasonable treatment. An experiment should test an approach that is supported by past research, and looks feasible operationally and politically (by which I mean that it is likely the relevant administrative systems will cooperate, people will participate enough for the program to make a difference, and the costs will not be so high as to rule out replication).

Third: design a real-world test. The program should be tested fairly—not in its early, start-up period—and, if feasible, in multiple sites, since we have learned that context matters. It would be uniquely convincing to be able to say, for example, here, that similar results emerged from studies in St. Etienne, Bordeaux, and Paris, as it was in Arkansas, San Diego, and Baltimore.

Lesson four: address the questions people care about. Does the program work? For whom? Under what conditions? Why? What does it cost, etc.?

Fifth: fight for random assignment. A high-quality random assignment study is superior in providing reliable estimates of whether a program works. Yet if France is anything like the

United States, proponents of experiments will have to overcome opposition from administrators, politicians, and possibly other researchers, who will argue that such studies are some combination of unnecessary, unethical, illegal, burdensome, or unreasonably expensive. The easy response is to accept a weaker design, but a second-best study is often not worth the paper it is written on. While it is important to acknowledge that there are situations where random assignment studies cannot be used or will not address the right questions, these are much less numerous than opponents will claim. To implement experiments successfully and get the most out of them, experience suggests being very careful to meet ethical and legal standards, remaining sensitive to local context, and combining different research methods to examine which features of a program (or its implementation) account for success or failure.

Lesson six: no single experiment is definitive. One study cannot address all questions for all time. Certainty accumulates with replication. In social policy, as in medicine, the real pay-off is when there are enough high-quality experimental results to allow for various types of syntheses in an effort to identify the trade-offs, and find out what works best for whom in what conditions. In the U.S., sustaining a long-term program of experiments required building a community of funders, researchers, advocates, and the Media (the Press), that valued and could distinguish high-quality studies.

Seven: do not define success as "working miracles," or you are likely to fail. To sustain a program of experiments, you need some good news. The welfare studies delivered this, in that they found many reforms that produced positive changes; but they also showed that the magnitude of Change was often modest. You might increase employment rates by seven or ten percentage points, increase average earnings by twenty-five, thirty, or forty percent. Welfare roles might decline by five or six percentage points. These modest impacts show progress, but they are not magic bullets, and I think they are less a caution about experiments than evidence of how hard it is to change people's behavior, of the importance of the economy (and other factors) in affecting behavior, and that the policies being tested in these experiments were often not so different from the services available to people in the Control group. For this audience, I would ask, "In France, are reforms likely to produce bigger impacts? And if not, would findings of modest success be viewed as building blocks to progress, as they are in medicine (or were in welfare, in many cases in the United States), or as evidence of failure?"

Eighth: simplify. One of the beauties of an experiment, as one of the speakers yesterday mentioned, is that anyone can understand what you did. There are no "fancy" statistics. In our experience, if an advanced degree is needed to understand the results, they are not likely to reach policy-makers in the United States. We built on this simplicity in a number of ways. We presented the results in a standardized—almost boringly simple—framework, we used the same outcome measures across studies, and we put most of the equations in appendices. We also avoided overly complex research designs (although they certainly got more complex over time), but we didn't avoid telling people that the findings were complex, or involved trade-offs, or needed to be understood in the context of other research.

Nine: actively communicate the results. Politicians and funders are impatient consumers. The welfare projects were structured to produce some findings quickly—in a year or two—which were aggressively shared with multiple audiences; but at the same time, we were very conscious to resist pressure to produce results so early that we subsequently risked reversing the conclusion. It is very hard to withdraw a conclusion.

Lesson ten: don't confuse dissemination with advocacy. The key to long-term, successful communication is trust. If you overstate your findings, or distort them to fit an agenda, people will know it, and will ultimately reject what you have to say. The researcher's role is to determine whether something works—not to prove that it works.

Lesson eleven: be honest about failure. Public officials and program operators want good news. Let's face it—they don't welcome hearing that progress depends upon identifying and discarding approaches that do not work, and that their program was one of those. To their credit, however, we have found many people able to learn and move on from disappointing results; but I will say that, if we had only found negative results, I wouldn't have been in business for thirty-eight years, and this whole endeavor would have folded.

Lesson twelve: get partners to buy in from the beginning. In the U.S., it is important to involve the major actors and interest groups from the beginning (as Jonathan Portes mentioned in the English context yesterday), so that they understand and have a stake in the research and are less likely to attack the methods or the findings later. You may also learn a lot from them, and be able to improve the study.

These twelve lessons emerged from the U.S. context of a lot of skepticism about whether social programs work; and I leave it to you, the experts on France, to judge whether they apply here. Beyond that, it's always useful to retain some humility. While it's not necessarily pleasant, researchers should remember that their work is only one ingredient in the policy process, and when the stakes are high enough (at least in the U.S. context), politics trumps research. Our job is to bring truth to power. Experiments are vital to doing that, but power lies elsewhere.

Thirty years ago, Europeans who visited me at MDRC typically had a two-part reaction when they heard about social policy experiments. The first response was "Wow!"—delivered with awe and admiration, "you're really testing social programs the way doctors test new drugs? That's fantastic!" And the second response was "Thank God *we* don't have to do that! In Europe, if we want to adopt a new policy, we can just pass a law." This conference suggests the atmosphere may be changing here. I hope it proves to be for the better. Thank you.

Yannick MOREAU
Merci beaucoup pour ces deux exposés extrêmement divers, qui peuvent susciter des questions également tout à fait diverses, mais qui, à travers des approches différentes, montrent très bien des contextes différents. Je voudrais tout de suite vous passer la parole pour les questions.

Jonathan PORTES
I think I'll talk in English, if that's okay. I would invite either or both of you (or indeed someone from the floor) to speculate on the likely policy and political impact of experiments. And - going to what you just said at the end - in the French context, where the normal practice is indeed simply to pass a law and implement on all the territory, will experiments in this context have the desired impact on policy and process? Also, I was sort of interested by a mission, in fact, that nobody has mentioned - what I regard as one of the most interesting Random Assignment experiments that I've seen in the last few years - which was the one conducted by the strategy unit of the French Prime Minister here on ethnic, gender, and geographical discrimination in hiring: a very sophisticated and interesting Random Assignment experiment, which has produced very strong and clear results about the extent of discrimination in the French labor market, but which has, as far as I know - and perhaps someone from the audience can correct me if I'm wrong - not appeared to have been translated into anything much in terms of policy action as of yet, nor to have received much political attention.

Judith GUERON
I think I'll leave that to the French to respond.

Gwénaële CALVES
Je ne vois pas de quelle étude il est question… Ah, vous parlez de la Halde ! Mais la Halde n'est pas la *Strategy Unit* du Premier Ministre, c'est une Autorité administrative indépendante. Et, surtout, elle ne conduit pas d'expérimentations. Elle collecte des réclamations qui, pour le coup n'ont vraiment aucune valeur statistique. La Halde existe depuis trois ans. Ce que mesure la hausse des réclamations qui lui parviennent, c'est essentiellement la progression de sa notoriété.

Interventions dans la salle hors micro.

Ah excusez-moi ! Vous parliez du *testing* mené par le Centre d'analyse stratégique ? Mais c'est moi qui l'ai commandé ! Oui, les résultats étaient intéressants, mais en entendant tout à l'heure Esther Duflo dire qu'il fallait des expérimentations sur une certaine échelle pour obtenir quelque chose de significatif, je me suis un peu tassée sur ma chaise… La population testée était très réduite, et le nombre de réponses valables obtenues (valables au sens du BIT) vraiment minuscule. Cette expérimentation a d'ailleurs été assez contestée au plan méthodologique, notamment dans le cadre d'une journée d'études que j'avais organisée au CAS[22].

Ce *testing* a-t-il eu des prolongements en termes de politiques publiques ? J'en doute. Je crois que les politiques publiques de lutte contre les discriminations, en France, sont très chahutées en ce moment. Elles hésitent entre plusieurs modèles. Un modèle classique, répressif, qui ne cesse de se renforcer au fil des années, et un modèle plus proactif de « promotion de la

[22] http://www.strategie.gouv.fr/article.php3?id_article=399

diversité » et de l'égalité des chances. Entre les deux, malheureusement, il existe une certaine contradiction. Ces politiques publiques ne fonctionnent pas bien, parce qu'elles ne savent pas sur quelle philosophie elles s'appuient. En tout cas, ce n'est pas le *testing*, du moins certainement pas le *testing* du CAS qui a déclenché une prise de conscience quelconque au niveau des pouvoirs publics. Il ne faut pas se leurrer…

Yannick MOREAU
Je suis moins sceptique que vous sur l'importance que cela a eu.

Gwénaële CALVES
Ce *testing*-là ? J'en serais ravie.

Yannick MOREAU
Celui-ci parmi d'autres. Je trouve qu'il y a un changement de prise en compte par l'opinion publique, de la réalité des discriminations, qui est extrêmement saisissant. Des faits qui étaient massifs mais très largement ignorés, me semblent aujourd'hui, non pas conduire à des politiques cohérentes, mais à ce que au moins une question soit très clairement posée sur la table. Alors, ce n'est certainement pas de l'expérimentation, ni de l'évaluation ; c'est plutôt du questionnement tout de même relativement efficace. Mais, pour des chercheurs, évidemment, cela ne doit pas être terrible.

D'autres questions ?

Marc GURGAND
Madame Calvès, vous avez indiqué que le RSA était en fait à ce jour la seule utilisation de la révision de la Constitution que vous avez présentée. Comme vous le savez, dans le RSA, des départements ont expérimenté le dispositif sur des sous-territoires qui représentent typiquement un cinquième de leur population. D'un point de vue juridique, est-ce qu'il était possible d'aller à un niveau plus fin, comme la commune, le quartier ? Est-ce que, par continuité, cela pourrait être des individus ? Ensuite, est-ce qu'en droit, il aurait été possible de déterminer au hasard les territoires sur lesquels aurait été mis en œuvre le dispositif ? Deux questions d'un point de vue strictement juridique et non politique ou mise en œuvre ?

Gwénaële CALVES
Sans connaître de manière suffisamment précise l'expérimentation RSA, j'ai constaté en lisant les délibérations des conseils généraux parues au *Journal officiel* que les collectivités, qui étaient libres de déterminer le périmètre de cette expérimentation, ont recouru à des formules assez diversifiées. Pour certaines, c'est le département tout entier qui est concerné. C'est le cas par exemple du département de l'Aude. Dans le cas du Rhône au contraire, le texte indique que les territoires d'expérimentation sont ceux des commissions locales d'insertion de Tarare, de Givors et de Villefranche-sur-Saône. Je ne suis même pas sûre que ces zones se superposent exactement au territoire communal. Conformément à la démarche de l'expérimentation locale, les territoires dérogatoires sont déterminés librement, sous le contrôle que j'ai présenté, par les expérimentateurs. Il n'y a pas de frein juridique, à ma connaissance.

Yannick MOREAU
Ces textes sont passés au Conseil d'Etat et au Conseil constitutionnel avant d'être promulgués. Mon sentiment est qu'il n'y a pas le souci de freiner les expérimentations, comme si c'était quelque chose d'horriblement dangereux, etc.

Le problème de tirage au sort est plus un problème politique qu'un problème de droit. En France, d'une manière générale, on n'aime pas trop cette procédure. Mais le juriste est probablement capable de dépasser cela, avec l'idée que sur des choses aussi lourdes que ce type de politique, l'expérimentation, c'est vraiment mieux que pas d'expérimentation. Donc, si cela est expliqué, les questions posées sur les modalités de l'évaluation ne sont pas extraordinairement exigeantes. Il ne faudrait pas tirer de l'existence d'un cadre juridique, l'idée d'un encadrement extrêmement rigide sur l'expérimentation. Mais, visiblement, vous connaissez mieux le droit que moi. Je vous en donne juste l'ambiance juridique. C'est un petit peu différent.

Gwénaële CALVES
L'adjectif clé est l'adjectif « objectif ». Il faut que le périmètre de l'expérimentation soit déterminé en fonction de critères objectifs. Opter pour l'ordre alphabétique - ceux dont le nom commence par les lettres A à L relèveront du périmètre et les autres en seront exclus - est-ce opter pour un critère « objectif » ? Ça dépend, c'est à voir….

Yannick MOREAU
C'est une forme d'objectivité. C'est aléatoire.

Gwénaële CALVES
L'essentiel est d'éliminer l'arbitraire, lors de l'application et la mise en œuvre de l'expérimentation.

Un intervenant
Existe-t-il une obligation d'informer les individus qui font partie de l'expérimentation, notamment aux Etats-Unis, qu'il s'agisse du groupe de traitement ou du groupe de contrôle ? Est-ce que si c'est le cas, cela pose des problèmes méthodologiques en matière sociale ? Est-ce qu'en France, il existe des perspectives juridiques sur ce point ? Où en est l'état de la loi, concernant les individus ?

Judith GUERON
Aux Etats-Unis, la réponse est toujours : « Ca dépend ». Alors, dans certains cas, quand les programmes qui ont été testés étaient volontaires, les individus étaient toujours informés et ont eu le choix d'entrer ou pas dans l'expérimentation. Mais, quand le programme était obligatoire (*mandatory*) et que l'Etat – de Californie notamment – a décidé de changer la loi et où tout le monde a été obligé d'être dans le nouveau système, les gens n'ont pas été informés. On "excusait" simplement certaines personnes de cette obligation. On a fait cela par *random assignment.* Les personnes étaient informées, toujours, sur la question de donner des informations dans un *survey*, à ce moment-là c'était volontaire, mais pas sur le fait de faire partie de l'expérience. Et cela ne va pas sans poser des questions aux Etats-Unis. Les gens

préfèreraient plutôt être informés et donner leur permission, mais il existe certaines questions auxquelles il n'est pas possible de répondre si le groupe choisit de manière volontaire.

Esther DUFLO
Aux Etats-Unis, toute expérience aléatoire, ou recherche impliquant des sujets humains, est encadrée par des comités d'éthique – chaque université doit avoir un comité d'éthique – qui appliquent les règles d'éthique qui ont été développées au niveau international – les principes de Belmont, qui sont les mêmes que ceux qui encadrent les essais cliniques médicaux ou toute recherche sur des sujets humains –, chaque expérience est donc passée en revue par un comité qui vérifie l'application du processus de Belmont dans le cadre de l'expérience. La réponse « Ca dépend » se trouve donc modulée en fonction de chaque expérience. Le principe de base est que les sujets doivent être informés. Mais il peut y avoir dérogation si l'on considère que le risque pour le sujet est minimum et qu'informer nuirait au protocole et réduirait l'expérience à néant. Alors, un comité d'éthique juge au cas pas cas l'adéquation du protocole au principe général de protection des sujets humains.

Judith GUERON
Ce n'est pas exactement le cas. Si cette étude est financée par un Etat qui veut une étude sur son propre programme, il a le droit d'administrer ce programme, de faire de la recherche, de payer pour ces recherches. Si c'est l'Etat même qui veut faire cette expérience et veut apprendre, le fait de devoir forcément passer par ces comités n'était ici pas du tout clair, du moins dans les années 80, et n'était pas toujours fait.

Esther DUFLO
Si cela implique un chercheur qui veut potentiellement utiliser ses travaux pour les publier, cela est indispensable. Ainsi, je n'ai pas le droit de participer à une expérience qui n'est pas passée en revue par un comité d'éthique. Alors, peut-être que s'il s'agit d'une expérience conduite par un Etat qui, ensuite, est sous-contractée à MDRC qui considère qu'il ne s'agit pas de recherche, mais de *consulting*, cela est peut-être possible. Mais, en principe, il y a au moins toujours la possibilité de faire passer par un comité d'éthique qui juge ces questions au cas par cas.

Yannick MOREAU
D'après ce que vous dites, il semble y avoir une claire différence avec la France. En France, les comités d'éthique pour la recherche médicale se sont beaucoup développés. Pour la recherche sociale, ce n'est pas une pratique courante. Probablement parce que nous n'en faisons pas assez. Quand il y a des problèmes d'éthique, ils sont traités politiquement. Vous ne pouvez pas collecter telle statistique, vous pouvez pour telle autre. Ces questions d'éthique sont intéressantes, mais elles sont à peine émergées en France, me semble-t-il.

Marie-Hélène CARLAT, direction départementale du Travail d'Indre et Loire
Madame Gueron, aux Etats-Unis, les lobbies ont une forte influence sur les politiques publiques. Peut-on dire qu'il existe un lobby de l'expérimentation ou est-ce que, finalement,

l'expérimentation réalise un consensus tel qu'il peut traverser les clivages droite-gauche ou démocrates-républicains ?

Judith GUERON
Ce qui fait la beauté de ces expérimentations, c'est que les gens arrêtent de parler, qu'ils acceptent les résultats, changent de sujet et parlent de l'application et de ce qu'il faut faire avec les résultats. Et comme les personnes de droite comme celles de gauche, en Amérique, ont des valeurs différentes – les personnes de gauche veulent plutôt baisser la pauvreté, les personnes de droite que les gens travaillent –, alors ils regardent les mêmes résultats, et ils leur trouvent des conclusions différentes. Mais, ce n'est pas une divergence sur les faits, ce sont les politiques qui diffèrent.

Yannick MOREAU
Merci à tous, et notamment à nos deux très intéressantes intervenantes.

Retour sommaire

Pierre CAHUC, CREST
Cette séance de l'après-midi est consacrée, comme ce matin, aux aspects opérationnels des expérimentations. Nous allons commencer par un exposé de Bruno Crépon, qui est au CREST. Tu vas nous présenter des expérimentations d'accompagnement menées en France auprès des demandeurs d'emploi, travaux que tu poursuis depuis plusieurs années.

Les expérimentations d'accompagnement menées en France auprès des demandeurs d'emploi

Bruno CREPON, CREST

Bonjour. Je suis particulièrement content de venir vous présenter les résultats des programmes d'accompagnement que l'on cherche à évaluer en France. Je tiens particulièrement aussi à remercier la DARES pour avoir organisé ce colloque. Il me semble bien que si la DARES organise un colloque de cette nature aujourd'hui en France, c'est peut-être parce qu'on est en train de sortir du sombre âge des ténèbres dont parlait Judith Gueron ce matin. Je les remercie, parce que ce sont vraiment des manifestations importantes.

En effet, voilà déjà plusieurs années qu'on essaie, en particulier avec Marc Gurgand, de mener des évaluations par échantillonnage aléatoire. Aujourd'hui, après plusieurs tentatives infructueuses, on est en train d'évaluer trois programmes. D'abord, le programme d'accompagnement des RMIstes dans les Hauts-de-Seine, qui a été mis en œuvre par le Conseil général. C'est le seul programme pour lequel on ait déjà des résultats. Un deuxième programme, de grande ampleur, est mené conjointement avec l'ANPE et l'Unédic. Nous cherchons à l'évaluer depuis déjà un an et demi. Un troisième programme porte sur l'accompagnement des jeunes diplômés qui sont au chômage.

L'accompagnement des RMIstes, c'est un programme qui a débuté en octobre 2005. Il concerne les RMIstes de plus de deux ans d'ancienneté dans les Hauts-de-Seine. C'est un programme qui a été confié à un opérateur privé. Il fournit une prestation d'accompagnement à la recherche durant six mois et éventuellement ensuite, une fois que les RMIstes ont repris un emploi. Le contrat de l'accompagnateur est formé d'une rémunération à la prise en charge, au placement et à la confirmation du retour à l'emploi. La population initiale était une population de quatorze mille RMIstes en stock plus le flux entrant. Au total, l'évaluation a porté sur dix-sept mille RMIstes.

Le deuxième programme est le programme des opérateurs privés et « Cap vers l'entreprise ». Ce sont des programmes d'accompagnement pour les chômeurs présentant un risque de chômage de longue durée (orientés vers le Parcours 3). L'idée générale de ces accompagnements consiste à réduire la taille des portefeuilles des agents de l'ANPE ou des personnes qui accompagnent les demandeurs d'emploi, depuis quelque chose comme cent vingt demandeurs d'emploi par agent jusqu'à près de soixante. C'est donc une réduction considérable, qui consiste en une concentration assez importante des moyens, sur des demandeurs d'emploi jugés comme présentant des risques de chômage de longue durée.

Ces deux programmes, opérateurs privés et « Cap vers l'entreprise », sont très proches l'un de l'autre dans leur esprit. Celui qui est proposé par l'Unédic a été confié à des opérateurs privés. Le public qu'il concerne est celui des demandeurs d'emploi issus du flux indemnisable. Ce sont ceux qui ont au moins cent vingt jours d'indemnisation devant eux lorsqu'ils s'inscrivent au chômage. L'objectif de l'Unedic est d'accompagner quarante mille demandeurs d'emploi.

Le programme proposé par l'ANPE, de niveau assez similaire, s'appelle « Cap vers l'entreprise ». Son public est constitué de tous les demandeurs d'emploi qui relèvent du Parcours 3, c'est-à-dire qui présentent un risque de chômage de longue durée. Il comprend donc ceux qui sont du flux indemnisable, mais aussi d'autres populations issues du flux non indemnisable, à savoir les demandeurs d'emploi qui, lors de leur inscription au chômage, ont à leur disposition moins de cent vingt jours d'indemnisation devant eux. Il y a également les demandeurs d'emploi qui sont "en stock", donc pas dans le flux indemnisable, mais ceux qui sont déjà au chômage depuis quelque temps. L'objectif est similaire. Il s'agit d'accompagner quarante mille demandeurs d'emploi également.

Enfin, le troisième programme que nous sommes en train d'évaluer est un programme d'accompagnement des jeunes diplômés. Il s'agit d'accompagner dix mille jeunes et de les réinsérer dans l'emploi. C'est un programme qui a commencé en 2007 et qui finira bientôt. Il s'agit des individus qui ont moins de trente ans, un diplôme du premier cycle et qui sont au chômage depuis plus de six mois. Le public éligible était de quarante mille demandeurs d'emploi. On veut en accompagner dix mille.

Qu'est-ce qui va nous intéresser dans ces évaluations ? Ce n'est pas ce que l'on entend en général par l'évaluation en France, qui consiste à savoir si les services que l'on avait prévu d'offrir à une population ont effectivement été offerts. Cela consiste aussi en général à compter le nombre d'individus qui vont rentrer dans les dispositifs, à savoir si tout ce que l'on avait prévu, aussi bien les objectifs quantitatifs que qualitatifs, s'est bien passé exactement comme on avait prévu de le faire.

Ce qui nous intéresse, c'est de mesurer l'impact du programme sur les bénéficiaires. Il s'agit donc plutôt de mesurer la plus-value apportée par le programme. Depuis déjà un jour et demi, nous avons eu de nombreuses précisions sur ce que pouvait être cette plus-value. Il s'agit de comparer la situation des individus avec ce qu'aurait été la situation des individus s'ils n'étaient pas passés par le programme. Ainsi, il peut s'agir de choses telles que l'accroissement du retour à l'emploi des bénéficiaires, leur maintien dans l'emploi durable. Est-ce que la participation au programme a réduit le nombre de jours passés hors emploi ? Est-ce que cela a aussi réduit les transferts dont ont bénéficié les bénéficiaires du programme : les indemnités chômage, le RMI ? En d'autres termes, ce qui nous intéresse, c'est l'impact du programme sur ces bénéficiaires.

Mais, ce qui va nous intéresser aussi – dans le dernier programme, c'est ce que l'on cherche à mesurer – c'est l'impact du programme sur les non-bénéficiaires. Peut-être qu'un programme a des effets induits sur les individus qui ne bénéficient pas du programme. Ce sont ces impacts que l'on cherche à mesurer également.

Comment est-ce que l'on mène ces évaluations d'impact ? L'idée générale des évaluations d'impact consiste à comparer la situation des bénéficiaires avec la situation des non-bénéficiaires. C'est quelque chose que l'on comprend assez facilement mais cela présente une difficulté méthodologique importante, dans la mesure où la population des bénéficiaires est très particulière. Les gens qui décident de bénéficier d'un accompagnement sont des gens qui ont des caractéristiques très particulières et l'on ne peut les comparer à n'importe qui, parce que, à ce moment-là, si l'on fait cette comparaison, l'on va mesurer deux choses à la fois, qui seront totalement indiscernables l'une de l'autre. La première, c'est l'effet du programme qui nous intéresse. La deuxième, c'est l'ensemble de toutes les caractéristiques, les spécificités de la population des bénéficiaires, qui font qu'ils ne sont pas pareils que les autres et que cela peut éventuellement affecter directement leur capacité de reprendre un emploi.

Pour pouvoir faire une comparaison correcte, il faut un dispositif spécifique. C'est tout ce qu'a montré et détaillé ce matin Esther Duflo. Il est nécessaire de comparer deux populations semblables. Pour ce faire, le tirage aléatoire s'avère extrêmement important. Ce genre d'évaluation, à tirage aléatoire, dans laquelle on cherche à comparer deux populations statistiquement importantes et identiques, est une première dans notre pays.

Les idées générales que je voudrais développer sont d'abord qu'il est possible de le faire en pratique. Je vais donc vous montrer comment nous avons fait, en pratique, pour mettre en œuvre ce genre de méthode : mettre des demandeurs d'emploi dans un groupe de traitement potentiel et dans un groupe de contrôle, avec un degré de rigueur suffisant. Je voudrais vous expliquer que cette idée de l'échantillonnage aléatoire n'est pas une idée bizarre, mais que c'est plutôt un principe général qui peut se décliner et s'insérer très facilement, ou en tout cas de façon possible, dans les contraintes opérationnelles de la conduite d'un projet.

Il y a aussi des contraintes, en particulier la question du suivi, très importante : quelle est la nature des informations que l'on recueille sur les individus qui passent par les dispositifs ? Ce sont des aspects en général importants parce que beaucoup d'entre eux sont associés au budget. Toutes ces opérations sont coûteuses. Le coût de collecte des informations représente une grande partie du budget de ces opérations.

Il existe aussi des contraintes liées aux délais. Il faut attendre pour obtenir les résultats. Il existe aussi des problèmes : parfois, les choses ne se passent pas exactement comme on le voudrait. Se pose ainsi la question du taux de participation : on propose à des gens de participer à un programme et puis, finalement, ils ne sont pas tellement intéressés. Cela pose des problèmes. Il y a encore la question de l'alimentation des dispositifs. Ainsi, quarante mille demandeurs d'emploi devaient passer dans le dispositif OPP ; quarante mille dans le dispositif « Cap vers l'entreprise ». Au bout du compte, il faut que les quarante mille soient passés et que, malgré tout, on ait réussi à constituer un groupe de contrôle et un groupe de traitement potentiel. Ce sont des problèmes qu'il faut gérer au quotidien, des aspects importants.

Malgré ces contraintes et ces problèmes, ce type d'analyse produit des résultats de façon transparente. Et ce que je trouve intéressant et important, c'est que ce sont des résultats qui ont la force de l'évidence. C'est-à-dire qu'ils vont s'imposer à tous et, plutôt que d'avoir des résultats d'évaluation et d'étude qui vont faire débat, qui vont sans cesse être l'objet de

disputes ou de querelles de spécialistes, on va avoir des résultats qui, au contraire, sont plutôt de nature à alimenter le débat et qui sont susceptibles de faire progresser la connaissance et la conduite des politiques économiques.

Un mot sur qui nous sommes : nous sommes un ensemble de chercheurs, principalement du CREST et de l'Ecole d'économie de Paris, qui essayons de développer ces méthodes d'évaluation depuis plusieurs années. Parmi toute cette liste, on est très nombreux à être membres du laboratoire J-PAL Europe, qui vous sera présenté un peu plus tard dans l'après-midi.

Il y a un point important auquel je crois particulièrement – je suis administrateur de l'Insee – dans la conduite de ces projets : c'est le partenariat de qualité que l'on peut avoir avec les porteurs de projets et les services statistiques. Nombre de projets que l'on conduit actuellement ou que l'on a conduits sont nés de ce type de partenariat et en particulier le partenariat que l'on a avec le DARES, qui nous aide énormément à mener ces projets.

Pour vous expliquer comment on fait, en pratique, je vais prendre l'exemple du RMI. La population totale des bénéficiaires dans le stock, par exemple, d'individus qui étaient au RMI depuis plus de deux ans, est de quatorze mille individus. Cette population, on la répartit aléatoirement en deux sous-populations, la population A et la population B, et dans les proportions 25 % et 75 %. Le choix de 75 % nous a été dicté parce que le prestataire de services a une certaine capacité de production, de RMIstes qu'il peut accompagner, et il lui a semblé qu'il fallait qu'il puisse toucher ou chercher à toucher 75 % des quatorze mille bénéficiaires potentiels. Sur la population A, on donne des incitations réduites à entrer dans le dispositif. Par contre, sur la population B, on donne des incitations renforcées. Par exemple, les incitations réduites sont les affiches qui se trouvaient partout dans les villes des Hauts-de-Seine pour annoncer ce programme d'accompagnement. C'était aussi le bouche-à-oreille, le fait que les personnes se parlent, et se parlent éventuellement du programme, ainsi que les travailleurs sociaux, qui sont au courant de l'existence de ces programmes et qui, naturellement, vont aller spontanément en parler aux bénéficiaires potentiels. La population B, elle, reçoit des incitations renforcées. De la même façon, elle regarde les affiches, entend les autres parler des programmes, voit des travailleurs sociaux. Mais, en plus, elle reçoit un courrier, on leur passer des coups de téléphone et, si cela ne suffit pas, on recommence, on les relance, jusqu'à obtenir une adhésion de leur part.

Comment cela fonctionne-t-il, en pratique ? Dans la population A, 6 % des individus sont entrés dans le programme, malgré l'absence d'incitations renforcées. Dans la population B, 17 % sont entrés en présence d'incitations renforcées, grâce aux affiches, au bouche-à-oreille mais aussi aux coups de téléphone et aux lettres qu'on leur a envoyées. 83 % ne sont pas rentrés même en présence d'incitations renforcées. Ces 17 % sont malgré tout assez faibles. C'est déjà un objet important que de s'en rendre compte.

Ces 17 % peuvent être répartis en deux sous-populations, qu'on ne pourra pas distinguer : ceux qui seraient entrés même en l'absence d'incitation renforcée et une partie entrée grâce aux incitations renforcées. De la même façon, dans la population A, tous ceux qui ne sont pas entrés se décomposent en deux catégories d'individus : ceux qui ne sont pas entrés, mais qui

seraient entrés en présence d'incitations renforcées ; ceux qui ne seraient pas entrés même en présence d'incitations renforcées. Les 6 % du haut sont identiques. Ceux qui sont entrés malgré l'absence d'incitations renforcées dans la population A, on les observe, mais ils sont aussi présents dans la population B et ils sont statistiquement identiques. Il y a la même proportion d'individus, de quelque nature que ce soit, dans chacune des deux populations. Et puis, de la même façon, ceux qui ne sont pas entrés, même en présence d'incitations renforcées et ceux qui ne seraient pas entrés, même en présence d'incitations renforcées, sont identiques également. Ceux du milieu, les 17 – 6 %, les 11 % qui sont rentrés grâce à l'incitation renforcée, ou qui ne sont pas rentrés mais seraient rentrés en présence d'incitations renforcées vont être différents, à cause du programme. La seule chose qui va distinguer ces deux populations, c'est que cette sous-population des 11 % est différente à cause du programme. Pour toutes les autres caractéristiques, ils sont identiques. Cela nous conduit à mesurer l'effet, par exemple si l'on s'intéresse au RMI qui a été versé. Le RMI moyen versé dans le groupe B, moins le RMI moyen versé dans la population A – ce que Esther Duflo avait appelé ce matin *intention to treat* – est rapporté à la différence de taux d'entrée, le taux d'entrée dans B, moins le taux d'entrée dans A, c'est-à-dire les 11 % dans le dispositif, qui sont rentrés dans le dispositif à cause des incitations renforcées.

En pratique, cela est très simple à mettre en œuvre. Cela fonctionne très bien et donne des résultats qui permettent, sans ambiguïté, de mesurer la plus-value apportée par le dispositif de façon transparente. Il est vrai que ce n'est pas n'importe quelle population. C'est la population des 11 % qui sont rentrés à cause des incitations renforcées. C'est une population particulière. C'est ce schéma que nous avons suivi dans les évaluations que nous avons menées – Esther Duflo a montré ce matin de nombreux schémas possibles –, qui se résume facilement : la constitution en amont de deux populations statistiquement identiques ; ensuite, des incitations différentes à rentrer dans le programme sur chacune des deux populations.

Je vous ai mis ici un tableau qui vous montre que les caractéristiques dans la population A et dans la population B sont identiques. Quand on a fait cette opération de mettre aléatoirement les individus dans la population A ou dans la population B, on a deux populations qui sont statistiquement identiques. C'est-à-dire que c'est la même proportion de femmes, la même proportion d'individus qui ont moins de trente ans ; la même proportion d'individus qui ont entre trente et quarante ans ; la même proportion d'individus qui ont une ancienneté de deux à quatre ans ; vous le voyez aussi, les bénéfices qui ont été perçus au trimestre antérieur à l'entrée dans le dispositif sont les mêmes. Cette opération de randomisation permet d'avoir deux populations statistiquement identiques.

Voici les résultats que l'on trouve. Vous avez, dans la première colonne, l'effet global sur toute la population. Trimestre 1, 2, 3, 4, 5, 6 et 1, 2, 3, 4, 5 ou 6 trimestres après être rentrés dans le dispositif. Ce que vous voyez, ce sont les économies en termes de RMI versé. C'est la seule variable que l'on avait. Le fait de participer au programme, au bout d'un trimestre, réduit les versements du RMI de 46 €. Ce n'est pas énorme. Au deuxième trimestre, 181 € ; au troisième, 282 €. Celui-là, en plus, a une étoile, parce que l'on peut accepter l'hypothèse qu'il est différent de zéro, statistiquement. L'écart type, à côté, mesure la précision de notre estimation. Ensuite, pour les trimestres 4, 5 et 6, on voit que l'effet diminue. Ainsi, ce que

montre sans ambiguïté notre évaluation, c'est que l'effet du traitement, de l'accompagnement sur la population globale des bénéficiaires qui sont rentrés à cause de nos incitations renforcées sur cette population est nul. Ce programme ne permet pas de réduire de façon significative le RMI versé en moyenne aux individus qui en ont bénéficié. Ce n'est évidemment pas une bonne nouvelle pour le programme.

Vous avez ensuite une deuxième colonne, qui vous montre l'effet sur une sous-population. Cette sous-population est celle qui, avant d'entrer dans le dispositif, percevait un RMI d'au moins 400 €. C'est la seule information dont nous disposons dans ce fichier. Un RMI de plus de 400 € signifie qu'il s'agit vraisemblablement d'individus en couple et qui ont des enfants. Leur situation n'est pas trop dégradée sur le plan social. L'on constate ici des choses tout à fait différentes : dès le premier trimestre, la réduction de RMI versé est de 493 €. Sur le deuxième trimestre, elle est de 826 € ; au troisième trimestre, 879 € et cela continue ainsi. A chaque fois, ou presque, j'ai mis une étoile, ce qui signifie que cet effet peut être jugé statistiquement significatif, ce qui est un critère très important sur le plan de l'appréciation de l'effet. Au total, pour cette population, les économies de RMI cumulées sur six trimestres sont de 4258 €. Voilà ce que l'on peut apprendre avec ce type d'évaluations et comment elles peuvent être utiles, puisque cette évaluation montre que, globalement, le programme est inefficace – ce qui est déjà une information extrêmement importante – mais que par ailleurs, lorsqu'il est ciblé sur certaines populations, il peut être très efficace.

Comment a-t-on fait dans les autres programmes ? Je vais vous montrer les systèmes d'échantillonnage que l'on a utilisés. Dans le premier, le RMI, c'est ce que je viens d'expliquer. Dans le deuxième, pour les opérateurs privés et « Cap vers l'entreprise », on a constitué à chaque fois deux ou trois listes suivant que les individus sont dans le flux indemnisable ou non. Une caractéristique importante à ce deuxième cas, c'est que dans le groupe de contrôle il n'y avait pas d'entrée possible dans le programme. Enfin, on a utilisé un outil spécifique pour l'affectation, qui a été développé par l'ANPE et qui s'appelle l'outil de constitution des cohortes, OCC. Cet outil a été très utile.

Enfin, pour les jeunes diplômés, on a mis en œuvre un dispositif particulier qui permet de mesurer non seulement l'effet direct, mais aussi l'existence d'un effet indirect.

Pour les opérateurs privés de placement, il y a quarante mille places ; pour « Cap vers l'entreprise » également. Le public était les demandeurs d'emploi avec risque de chômage de longue durée. C'est une population identifiée lors d'un entretien que les demandeurs d'emploi ont régulièrement avec les conseillers ANPE. Le potentiel des personnes éligibles est estimé à quatre cent cinquante mille demandeurs d'emploi, avec trois sous-populations : celles qui ont un flux de plus de cent vingt jours d'indemnisation, celles qui ont moins de cent vingt jours d'indemnisation et puis la population des demandeurs d'emploi en stock. Il n'y avait pas de participation obligatoire au programme « Cap vers l'entreprise ».

Ici, un tableau vous montre les potentiels. Pour « Cap vers l'entreprise », il y a quarante mille places ; pour OPP également. Enfin, dans la dernière ligne, vous voyez le potentiel : cent vingt mille pour le flux indemnisable, soixante mille pour le flux non indemnisable et deux cent soixante-dix mille pour le stock. Ce sont les chiffres qu'on avait calculés à l'origine,

lorsqu'on a démarré ce programme. Il y a donc beaucoup plus de bénéficiaires potentiels qu'il n'y a de place pour chacune des populations.

On a ventilé les demandeurs d'emploi du flux indemnisable en trois catégories : pour les opérateurs privés, avec une certaine probabilité ; pour « Cap vers l'entreprise » avec une autre probabilité ; enfin, ceux qui étaient envoyés en parcours classique, avec une troisième probabilité. La somme de ces probabilités fait un.

Ces probabilités ont été fixées à 85 %, 7,5 % et 7,5 % dans le flux indemnisable. Ces probabilités sont très déséquilibrées, parce qu'il était nécessaire que l'affectation que nous avions des demandeurs d'emploi satisfasse également les critères quantitatifs d'entrée dans le dispositif qui étaient fixés par l'ANPE et par l'Unédic.

Pour faire cette ventilation, on a utilisé un outil informatique qui a été développé par l'ANPE, installé dans quatre cent cinquante agences et qui a ventilé au hasard les demandeurs d'emploi dans les différents programmes, pour ensuite enregistrer l'affectation. Un seul passage était possible : une fois qu'un individu était passé par cet outil, il était impossible de le réaffecter à un autre public. Cet outil a été très puissant et sans lui l'évaluation n'aurait pas été possible. Il a été assez compliqué, pour l'ANPE, de faire accepter cet outil, parce que cela a beaucoup changé les méthodes de travail des personnes de l'ANPE, qui étaient quotidiennement au contact des demandeurs d'emploi. Cela a été assez courageux, de la part de l'ANPE, de se plier à la contrainte du développement de cet outil et de le maintenir, malgré l'ensemble des difficultés qu'il a présentées.

Jusqu'à présent, voilà la montée en charge du dispositif, jusqu'au quatrième trimestre 2007. Vous voyez que dans le stock, on a fait passer quatre-vingt-seize mille individus dans le dispositif, soixante mille dans le flux non indemnisable et soixante-quatorze mille dans le flux indemnisables. Ces chiffres doivent vous paraître faramineux par rapport aux chiffres qui ont été présentés ce matin, en particulier par Judith Gueron. Cela est essentiellement lié au fait que dans cette opération les demandeurs d'emploi peuvent refuser de rentrer dans le dispositif, ce qui nécessite que beaucoup de monde passe par le dispositif. De plus, les probabilités étaient très fortement déséquilibrées en faveur des opérateurs privés : cinquante mille personnes ont été envoyées aux opérateurs privés.

De la même façon que précédemment, on a comme ça des populations identiques.

Pour l'accompagnement des jeunes diplômés, la question qui nous a intéressés est celle de l'existence d'externalité. Une question récurrente sur l'accompagnement des demandeurs d'emploi consiste à déterminer si, lorsque l'on accompagne quelqu'un et qu'il retrouve un emploi, finalement, on ne fait pas que faire tourner les individus dans la file d'attente ? C'est-à-dire que l'on prend ceux qui étaient à la queue, on les aide, ils sortent, mais ceux qui allaient sortir, voient leur sortie du chômage retardée. Alors, on a mis en place un dispositif spécial, qui va nous permettre de mesurer non seulement la sortie vers l'emploi des bénéficiaires, mais également le fait que les non-bénéficiaires puissent voir leur situation affectée. Sur l'ensemble des agences locales de l'emploi dans lesquelles le programme a été développé, on a créé quarante quintuplés d'agence. Près de deux cents agences étaient concernées par le

programme. On a fait quarante quintuplés et au sein de chacun d'eux, on a tiré au sort les agences locales pour l'emploi dans lesquelles on n'allait envoyer personne à l'opérateur qui était chargé de développer le programme, celles dans lesquelles on allait en envoyer 25 %, 50 %, 75 % ou enfin 100 %. Ainsi, cela nous permet de mesurer, en comparant l'ensemble des quintuplés entre eux, non seulement l'effet du programme pour les bénéficiaires – les gens qui étaient dans une agences locales pour l'emploi avec 50 % par exemple, comparés à ceux qui étaient dans une agences locales pour l'emploi dans laquelle il n'y a eu personne – et donc la plus-value spécifique apportée par le programme aux individus qui étaient dans une agences locales pour l'emploi avec 50 % envoyés. Mais, dans cette même agences locales pour l'emploi, 50 % des individus n'ont pas été envoyés au programme. Ceux-là ont éventuellement subi un effet négatif du fait que 50 % des demandeurs d'emploi éligibles avaient été envoyés à l'opérateur privé. Donc, en comparant les 50 % qui n'ont pas été envoyés à l'opérateur privé avec les demandeurs d'emploi qui étaient dans l' agences locales pour l'emploi sans envoi aux opérateurs privés de placement, on va être capable de mesurer l'existence potentielle d'un effet négatif de l'accompagnement sur les non-bénéficiaires. Cela pour vous montrer qu'avec ce principe général de l'échantillonnage aléatoire, on peut mesurer non seulement la plus-value directe, mais aussi essayer de mesurer des plus-values indirectes sur le dispositif.

Je voudrais maintenant passer à l'aspect suivi des demandeurs d'emploi. C'est une question extrêmement importante et tout à fait complémentaire de celle du *design* expérimental. C'est-à-dire qu'il existe une complémentarité très forte, c'est une très bonne chose d'avoir un *design* expérimental qui soit très sophistiqué et que l'on puisse mettre en œuvre de façon parfaitement rigoureuse, avec des outils puissants comme l'outil de constitution des cohortes de l'ANPE. Mais ensuite, par ailleurs, il est extrêmement important de développer un système de mesure et de suivi des demandeurs d'emploi, qui soit de grande qualité. Ainsi, j'ai été extrêmement frappé, hier, par l'opération *Moving to Opportunity* qui avait un système d'information particulièrement développé et très sophistiqué. Il est certainement très intéressant de s'inspirer de ce genre de tentative pour les opérations futures.

On en est resté à des questions très traditionnelles. On s'intéressait à une seule chose, qui était le maintien dans l'emploi. Un principe était essentiel : il ne fallait pas de non-réponse. On a concentré tous les moyens qu'on avait – on a un budget limité, mais tout de même assez important grâce à la DARES notamment – pour pouvoir suivre les demandeurs d'emploi. On s'est appuyé, pour les opérateurs privés et « Cap vers l'entreprise », en partie sur le fichier historique statistique de l'ANPE et en partie sur un système d'enquêtes complémentaires, qui nous a permis de mesurer précisément la situation des demandeurs d'emploi.

On a rencontré différents types de difficultés, en particulier de convaincre nos partenaires que c'était bien cela qu'il fallait faire, plutôt que d'avoir des enquêtes très larges, sur lesquelles le taux de non-réponse serait important. C'était résister à la tentation des économies d'échelle. On a eu des problèmes avec le Code des marchés publics parce que, souvent, il faut faire ces enquêtes très rapidement, parce que ces projets ont une durée. Il faut faire l'enquête au bon moment. Il ne sert à rien de la faire plus tard. Il est donc extrêmement important de pouvoir

contracter dans un bon *timing* avec une entreprise qui va faire les enquêtes. De ce fait, le Code des marchés publics présente un certain nombre de problèmes.

Par exemple, pour le RMI, on avait un système d'information très pauvre, parce que l'on voulait absolument se lancer dans cette opération et avoir un tout petit budget. Donc, on avait seulement les données administratives de la Caisse d'allocations familiales. Il y avait très peu de variables. La seule variable de résultat était le RMI. Il y avait très peu de variables, aussi, qui nous permettent de caractériser la population.

Pour le programme opérateurs privés et « Cap vers l'entreprise », les données dont on dispose sont celles du fichier historique statistique de l'ANPE. Mais, là, il existe un problème bien connu de ceux qui utilisent ce fichier, c'est qu'il a des sorties inconnues. Beaucoup quittent le chômage sans dire où ils vont. On a mis un système d'enquêtes complémentaires. Au bout du compte, avec tous ces efforts, on a réussi à avoir des informations sur 90 % des individus présents dans l'échantillon à l'origine.

Enfin, pour les jeunes diplômés, on a deux systèmes parallèles : un système avec des enquêtes et un système administratif. On va essayer de faire plusieurs enquêtes, transversales sur l'ensemble de la population.

Les problèmes que l'on a sont en termes de taux de participation. Pour le RMI, par exemple, 17 % seulement des personnes étaient rentrées dans le programme. Pour les opérateurs privés et « Cap vers l'entreprise », ce sont 50 % des personnes. Ces problèmes sont importants. Esther Duflo a parlé ce matin de la puissance. Cela fait perdre beaucoup de puissance à nos évaluations. Il va donc falloir que l'effet soit extrêmement important pour que l'on puisse le détecter. C'est aussi l'une des raisons pour lesquelles on n'a pas hésité à avoir des échantillons assez nombreux pour l'opérateur privé et « Cap vers l'entreprise ».

Le deuxième problème est bien évidemment celui de l'alimentation. Il y a eu des tensions un peu récurrentes pour fournir aux opérateurs privés les demandeurs d'emploi qui devaient leur être envoyés contractuellement. C'est aussi la raison pour laquelle les probabilités ont été très déséquilibrées et que, au bout du compte, on a un système pas très productif. Enfin, il existe des problèmes d'autre nature, de contamination. Il y a eu des envois directs aux opérateurs privés, en dehors de l'outil de constitution des cohortes. C'est de nature à perturber nos évaluations. Enfin, les délais : il faut attendre pour obtenir des résultats et pourtant maintenir l'effort et l'adhésion de tous. En dépit de ces problèmes, au bout du compte, on aura des résultats qui permettent de mesurer de façon certaine l'effet des politiques que l'on met en œuvre sur leurs bénéficiaires. En particulier, pour le mois de juin, on va avoir des résultats sur opérateurs privés et « Cap vers l'entreprise ». Il faut donc attendre, mais là, on va être capables de mesurer de façon précise et sans ambiguïté, la plus-value apportée par ces programmes.

En conclusion, c'est un début. On sort juste des ténèbres. Cela marche bien et ce n'est pas compliqué. Il est vrai qu'il faut coordonner de nombreux efforts. On a certainement des marges et des gains d'efficacité dans la conduite de ces opérations mais on va essayer de s'inspirer d'autres expériences. L'outil de constitution des cohortes qui a été développé par

l'ANPE a certainement joué un rôle central. Il est souhaitable que ce type d'outil, capable de graver un peu dans le marbre l'affectation initiale des individus, puisse se diffuser et être plus utilisé. Enfin, on ne peut que saluer l'appel d'offres du Haut Commissariat aux Solidarités actives, qui ouvre de nouvelles perspectives de développement pour ce type de politique.

Pierre CAHUC
Merci beaucoup. Nous avons le temps pour quelques questions, s'il y en a.

Claude SEIBEL
Je ne devais pas être initialement présent, mais je dois dire que l'exposé de Bruno est tout à fait intéressant. J'interviens au titre de président du comité de pilotage de l'évaluation Expérimentation de l'accompagnement renforcé. Cette démarche est, comme vous l'avez bien compris, particulièrement innovante sur plusieurs dimensions. La dimension de l'importance du chantier : en fait, dans l'outil de constitution de cohorte, il est passé deux cent trente mille personnes qui ont été "moulinées" et affectées dans ces trois voies. La deuxième, c'est évidemment le recours à une affectation aléatoire, qui est un peu le centre de cette rencontre. La troisième, c'est aussi la complexité d'un chantier dans lequel interviennent, avec des enjeux quand même assez différents, l'Unédic, l'ANPE, une direction d'études de la DARES qui essaie de jouer un peu le rôle d'architecte de l'ensemble, sachant que le thème que nous présente Bruno, c'est un peu la racine qui organise le reste de l'évaluation, qui comporte notamment des aspects de monographie auprès d'analystes des processus, à l'intérieur de l'ANPE ou du point de vue des demandeurs d'emploi, des cellules CVE ou d'OPP. Ces deux monographies sont particulièrement passionnantes, parce qu'elles peuvent être réinjectées très vite dans une amélioration de l'ensemble du dispositif. Voilà les trois dimensions d'innovations qui me semblent importantes.

Cela étant, je voudrais rajouter brièvement quelque chose, qui a peut-être été évoqué ce matin, mais je ne pouvais pas être là, c'est l'aspect éthique sous-jacent à ces dispositifs. L'utilisation d'un outil de tirage au hasard plaît beaucoup à tous les chercheurs ici présents, comme j'ai cru le comprendre. Mais, ce n'est pas du tout le cas des acteurs, qui se trouvent souvent dans une situation difficile, soit parce qu'ils n'ont pas perçu l'importance de faire le couple expérimentation-évaluation, soit parce que, pour toute une série de raisons, ils ne veulent pas rentrer dans ce jeu. Ces travaux, nous devons absolument les accompagner d'une réflexion éthique qui puisse être dans la négociation du protocole d'évaluation.

Dans le cas présent de l'accompagnement renforcé des demandeurs d'emploi, en réalité, deux groupes sont dans des situations un peu "sportives" par rapport à l'outil de constitution des cohortes. Vous avez des gens qui le voient tout de suite : les conseillers de l'ANPE, dont le métier est remplacé par une table de tirage au hasard, même s'ils ne savent pas très bien comment cela fonctionne précisément. D'ailleurs, le OCC a été appelé au sein de l'ANPE, "la roue de la fortune", "Madame Soleil", etc. La réponse à cela consiste à dire que cela est limité dans le temps. En particulier, cela fait quinze mois. Ensuite, vous allez reprendre les rênes – c'est ce qu'il s'est fait, d'ailleurs au mois de mars – et il sera très intéressant d'observer, au sein des agences ALE, la sortie d'OCC, la sortie de l'outil de constitution de cohortes. Parce que, *a priori*, il ne sera pas forcément si facile que cela de retrouver la maîtrise de

l'affectation vers les opérateurs privés ou vers les cellules CVE. Cela est gérable, mais encore faut-il l'avoir discuté et négocié.

Par contre, là où cela est plus compliqué, c'est vis-à-vis des demandeurs d'emploi. En fait, tu nous as montré ces parties qui acceptent, refusent, etc. Au niveau du comité de pilotage, nous avons proposé à l'ANPE que les personnes qui auraient été désireuses d'entrer dans un accompagnement renforcé, l'on puisse le leur proposer, même s'il y a un décalage dans le temps, puisqu'il se sera écoulé plusieurs mois. Ils seront prioritaires pour un accompagnement renforcé. Je vous remercie.

Pierre CAHUC
Avez-vous d'autres questions ?

Alberto LOPEZ, CEREQ
J'avais une question sur tout ce qui concerne le protocole sur les jeunes demandeurs d'emploi diplômés. Je n'ai pas très bien compris sur quels champs on allait observer ou tenter d'observer les éventuels effets de substitution. Effectivement, on se dit que les jeunes demandeurs d'emploi diplômés qui sortent du système éducatif sont en concurrence entre eux, mais aussi avec une foule de gens qui ont un niveau de qualification correspondant. La question est de savoir si cela veut dire que l'on va observer la trajectoire de pratiquement l'ensemble des demandeurs d'emploi qualifiés des ANPE ou si l'on reste limité sur le champ des jeunes récemment sortis du système éducatif.

Bruno CREPON
Dans l'étude telle qu'on la mène, on va s'intéresser à l'effet direct sur les bénéficiaires potentiels, qui ont effectivement bénéficié, et puis à l'effet indirect sur les bénéficiaires potentiels – les mêmes, les jeunes de moins de trente ans – mais qui, eux, avaient été mis dans le groupe de contrôle, c'est-à-dire dans les 50 % ou dans ceux qui ont refusé la participation au programme. Ce n'est effectivement pas la totalité de l'effet de diffusion. L'effet de diffusion pourrait aussi aller vers d'autres publics. Certainement, l'on doit pouvoir examiner un petit peu, avec un degré de fiabilité raisonnable, l'effet sur d'autres publics, qui ne sont pas exactement les éligibles. Il nous semblait intéressant de nous focaliser sur cette population.

Jonathan PORTES
Yes, just to follow that up, I'm still not sure what the theory is here. Is it that the people who are not benefiting from the program (who are in the Control) are competing for the same jobs, and are therefore disadvantaged because the people helped get those jobs, in which case the probability of detecting an effect must be very small because the whole labor market is much larger, and even if the program itself is very effective, it's still taking only a very small portion of jobs? Or is the theory that the agencies will be spending so much time on the people in the program that they will not worry about the people who have been "Controlled out," and hence those people will get a worse service as a consequence, which is more plausible?

Bruno CREPON

No, the idea is that first, you are crowding out effect, related to the fact that individuals who are counseled may find a job that would have been taken by people that have not been counseled; and it is possible that this will be the case because the individuals or the labor markets we are considering are very local. It is local employment agencies in a very, very small area, like a small town: a local labor market. Some effect may exist.

Pierre CAHUC

Je propose que nous passions à la prochaine intervention. Merci beaucoup, Bruno. C'est au tour d'Eric Maurin et Olivier Noblecourt.

Retour sommaire

L'appel à projets d'expérimentation sociale du Haut Commissariat aux Solidarités actives : présentation des projets sélectionnés

Olivier NOBLECOURT, mairie de Grenoble

Bonjour à toutes et à tous. Je vais commencer cet exposé à deux voix, puisque la transition a été parfaitement faite par la précédente conclusion, d'abord en me présentant.

Je m'appelle Olivier Noblecourt. Pendant un peu plus de dix ans, j'ai été conseiller au cabinet du maire du Grenoble et le directeur de son cabinet pendant plus de sept ans. J'ai notamment travaillé sur le suivi du Conseil national des missions locales, pendant quatre années. Je suis aujourd'hui adjoint en charge des Affaires sociales à la Ville de Grenoble et notamment du CCAS.

Si j'ai été membre du jury de l'appel à projets, c'est parce que c'est à Grenoble qu'a été lancé cet appel à projets et qu'a été lancé le Grenelle de l'insertion, sachant que la démarche de valorisation des expérimentations sociales que nous avions engagée à ce moment-là avec Solidarités actives, a été rejointe par la volonté du président de la République de confier à Martin Hirsch l'organisation du Grenelle de l'insertion.

Grenoble, en deux mots, comme vous le savez certainement, est une ville qui a un rapport identitaire à l'innovation sociale. C'est la ville du rapport Dubedout, qui a lancé le développement social des quartiers ; celle du rapport Schwartz sur les missions locales ; celle de la médecine salariée avec les centres de santé depuis le milieu des années 70. D'une certaine manière, nous avons, non pas des leçons à donner, mais travailler peut-être davantage que les autres sur ces aspects.

En tant qu'élu local, et c'est un peu la conviction que j'ai voulu porter au sein du jury : nous sommes quand même sur des politiques qui sont à la fois en crise de financement et en crise de légitimité politique. Je parle des politiques en général de solidarité. Ainsi, j'ai beaucoup aimé l'expression de Bruno Crépon sur les résultats ayant la force de l'évidence. Je le dis aujourd'hui, notre nécessité absolue, c'est la force de la preuve de l'efficacité de chaque euro investi dans les politiques sociales. Pour avoir vécu un certain nombre de négociations budgétaires difficiles, on sait bien que la culture d'un ministère comme Bercy ou d'un certain nombre de collectivités territoriales est quand même de s'interroger très fortement sur les budgets souvent considérables qui sont alloués aux politiques sociales. Donc, pour l'avenir et dans un contexte idéologique qui, selon moi, est assez peu porteur, il y a un enjeu tout à fait majeur à promouvoir l'expérimentation sociale. Cette expérimentation – cela a dû être dit largement entre hier et aujourd'hui – est entendue comme la combinaison entre une innovation sociale réelle, c'est-à-dire modélisable, et une évaluation qui permette de dégager des résultats incontestables.

Dans cette démarche d'appel à projets, il y avait pour moi, au départ, une double démarche. Une démarche à la fois qu'il fallait saluer, du Haut Commissariat, de promotion de l'expérimentation sociale dans les territoires. Mais aussi une démarche, pour les territoires eux-mêmes, d'étalonnement des pratiques d'innovation sociale, d'étalonnement des pratiques

en termes de nouvelles politiques publiques et donc la nécessité pour nous, acteurs locaux, de sortir du microlocal, de la seule adaptation au territoire, qui justifie souvent beaucoup de bricolage, et donc d'être réellement dans la recherche de canons, de normes d'expérimentation. Et, là-dessus, les territoires ont réellement besoin que l'Etat leur donne un cadre et des moyens.

De ce point de vue-là, l'appel à projets du Haut Commissariat avait certaines similitudes avec l'appel à projets pour les pôles de compétitivité, dont les enjeux, à la fois en termes de politique industrielle pour l'emploi et les moyens financiers, étaient sans commune mesure. Mais, d'une certaine manière, il y avait un effet de test sur les territoires, consistant à dire : « Nous vous ouvrons la possibilité de valoriser vos initiatives : défendez-les, montrez-nous comment vous savez aussi monter localement des protocoles d'évaluation pour garantir leur transférabilité et leur modélisation ». Cette démarche innovante du Haut Commissariat a été tout à fait majeure, d'autant qu'aux deux enjeux que j'évoquais tout à l'heure, à la fois pour l'Etat et pour les territoires, se greffe quand même un autre enjeu de nature plus politique, qui sont ces six millions d'euros de crédits qui ont été alloués à l'appel à projets, par redéploiement de crédits. Nous sommes malheureusement sur des politiques où les crédits nouveaux sont assez rares à trouver. Dans les débats qu'on a eus à l'intérieur du jury, une question importante était celle de tout dépenser ou non. Je plaidais bien évidemment pour tout dépenser, puisque nous étions dans un secteur sous-financé. L'Histoire a montré que nous n'avons pas tout dépensé *stricto sensu* sur les trente-sept projets qui ont été retenus, mais nous avons l'engagement tout à fait formel du Haut Commissaire que les six millions d'euros seront intégralement dépensés.

L'autre contrainte est plus banale : c'est de réussir. Parce qu'on imagine bien que si sur ces seuls six millions d'euros et sur les seuls trente-sept projets qui ont été sélectionnés on ne tire pas de démonstration importante pour les politiques publiques, on pourra être extrêmement inquiets pour l'avenir du financement d'un certain nombre de politiques.

Le travail du jury, je ne vais pas insister très lourdement dessus, d'une part parce qu'un excellent dossier vous a été distribué, d'autre part, parce qu'Eric Maurin en développera les contraintes et les exigences scientifiques bien mieux que je ne saurais le faire. Juste dire rapidement que nous avons très vite obtenu quatre cent cinquante manifestations d'intention, ce qui est un beau chiffre ; qu'à l'issue d'une première sélection et d'une rencontre avec les porteurs de projet, deux cent quarante et un dossiers ont été examinés ; chaque membre du jury a eu entre vingt et quarante dossiers à regarder, sur lesquels il a rapporté – évidemment, chaque membre du jury ne pouvait pas rapporter sur les dossiers où il pouvait avoir un conflit d'intérêt – et puis ces trente-sept projets sélectionnés qui marquent une sélectivité importante.

Dans les débats internes au jury, on a eu trois types de grandes questions. La première était : qui devons-nous financer ? Est-ce que notre rôle est d'apporter des compléments de financement à de l'action de collectivités locales qui, elles-mêmes pourraient, dans leurs arbitrages financiers, mettre les sommes nécessaires ? Est-ce que notre rôle est de financer de la recherche universitaire *stricto sensu* ? On a finalement accepté de financer des projets qui étaient véritablement des projets de thèse, en réalité, ou pas loin. Et puis, est-ce que notre rôle

est de financer par exemple les grands réseaux d'économie sociale et solidaire, pour les accompagner dans leurs mutations et dans leurs enjeux – on a eu notamment le débat sur le COORACE. C'étaient véritablement des questions qui ont jalonné le travail du jury, notamment tout au long des deux journées de sélection finale.

Deux autres questions importantes : évaluer pour nous et se mettre d'accord sur ce que recouvre réellement la notion d'innovation. L'innovation, c'est quand même un peu le mot "tarte à la crème", en particulier, je suis bien placé pour le savoir, dans le discours des élus, parce que l'on veut voir l'innovation partout, là où souvent c'est de l'adaptation au territoire, du compromis interinstitutionnel pour des raisons financières. L'innovation sociale, pour nous – et c'est comme cela qu'on l'a traduit – c'est que n'est innovant que ce qui est transférable et modélisable. On a donc éliminé énormément de projets sur le critère de l'innovation réelle.

Le deuxième débat – et je n'insisterai pas non plus, Eric en parlera, ainsi que Monsieur Bourguignon juste après – repose sur la question de quels modes d'évaluation on veut privilégier. Grandes querelles des sciences sociales, l'évaluation expérimentale est-elle la seule à devoir être reconnue par le jury ? Est-ce que nous admettons d'autres types d'évaluation ? Là-dessus, nous avons adopté une position commune, qui a consisté à privilégier très largement – et la constitution des dossiers de candidature y invitait fortement – la méthode expérimentale d'évaluation. Mais de ne pas s'y restreindre sur le principe.

Un dernier mot, avant d'échanger plus directement, sur un constat général et quelques enseignements de la démarche à ce stade. D'abord, on a quand même été très frappés par le grand dynamisme du secteur de l'insertion et des politiques sociales, leur forte réactivité sur ce type d'appel à projets. C'est quelque chose d'extrêmement encourageant et qui rejoint, d'une certaine manière, le dynamisme qui avait été observé dans les territoires sur les projets économiques.

Quelque part, dans un état très centralisé – je pense que l'on porte tous ici cette conviction – c'était tout à fait majeur que les territoires puissent faire la démonstration de cette capacité, en peu de temps, de se fédérer et de porter des projets. Malheureusement, à part ce premier constat très positif, deux constats sont assez cruels : le premier est la faiblesse de l'innovation de manière générale, traduite par la sélectivité des projets. On a eu quand même à la fois des projets présentés comme innovants qui ne l'étaient pas et puis des problématiques auxquelles on s'attendait, qui n'étaient pas traitées, par exemple la fracture numérique, les problématiques logement et insertion, insertion et conduites addictives, l'accompagnement renforcé des RMIstes. De ce point de vue-là aussi, on peut constater avec regret, le peu de porosité des expériences étrangères sur la réflexion des acteurs locaux. Là-dessus, on a quand même le sentiment, toujours, d'être dans quelque chose de très franco-français ou dans le localo-local.

Le deuxième enseignement, c'est la faiblesse de l'évaluation. Cela est dit dans le dossier qui vous a été distribué, qui illustre la faiblesse des pratiques sur les territoires, les pratiques de travail et les liens entre équipes universitaires, entre scientifiques et acteurs de terrain. Cela est aussi tout à fait valable pour une ville comme Grenoble, qui se targue d'être à la pointe. Il y a une identification des acteurs scientifiques par les praticiens qui est tout de même

relativement faible et qui, me semble-t-il, doit interroger les universités en sciences sociales sur le territoire.

De la même façon, il faut que les acteurs s'interrogent eux-mêmes sur le caractère extrêmement endogène de leurs pratiques d'évaluation. La caricature classique, c'est l'élu qui veut monter un comité de pilotage et qui va s'autoévaluer avec ses techniciens une à deux fois par an, dans une réunion grand-messe. On a quand même une tradition franco-française d'autoévaluation, et où, en définitive, le recours, souvent cruel, à des scientifiques pour avoir une évaluation réellement objective est peu développé et peu admis. Les questions d'éthique ont été évoquées récemment. Il y a évidemment les questions d'égalité de traitement des citoyens, cette tradition française, mais, véritablement, l'enjeu, aussi, d'une démarche comme ces deux journées de colloque, consiste à se dire – et ce sera ma conclusion – que d'une part il est important que la réflexion ne soit pas circonscrite aux décideurs parisiens et que, véritablement, on aille plus loin dans la capacité à traduire cette conviction sur l'expérimentation sociale dans les territoires et que l'on mette en place des outils pour que, dans les territoires, les acteurs qui ont besoin d'évaluation de leurs politiques et de leurs actions puissent beaucoup plus facilement avoir accès à des financements et à de la méthodologie.

Là-dessus, on aura, je le crois, des enseignements généraux à tirer d'ici quelques mois ou quelques années, sur cet appel à projets. Mais il faut bien le considérer comme le début d'une démarche, et certainement pas comme un *one-shot* ou une fin en soi. C'est un petit volume budgétaire. C'est le début de quelque chose, mais il faut être extrêmement modeste sur l'impact que l'on peut en attendre. Il faudra aller beaucoup plus loin et donc insister là-dessus, sur la nécessité d'impliquer les élus et les réseaux d'élus là-dedans. Nous avons une démarche sur laquelle, là encore, même si sur le Grenelle de l'insertion on a pu évidemment aller beaucoup plus loin, nous marchons encore de manière trop séparée entre les acteurs administratifs et décideurs, les scientifiques et les élus locaux. C'est plutôt une invitation à renouveler ce type de temps d'échange et de colloque prochainement, si possible dans les territoires, pour mieux diffuser les quelques premières conclusions de ce travail de sélection que j'esquissais.

Retour sommaire

Eric MAURIN, EHESS

Je vais rapidement et concrètement vous parler de la façon dont se sont déroulées ces séances du comité d'évaluation des projets. On a défini des grands thèmes : éducation, insertion, quatre ou cinq grands thèmes dont le périmètre avait été dessiné par l'appel d'offres lui-même. On s'est partagé ces grands thèmes entre nous. Chaque projet a été évalué par deux coévaluateurs. Par exemple, j'étais avec Denis Meuret, un économiste de l'éducation – je suis économiste du travail – et tous les deux, nous avons regardé de manière indépendante l'ensemble des projets qui avaient trait à l'éducation et, dans le groupe de ces projets, l'ensemble de ceux qui avaient trait au soutien, à l'aide et à la sensibilisation des parents au

problème de l'école, à la formation des parents, au soutien à la parentalité. Au total, cela nous faisait, par paires, une vingtaine, une trentaine de dossiers à évaluer.

J'ai retrouvé la note que j'avais sur mon ordinateur, avec les critères que je m'étais dit que j'appliquerais pour noter les projets que j'allais recevoir. J'avais essayé de me parler à moi-même en français.

Premier critère, la pertinence de l'action publique évaluée. Il y a deux types de pertinence, à mes yeux, dans ce type de projet : une pertinence scientifique ou une pertinence politique. Une pertinence scientifique, c'est par exemple un projet qui va tester quelque chose dont on sent bien que cela est trop coûteux et trop compliqué pour en imaginer une généralisation mais qui, s'il aboutit à son terme, va nous apprendre quelque chose d'important et va nous donner un repère important. Pour moi, le paradigme de ce genre de projet, c'est par exemple tout ce qu'il s'est fait sur la petite enfance aux Etats-Unis. Ce sont des programmes qui traitent des groupes particuliers d'enfants très désavantagés, qui font un effort considérable sur ces enfants, qui mettent beaucoup de ressources sur ces enfants, beaucoup plus que celles que l'on peut imaginer pouvoir mettre *via* une école maternelle, un jour, sur des enfants en bas âge. Ils font quelque chose qui n'est politiquement pas pertinent ; mais, comme ils trouvent, de fait, qu'une action massive peut changer le destin de ces enfants, ils nous donnent un horizon, un horizon scientifique. Ils nous disent : ce n'est pas irrémédiable. Si vraiment l'on met les moyens sur des enfants en bas âge, aussi défavorisés soient-ils, aussi handicapés soient-ils, on peut changer leur destin. Ces expériences contrôlées sur les enfants en bas âge ont d'ailleurs eu une importance idéologique très forte tout au long du temps. Il reste des repères forts dans la représentation que l'on a, finalement, de ce qu'il est possible de changer ou pas avec la politique publique. C'est donc une pertinence scientifique.

Après, il y a d'autres types de projets : avant même de parler d'innovation, des projets beaucoup plus modestes, mais où l'intérêt est justement la réplicabilité. Ils vont tester quelque chose d'assez peu coûteux, dans des conditions assez peu extraordinaires. L'idée est de voir si l'on en tire quelque chose et si, finalement, le bilan coûts-bénéfices n'est pas si inintéressant que cela. C'est le deuxième type de pertinence d'un projet, la pertinence politique, plus que fondamentalement scientifique. Dans le cas du projet politique, on ne sait pas très bien ce que l'on identifie, au bout du compte, sinon une politique globale, un paquet, et on se demande s'il est, finalement, pour la collectivité, bénéfique de l'adopter ou pas.

Le deuxième critère, c'est : pourra-t-on, à l'issue de cette expérimentation, comparer des bénéficiaires avec des gens qui ne le sont pas ? Est-ce qu'il y aura que des bénéficiaires ou est-ce qu'il y aura, finalement, un élément de comparaison entre des personnes ? De fait, 90 % des projets qui nous arrivaient n'avaient pas cette propriété. Il n'y avait que des bénéficiaires et donc ce simple critère n'était pas respecté.

Le troisième critère, c'est, avant même de parler de tirage aléatoire – parce qu'on regarde beaucoup sur les exigences –, étant donné la façon dont est conçue l'expérimentation, existera-t-il de bonnes raisons de penser que ce qui arrive aux personnes qui ne bénéficient pas du dispositif est représentatif de ce qui serait arrivé à celles qui en bénéficient si elles n'avaient pas bénéficié de l'action particulière que l'on veut expérimenter ? Il y a ici une

condition d'or, le tirage aléatoire. Mais l'on peut imaginer d'autres formes de décalage dans le temps, de passages, qui peuvent finalement être intéressantes, faute de mieux. Cela dit, ce sont des conditions nécessaires mais, comme l'a dit Bruno au cours de son exposé, ce ne sont pas des conditions suffisantes non plus, dans la mesure où il y a des effets de contamination, où l'action dont on fait bénéficier les bénéficiaires a indirectement un effet sur les gens qui n'en bénéficient pas, par exemple dans le cas de l'éducation et des actions en direction des enfants.

Si les programmes de soutien dont vous faites bénéficier certains enfants, dans certaines classes, ont par le fait de l'amélioration qu'ils génèrent dans la classe, de l'ambiance de la classe, un effet bénéfique sur les enfants qui ne bénéficient pas directement de ces programmes de soutien, par la suite, si vous comparez ce qui arrive aux enfants qui ont vraiment bénéficié du soutien scolaire à ce qui arrive aux enfants qui n'en ont pas bénéficié au sein des mêmes classes, vous pouvez avoir une différence très faible, alors que pourtant, vous avez tiré au hasard vos enfants soutenus, bien que la politique soit doublement bénéfique, voire précisément quand elle est doublement bénéfique, à ceux qui sont soutenus et à ceux qui se trouvent avoir des camarades en moins mauvaise posture. Dans ce cas, le traitement aléatoire en présence de contamination est presque un remède pire que le mal, parce qu'il va nous faire croire dur comme fer à quelque chose qui, en fait, est biaisé. C'est le troisième critère, sur la qualité du groupe de contrôle et la qualité de l'indépendance du traitement entre le groupe de contrôle et le groupe traité.

L'autre critère que j'avais prévu de prendre en compte, c'est un critère qui est, finalement – Bruno l'a dit aussi entre les lignes –, quand on essaie de faire ces expérimentations, l'un des plus difficiles à réaliser. C'est le suivi des bénéficiaires et surtout celui des non-bénéficiaires, c'est-à-dire aboutir à un suivi de la même qualité, voir ce qui arrive après l'action publique que l'on cherche à analyser, aux gens qui en ont bénéficié – cela va encore, parce qu'ils sont dans une certaine façon en relation avec nous *via* le dispositif – mais aussi à ceux qui n'ont pas bénéficié de l'action, et pourtant, c'est une condition importante de la possibilité d'évaluation…

Le cinquième critère, c'est la taille de l'échantillon. L'expérience sera-t-elle capable de capter des effets raisonnablement faibles ? Les politiques que l'on cherche à évaluer sont souvent des politiques qui ne modifient qu'à la marge. Cela ne veut pas dire qu'elles n'ont pas un rapport coûts-bénéfices intéressant, parce qu'elles ne sont pas très coûteuses, mais les calculs de puissance nous révèlent qu'il faut quand même souvent plusieurs centaines de personnes, *minimum*, pour arriver à détecter des effets raisonnablement faibles.

Et puis le coût. J'avais prévu de penser au coût de ces expérimentations, en mettant en regard le coût d'une bonne enquête classique qui, aussi, nous apporte des connaissances sur le monde qui nous entoure. Plutôt que sur une expérimentation pharaonique, mais qui, en fait, ne permettra de comparer que dix traités et dix contrôlés, c'est-à-dire au bout du compte, a très peu de chances de trancher quoi que ce soit et qui coûte extrêmement cher, mieux vaut mettre l'argent sur une bonne enquête classique qui nous apprendra elle aussi des choses intéressantes.

Cela faisait six critères. Quand j'ai reçu les dossiers, l'écrasante majorité des projets que j'ai eu à évaluer ne respectait à peu près aucun d'eux. Sur les trente projets que j'ai pu évaluer, une vingtaine étaient conçus de la manière suivante : une association ou une antenne d'une institution nationale avec visiblement beaucoup d'intérêt pour la question qui proposait une action, quelquefois innovante, quelquefois non, auprès d'une centaine ou d'une cinquantaine de personnes et qui se proposait, à la fin de cette action, ce soutien aux parents, cette formation particulière à tel ou tel type, à tel ou tel problème, de demander aux bénéficiaires ce qu'ils avaient pensé, finalement, de ce qu'on leur avait proposé, s'ils trouvaient cela bien. Donc, beaucoup d'énergie et, en même temps, quelque chose de très loin de mes critères sans doute un peu "à côté de la plaque" mais qui me semblaient pourtant *ex ante* correspondre à l'esprit de cet appel d'offres.

Pour ne pas être totalement pessimiste, il y a quand même quelques professionnels qui s'étaient glissés dans le groupe. J'ai eu au moins pour ma part un projet qui respectait beaucoup des critères cités à évaluer. C'est un projet piloté par l'Université de Marne-la-Vallée, par Yannick L'Horty et Manon Dos Santos, visiblement très bien conçu et qui a été retenu. Il s'agit d'expérimenter un soutien à l'orientation auprès des jeunes lycéens, à l'entrée de l'université. C'est une opération qui se propose de tester l'efficacité de trois types d'action : une action d'aide à l'orientation au moment où les lycéens postulent pour accéder à l'université, avec un groupe de contrôle, un groupe traité, plusieurs centaines d'étudiants ; une action d'aide à l'orientation, à la réorientation, une forme de bilan avec un conseiller pédagogique en cours de la première année d'université ; ensuite, un troisième traitement qui s'imbrique dans les deux premiers, plus fondamental : le soutien à ceux qui en ont besoin, avec tous les canons de l'expérience contrôlée qui sont respectés. Inutile de préciser que c'est une action très pertinente sur le plan politique. Quand on considère l'ensemble de son enseignement supérieur, les taux d'échec à l'entrée en France ne sont pas si dramatiquement élevés par rapport à nos autres concurrents. Mais la structure de notre enseignement supérieur est telle que ces taux d'échecs sont concentrés et sont particulièrement visibles et massifs à certains endroits. Typiquement, les premiers cycles des universités classiques, non sélectives, dont l'université en question. Il y a vraiment une question importante qui est de savoir si, à moindre coût, l'on peut changer et améliorer les choses sur ce segment particulier de notre enseignement supérieur.

Les potentielles faiblesses de ce dispositif-là sont qu'à mon sens, il y aura une difficulté sensible à suivre les gens que l'on ne peut pas traiter – par exemple, les gens qui ont postulé, qui ne vont pas aller à l'université – ; il n'est pas clair, pour moi, de savoir dans quelle mesure il n'y aura pas une attrition assez sévère dans ce dispositif, dans la mesure où les gens vont aller à l'extérieur de l'université de Marne-la-Vallée. Il va y avoir des mobilités. Cela dit, je reste très optimiste sur ce projet, de par la qualité des gens qui le portent, mais aussi pour un deuxième aspect institutionnel important qu'il contient. Il y a une grande proximité, pour ne pas dire une identité, entre l'institution qui va porter l'évaluation et celle qui va piloter directement les actions que l'on cherche à évaluer. Les évaluateurs n'ont pas à négocier de loin avec une administration qui n'est pas nécessairement en phase, dans toute sa profondeur, avec les objectifs d'une évaluation.

D'autres projets ont été retenus, qui me concernent plus directement, qui n'ont pas pour propriété cette proximité entre les évaluateurs, les administrations et ceux qui ont piloté l'action ; là, je suis beaucoup plus pessimiste sur ce que l'on sera, au bout du compte, à même de produire comme évaluation.

Concrètement, un des projets qui a été initié par le rectorat de Créteil est par exemple une expérimentation aléatoire de soutien auprès des parents. Il est extrêmement difficile pour nous, évaluateurs, de nous assurer qu'au niveau des établissements, ce que l'on dit devoir être le traitement, c'est-à-dire des réunions spécifiques pour les parents non francophones, des propositions de stages de remise à niveau ou de sensibilisation à l'école, de nous assurer, dans des écoles dans lesquelles on n'a même pas le droit de rentrer, qu'au bout du compte, le traitement que l'on pense devoir être mis en œuvre le sera. Un biais très simple peut être à l'œuvre : vous demandez à une administration de faire cet effort particulier-là pour cette expérimentation-là, elle va faire quelque chose en moins par ailleurs. Typiquement, elle organisait déjà quelque chose, elle ne va plus l'organiser pour organiser ces réunions. On essaie bien sûr de parer à ce genre de défauts, mais plus il y a distance entre l'évaluateur et l'institution qui met en place, plus il y a de souci à se faire sur la qualité de ce que l'on obtient. Je rejoins ce que disait Olivier tout à l'heure : dans tous ces projets qui ne respectaient absolument pas les canons de l'expérimentation sociale, il y a quand même un dynamisme, une énergie au niveau local, un peu gâchée. Si on arrivait à mieux l'utiliser et la coordonner, elle serait à même de produire des évaluations plus rigoureuses que ce que l'on finit par négocier entre administration et évaluateurs, qui est pour l'instant un peu insatisfaisant à mes yeux.

Pierre CAHUC
Merci. Avez-vous des réactions ?

Marie-France TOMAS, direction du Travail, La Réunion
Sur l'île de la Réunion, nous sommes concernés par le problème des jeunes, notamment par l'expérimentation sur les jeunes diplômés dont on parlait tout à l'heure. Par rapport à ce que vous venez de dire, Monsieur, je pense qu'effectivement, par exemple, il se trouve que j'étais par hasard en métropole dans cette période et que je participe à cette séance de travail, à laquelle je n'ai pas du tout été conviée, pour laquelle je n'ai eu une information que tout à fait par hasard, alors que nous sommes, au niveau du ministère de l'Emploi et en particulier dans notre région, concernés par plusieurs dispositifs, dont celui-ci.

Par rapport à ce que disait Monsieur, aussi, sur la réaction des acteurs locaux et le lien entre ce que l'on peut vouloir, dans des instances et ensuite, ce que l'on demande dans les régions de faire, les énergies s'épuisent effectivement. On a un gros travail à faire. Je demande à cor et à cris à être formée, à ce que l'on me dise quand je dois piloter et mettre en place des comités de pilotage régionaux pour accompagner – parce que l'on n'a pas dit que le marché jeunes diplômés a été lancé par le ministère – ce projet. Je dois pouvoir relayer auprès des partenaires locaux ce qu'est réellement cette expérimentation, comment elle se passe, pour que, effectivement, il y ait bien une association. On peut comprendre l'aléatoire, si c'est bien expliqué et que les acteurs comprennent les effets qu'ils vont pouvoir obtenir, puisque, sur le

problème que posait Monsieur, sur la question de savoir si les gens sont informés, si on prend le cas des jeunes diplômés, celui qui est informé de la situation va dans la mission locale, dit qu'il aimerait bénéficier de la prestation, et on va lui dire : « Non, tu ne peux pas bénéficier de la prestation, puisque c'est aléatoire ». Il faut quand même bien pouvoir expliquer cela aux gens. C'est quand même très important. Il y a les expérimentations, et puis il y a ce que l'on doit évaluer dans le cadre des politiques publiques, de manière générale. Si l'on parle de l'accompagnement des jeunes, on finance très largement un dispositif qui s'appelle CIVIS, qui coûte cher ; on va avoir le contrat d'autonomie – à la Réunion, j'ai les trois. Comment tout cela se coordonne-t-il ? Comment faire pour expliquer que là il y a une expérimentation, là il n'y en a pas et que tout cela, de toute façon, s'évalue collectivement ?

Pour terminer, on parlait tout à l'heure de l'impact sur les bénéficiaires. Dans les résultats qu'a donnés l'intervenant précédent sur l'action sur les bénéficiaires du RMI, j'ai plus entendu les résultats obtenus sur les économies éventuellement faites que sur les bénéficiaires eux-mêmes ? Il ne faut pas se leurrer sur ce que contient réellement la politique publique mise en œuvre. Est-ce un vrai souci de résultat sur les bénéficiaires ou est-ce autre chose ? A ce moment-là, on évite beaucoup de malentendus.

Pierre CAHUC
Merci. Avez-vous des réactions à cette intervention ?

Eric MAURIN
A Grenoble, par exemple, je ne pense pas que cela soit tout à fait pareil : vous pouvez avoir le maire de Grenoble qui, lors d'un déjeuner, vous dit : « Cette expérimentation est passionnante, il faut absolument qu'on ait des résultats clairs et nets dessus, cela est très important, politiquement, etc., je ferai tout pour que cela se passe ainsi » et, de fait, l'administration grenobloise, pourtant plus petite que celle de l'Education nationale, c'est ensuite l'opérateur local avec lequel il faut dessiner ce programme d'évaluation et, même si c'est quelque chose qui est parti avec une forte impulsion locale, enracinée dans des questions qui se posent vraiment au niveau local, qui ne sont pas imaginées dans un salon parisien, etc., dès lors qu'il faut actionner toute la diversité d'une administration pour mettre en œuvre cette action publique, cela devient très compliqué. Pour plein de raisons : il y a des problèmes éthiques qui, à chaque étage de l'administration, se reposent. Mais pas simplement. Il y a des questions toutes bêtes : si cette action est bien, pourquoi la réserver à certains ? C'est la question la plus immédiate. Au fil de la négociation du projet, on aboutit à des choses qui ne sont pas nécessairement informatives.

Olivier NOBLECOURT
Je ne peux pas manquer de réagir à ce que vient de dire Eric Maurin, d'une part pour dire que je ne voudrais pas que demain, l'adjectif grenoblois soit identifié à celui de parisien et d'autre part qu'il y aurait deux lieux en France, qui seraient comme des microclimats, ce qui serait faux. Par ailleurs, Eric Maurin parle de la construction d'une expérimentation sociale qui a été reprise dans l'appel à projets, qui est une étude sur l'impact du mode de garde des enfants, sur le travail des parents et en particulier le travail des femmes. Il est vrai que lorsque nous avons

commencé à travailler avec l'Ecole d'économie de Paris et Eric Maurin sur ce sujet, nous avions une ambition extrêmement importante : pouvoir revisiter les critères d'admission en crèche et notamment travailler sur les enjeux de mixité sociale dès la petite enfance, pourquoi pas avec de la mixité contrainte : des déplacements de jeunes, etc. Petit à petit, de réunion en réunion, on est arrivé sur une problématique très ciblée, riche, féconde mais très limitée. Nous allons donc travailler pour réélargir la problématique. En effet, il est compliqué, pour des élus et des administrations, de porter ce type de démarche – surtout quand elles arrivent d'un universitaire aussi exigeant qu'Eric Maurin, puisque vous avez bien compris qu'avec ses six critères, on aurait sélectionné cinq projets pour toute la France ; on a réussi à lui en faire accepter trente-sept, donc on a bien négocié – mais ce sont quand même deux cultures assez éloignées et je pense que dans les enjeux essentiels qui sont ceux de ces deux journées de réflexion, vous avez parlé, Madame, de la formation, je crois que c'est un enjeu majeur pour les acteurs sur les territoires, les directeurs de structure, les responsables d'administration ; il y aura aussi et c'était un petit peu le sens de mon propos liminaire, la pédagogie par rapport aux élus. Parce que si l'on n'arrive pas à convaincre les élus là-dessus, on n'y arrivera jamais. C'est peut-être en effet la petite chance qu'on a à Grenoble. Mais on n'a pas de leçon à donner !

Béatrice SEDILLOT, DARES

Je voudrais peut-être apporter deux ou trois compléments d'information sur le marché des jeunes diplômés, plus dans le champ de l'intervention précédente. Je suis de la DARES. La DARES participe à l'évaluation du marché jeunes diplômés avec les équipes de chercheurs. Ce qui est apparu depuis hier est qu'il est important, dans ce type d'évaluation menée avec un protocole relativement novateur par tirage aléatoire, que la compréhension de la démarche soit bien appropriée par les acteurs locaux. C'est un marché avec des comités de pilotage régionaux, parce qu'il est important d'animer les choses au niveau de la région.

Les équipes qui suivent de près ce marché au sein de la DARES, ont eu vraiment le souci de participer le plus possible aux comités de pilotage régionaux, pour expliquer la démarche. La Réunion, c'est un peu loin et il est vrai que nous n'avons pas été amenés à nous déplacer. En revanche, effectivement, sur l'ensemble des autres régions, les évaluateurs à la DARES et dans les équipes de recherche qui suivent le marché ont vraiment eu le souci, régulièrement, de participer à l'ensemble des comités de pilotage régionaux pour à la fois expliquer la démarche, son caractère novateur, comprendre les difficultés des opérateurs privés. J'entends bien que quand on est à la Réunion, c'est un peu difficile. Vous pouvez avoir le sentiment, justifié, d'être un peu tenus à l'écart par rapport aux autres démarches, mais il nous semble vraiment très important qu'il y ait cette communication et cette appropriation. Elle n'est sans doute pas parfaite. Certaines choses sont peut-être difficiles, l'on pourrait les améliorer. Mais il y a eu, quand même, ce souci constant et que l'on poursuit, de vraiment participer aux comités de pilotage régionaux, en tant qu'évaluateurs, pour expliquer la démarche et également lever les interrogations par rapport aux difficultés que celle-ci peut engendrer.

Pierre CAHUC

C'est un point très important qui n'était pas abordé de manière systématique, à savoir la gouvernance de l'évaluation des politiques publiques. On a l'impression d'une véritable ébullition de l'initiative, beaucoup portée par le ministère du Travail. Bien évidemment, à terme, on espère que les services du ministère du Travail n'évalueront plus eux-mêmes les politiques qu'ils mettent en place. Ce souci d'indépendance est vraiment important. Il existe, je crois, deux problèmes : l'indépendance des évaluateurs – en dehors de la création d'une agence indépendante qui évalue, il y a peu d'espoir de mettre en place des évaluations sur des politiques importantes– ; il y un vrai danger d'instrumentalisation des résultats. Il faut aussi que ces agences d'évaluation, si elles sont créées, aient énormément de pouvoir, pour pouvoir mettre en place, avec des acteurs locaux, des politiques et des protocoles. En dehors de cela, d'un point de vue purement opérationnel, l'évaluation restera quelque chose de l'ordre de la théorie et qui aura du mal à rentrer en pratique et à accroître l'efficacité de la dépense publique, sachant qu'il existe, en outre, des problèmes méthodologiques dans le contenu même des évaluations, qui seront présentés tout à l'heure par François Bourguignon, après la pause.

Je vous confie à une pause café, pour ensuite l'écouter, ainsi qu'Esther, qui nous présentera les programmes européens.

Retour sommaire

Pour une évaluation des méthodes d'évaluation

Remarques sur la démarche expérimentale dans le domaine de l'emploi, du travail et de la formation professionnelle

Pierre CAHUC

On va reprendre cette réflexion sur les méthodes d'évaluation. Nous allons maintenant évaluer les méthodes d'évaluation, sous l'éclairage de François Bourguignon.

François BOURGUIGNON, Ecole d'économie de Paris

Je voudrais d'abord remercier la DARES d'avoir organisé cette conférence, qui touche un sujet crucial. Dans le domaine de l'évaluation des politiques publiques, nous nous trouvons actuellement à un tournant et il est bien que l'on prenne un peu de vitesse dans ce tournant. Outre l'échange de connaissances, c'est là je crois l'une des grandes utilités de cette conférence.

La plupart d'entre vous connaissent mieux que moi les méthodes expérimentales d'évaluation des politiques publiques en matière d'emploi. Je dois en fait confesser que j'ai très peu pratiqué ces méthodes. Je fais partie d'une génération dans laquelle cette culture, en tout cas en France, était peu développée, et dans laquelle il aurait été difficile de convaincre les décideurs et le public de l'utilité et du bien-fondé de lancer, dans le domaine social, des expériences sur des individus isolés ou des groupes d'individus tirés au sort. On a parlé à plusieurs reprises de cette réticence. Quoique de façon atténuée, elle est encore présente aujourd'hui, et plusieurs interventions au cours de cette conférence l'ont souligné.

Le type d'évaluation que l'on pratiquait était d'une nature assez différente. Elle était fondée sur des modèles économiques représentant les comportements des agents. La représentation de ces comportements était elle-même basée sur certaines hypothèses, parfois très fortes, de rationalité des agents et sur l'estimation économétrique des paramètres de leur comportement à partir de données d'enquête ou de données administratives. L'estimation tenait compte bien entendu des contraintes sous lesquelles les agents devaient prendre leur décision telles qu'imposée par les dispositifs en vigueur d'aide à l'emploi ou de soutien au revenu. L'"évaluation" d'une réforme de ce dispositif consistait à "simuler" avec ce modèle économétrique les changements de décision des agents résultant de cette réforme, c'est-à-dire de la modification des contraintes sous lesquelles ils opèrent. L'évaluation pouvait être *a priori*, montrant les résultats anticipés d'un jeu de réformes envisagées. Elle pouvait être aussi *a posteriori*, en comparant les comportements observés après réforme à ceux correspondant au dispositif antérieur, obtenus eux aussi par "simulation" du modèle estimé. Je qualifierai ici cette démarche d'approche structurelle non-expérimentale, par opposition à l'approche purement expérimentale, qui n'a pas nécessairement besoin de référence structurelle.

Il ne s'agit pas d'opposer ici un type de démarche contre l'autre. Je voudrais seulement défendre l'idée que l'approche structurelle et l'approche expérimentale sont souvent

complémentaires l'une de l'autre et que cette complémentarité n'est pas toujours bien exploitée. Aussi puissante et aussi *incontestable* – c'est le terme qui convient– que soit la méthode expérimentale, on a souvent besoin d'un peu plus que la pure expérimentation pour pouvoir conseiller de façon efficace les décideurs, et ce "plus" exige, d'une façon ou d'une autre, la mise en œuvre d'une approche structurelle.

Cela étant, je ne suis tout de même pas étranger à l'expérimentation. Et ce qui me donne peut-être droit à la parole aujourd'hui, c'est le fait que j'aie été un promoteur très actif de ces méthodes lorsque j'étais à la Banque mondiale. J'ose espérer que les innovations que j'ai introduites dans cette institution, et avant tout en essayant de modifier la façon dont le terme évaluation y était entendu, c'est-à-dire comme évaluation des processus plutôt que de l'impact de projets, porteront leurs fruits et se perpétueront tout en s'améliorant.

Une anecdote révélatrice sur l'absence d'une véritable culture d'évaluation d'impact à la Banque mondiale est la suivante. Quand j'ai été nommé économiste en chef, j'ai tout de suite été assailli par les services de communication, qui m'ont dit : « Il faut immédiatement que tu communiques sur "ton message". Quel est ce message ? Il doit tenir au plus en deux mots! ». J'ai répondu : « Equité et évaluation ». On m'a alors rétorqué : « Equité, c'est très bien, c'est impeccable ; évaluation, tu enlèves, ce n'est pas accrocheur, cela ne sert à rien ; les gens vont penser qu'en plus de tous ses autres problèmes, la Banque mondiale n'est pas capable de faire les choses au mieux. Il n'y a pas besoin d'évaluer ce qu'elle fait parce que tout ce qu'elle fait, c'est tellement bien ! ». Heureusement, j'ai résisté et je me suis accroché à ce second thème autant qu'au premier. J'ai en particulier continué à communiquer sur ce terme et je pense que cela a eu une certaine utilité.

Je crois que nous vivons une époque où la demande sociale pour l'évaluation des politiques publiques est très forte. Cela est vrai partout, dans les pays développés comme dans les pays en développement. C'est une demande pour la transparence des gouvernements, pour la reddition de comptes, pas seulement sur le budget mais sur les résultats des politiques publiques. C'est une demande d'efficacité de la dépense publique. C'est aussi la constatation d'un certain nombre d'échecs de politiques qui ont été présentées au départ comme de bonnes idées et qui, finalement, n'ont pas donné les résultats escomptés.

Mais, ce n'est pas tellement le "pourquoi évaluer" dont je voudrais parler que le "comment évaluer". Les méthodes d'évaluation sont nombreuses. La méthode expérimentale sur échantillons aléatoires a été à juste titre qualifiée d'"étalon-or" de ces méthodes. Certainement, l'idée de baser l'évaluation d'une politique sur la comparaison de deux échantillons tirés de façon aléatoire, l'un soumis à la politique à évaluer, et l'autre non est une démarche scientifiquement incontestable pourvu qu'elle puisse être appliquée rigoureusement. L'adaptation à l'économique et au social de ces techniques qui ont fait leurs preuves depuis longtemps dans les sciences justement dites expérimentales est un grand pas en avant. Cela étant, il existe des limitations dans leur mise en œuvre et donc dans les résultats qu'elles fournissent. C'est ce sur quoi je voudrais essayer d'insister dans ma présentation.

Comme tout bon exposé, en France, il est divisé en trois partie. Dans une première partie, je voudrais ré-examiner brièvement certains exemples presque historiques ou emblématiques de

la démarche expérimentale, et en montrer les limitations et la nécessité de l'articuler avec l'approche structurelle. Dans un deuxième temps, je voudrais insister sur un certain nombre de contraintes pratiques que l'on rencontre lorsque l'on essaie de mettre en œuvre ces démarches d'évaluation expérimentale. J'insisterai enfin sur un certain nombre de limitations intrinsèques qui me conduisent à penser que, dans la pratique, il ne faut pas se reposer exclusivement sur un type d'évaluation.

Le premier exemple emblématique de la démarche expérimentale qui vient à l'esprit dans le domaine de l'économie du travail est l'expérimentation de l'impôt négatif menée aux Etats-Unis à la fin des années 1960. Lyndon Johnson, président, avait lancé la fameuse "guerre contre la pauvreté". Se posait donc la question de savoir comment lutter contre la pauvreté sans diminuer les incitations des individus à l'effort et au travail. On connaissait l'idée de l'"impôt négatif" de Milton Friedman, qui combinait un impôt sur le revenu à peu près porportionnel pour l'ensemble de la population salariée à un transfert forfaitaire jouant le rôle d'un revenu minimal. On disposait aussi, déjà à cette époque, de quelques estimations de l'élasticité revenu ou de l'élasticité salaire de l'offre de travail. De façon structurelle et *a priori*, on pouvait donc déjà évaluer ce que pourrait être l'effet d'un programme d'impôt négatif sur l'offre de travail – en ne supposant aucune contrainte du côté de la demande. L'incertitude sur les élasticités revenu ou salaire de l'offre de travail était cependant énorme, d'où l'idée d'expérimenter pour voir ce qu'il en était effectivement.

La première expérimentation eut lieu dans le New Jersey entre 1968 et 1972 et a donné des résultats intéressants : la réaction de l'offre de travail à ce programme d'impôt négatif se révéla beaucoup plus faible que ce qui était attendu sur la base des quelques estimations disponibles à ce moment-là. Une réflexion approfondie montra cependant que l'un des problèmes de l'expérimentation conduite dans le New Jersey était qu'il existait déjà des programmes locaux d'assistance sociale (Welfare) souvent assez généreux qui n'avaient pas été neutralisés lors de l'expérimentation. Cette difficulté de neutraliser les dispositifs en place qui peuvent gêner l'expérimentation est, sous une forme ou une autre, assez commune. Elle a pour effet d'introduire du bruit dans l'expérimentation. Dans le cas du New Jersey, il se trouve que, du fait des programmes locaux déjà existants, le dispositif d'impôt négatif ne modifiait pas radicalement la situation des gens les plus pauvres et avait effectivement peu d'impact sur leur offre de travail[23].

L'expérience de l'impôt négatif fut alors reprise dans d'autres Etats en prenant soin d'éviter la présence de programmes concurrents. Les dernières expérimentations eurent lieu à Seattle et à Denver. Dans les deux cas, les effets de l'impôt négatif sur l'offre de travail se révélèrent beaucoup plus forts qu'au New Jersey. En moyenne, la baisse de l'offre de travail était de 9 % pour les hommes, et proche de 20 % pour les femmes. En même temps, il s'avéra que le programme qui avait été testé dans ces deux villes était très généreux et difficilement

[23] Pour une synthèse de ces expérimentations, voir Munnell, Alicia H., ed. *Lessons from the Income Maintenance Experiments.* 1987.

généralisable à l'ensemble du pays. Il était donc difficile de tirer beaucoup d'implications pratiques de cette expérimentation. Le programme testé n'était pas réaliste, et révélait en outre que l'impact sur l'offre de travail pouvait être assez substantiel.

Comme vous le savez, ce programme n'a finalement jamais été mis en œuvre en tant que tel. Après beaucoup d'années et beaucoup de débat, il a débouché sur le programme du *Earned Income Tax Credit*, qui est simplement un impôt négatif pour les gens qui travaillent plus qu'un certain nombre d'heures dans l'année. La partie fixe de l'impôt négatif, le transfert forfaitaire qui assure un revenu minimum même en l'absence d'un revenu du travail est constituée par les programmes d'assistance sociale existants.

Cette expérience de l'impôt négatif aux Etats-Unis est intéressante. D'une part, parce qu'il s'agissait d'une expérience pionnière en termes de méthode expérimentale. D'autre part, parce qu'elle fait apparaître un certain nombre de difficultés liées à l'expérimentation dans le domaine économique et social: l'importance du contexte et le besoin de concevoir les dispositifs expérimentés à la lumière des contraintes budgétaires et de paramètres clés comme ces élasticités revenu-salaire de l'offre de travail de façon à être le plus proche possible d'un programme réalisable et, dans un certain sens, optimal.

Je parlerai peu du second exemple historique parce qu'il a été exposé en détail dans une séance précédente de cette conférence. C'est peut-être l'exemple type de l'expérimentation réussie dans ce domaine des transferts sociaux et de l'incitation au travail. Il s'agit du Programme d'Auto-Suffisance, le *Self Sufficiency Program*, au Canada, qui a été expérimenté au début des années 1990 et qui a effectivement montré quel impact sur l'offre de travail des parents isolés – essentiellement des femmes seules vivant avec des enfants - pouvait avoir un impôt négatif se substituant à un programme de revenu garanti avec un taux implicite d'imposition de 100%[24].

Deux conclusions peuvent être retenues de cette expérience. De façon structurelle, quand on parle d'offre de travail et d'inactivité, il faut faire la distinction entre l'inactivité volontaire d'individus qui bénéficient d'un revenu garanti qu'ils jugent suffisant, et l'inactivité involontaire de ceux qui, de toute façon, ne sont pas employables, par exemple parce que leur productivité est inférieure au salaire minimum. Un premier enseignement de l'expérience canadienne, c'est que la première situation est largement dominante, étant donné le taux élevé de reprise d'emploi dû au programme d'impôt négatif.

Un deuxième enseignement de l'évaluation du Programme d'Auto-Suffisance provient du fait qu'elle a couvert une période très longue. Les personnes sélectionnées dans le groupe de contrôle et dans le groupe de traitement ont été suivies sur quatre ans et demi, la durée du programme lui-même étant limitée à trois ans. Le résultat intéressant est que les effets du programme se dissipaient très vite au cours du temps une fois le programme fini, c'est-à-dire

[24] Pour une synthèse de ce programme et de ses résultats voir par exemple: Reuben Ford
David Gyarmati, Kelly Foley, Doug Tattrie (2003), Can Work Incentives Pay for Themselves? Final Report on the Self-Sufficiency project for Welfare Applicants, SRDC

lorsque le subventionnement de l'activité salariée avait cessé. En référence à la théorie des pièges de pauvreté et des équilibres multiples, on aurait pu penser que le programme pouvait générer une modification non réversible du comportement des parents isolés. Il n'en a rien été. Au bout de quatre ans, quatre ans et demi, il n'y avait pratiquement pas de différence entre le groupe de contrôle et le groupe de traitement.

Le troisième exemple, c'est l'évaluation des politiques d'aides ou d'accompagnement à l'emploi et notamment le *Job Training Partnership Act*, qui a donné lieu aux Etats-Unis à une expérimentation de grande taille : vingt mille personnes dans le groupe de traitement, dix mille dans le groupe de contrôle. Ce qui est intéressant dans la généalogie de cette expérimentation, c'est le fait qu'elle ait été causée par une profonde ambiguïté sur les effets de ces programmes de requalification de la main d'œuvre. Cette ambiguïté provenait de l'utilisation de techniques structurelles quasi-expérimentales plutôt qu'expérimentales, pour estimer l'effet de ces programmes. Comme les personnes peuvent décider de par elles-mêmes d'entreprendre une re-qualification, celles qui suivent ces programmes ne constituent pas un échantillon aléatoire de l'ensemble de la population. Par conséquent, il faut corriger les différences observées entre ceux qui suivent ces programmes et les autres d'un biais évident de sélection afin de pouvoir juger de l'effet du programme.

Un ensemble de techniques peuvent être utilisées à cet effet. Elles sont la plupart du temps discutables dans la mesure où chaque technique repose sur une certaine hypothèse qui peut être mise en doute. Ces hypothèses ont fait l'objet d'une discussion intense au sein de la profession. En même temps, les résultats obtenus semblaient montrer que ces politiques de formation avaient peu d'impact sur les carrières et salaires des re-qualifiés. C'est la raison pour laquelle ont été lancées des expérimentations sur échantillons aléatoires de grande taille à la fin des années 80. Elles ont effectivement montré une différence entre groupes de traitement – les re-qualifiés – et groupes de contrôle.

Cette expérimentation a déjà été abordée dans une présentation précédente. Son problème majeur réside dans la façon dont elle a été organisée. Des personnes ont accès au programme si elles le veulent (groupe de "traitement"); d'autres n'y ont pas accès (groupe de contrôle). Oublions le problème d'équité derrière cette différenciation. La difficulté réside dans le fait que les personnes qui ont accès au programme ne sont pas forcées de le suivre. Certaines ne sont pas intéressées et préfèrent continuer comme avant. On peut toujours comparer le nombre d'heures de travail ou le salaire moyen des personnes qui ont été exposées au programme, même si elles ne l'ont pas suivi, au salaire et à la durée de travail de celles qui n'ont pas été exposées au programme. Mais cette comparaison ne reflète pas l'effet du programme lui-même, puisque, au sein du groupe ayant eu accès au programme, certaines personnes ne l'ont pas suivi. Le problème de sélection, c'est à dire de prédire qui suivra le programme et qui ne le suivra pas, au sein du groupe qui y a accès, reste entier. Dans les

nombreuses discussions qui ont eu lieu par la suite sur les résultats de cette évaluation, ce problème, clairement de nature structurelle, est toujours présent[25].

Le quatrième et dernier exemple est celui du programme mexicain PROGRESA. C'est un programme de transfert de revenu conditionnel en faveur des familles pauvres. La conditionnalité réside en ce que les familles qui bénéficient du programme doivent envoyer leurs enfants à l'école et leur faire suivre deux examens médicaux par an. Cette initiative est intéressante dans la mesure où elle montre le développement de politiques de redistribution explicite des revenus dans les économies émergentes. Mais tout aussi intéressante est l'économie politique de ce programme – le transfert de revenu se justifiant après d'une certaine opinion par l'accumulation de capital humain permise par la conditionnalité – et plus encore les circonstances qui ont conduit à son évaluation sur échantillons aléatoires de groupes de traitement et de contrôle.

Il se trouve que, bien avant que j'ai une position permanente à la Banque mondiale, lors d'une visite à cet organisme j'ai été invité à une réunion avec les responsables mexicains de PROGRESA. Ceux-ci nous ont dit: « On n'a pas suffisamment de moyens budgétaires pour que le programme puisse s'appliquer à toutes les localités rurales d'un seul coup. On va donc l'appliquer de façon progressive à l'ensemble du pays, par tiers successifs». Chercheurs et fonctionnaires de la Banque leur ont alors suggéré: « Pourquoi ne profitez-vous pas de cela pour conduire une véritable expérimentation ? Choisissez votre premier tiers au hasard, sélectionnez-y un sous-groupe de localités, choisissez aléatoirement un groupe de localités de contrôles dans les deux autres tiers, et vous disposerez d'une expérimentation rigoureuse des effets du programme sur les variables que vous jugez pertinentes ». La Banque Interaméricaine de Développement était elle-même intéressée et a offert au Mexique de couvrir les frais de l'évaluation, pourvu qu'elle soit encadrée par des experts reconnus.

Cette évaluation de PROGRESA a eu un impact absolument considérable au Mexique. L'évaluation est même entrée dans la Constitution dont une clause demande que « … tout programme social doit être évalué ». Pour montrer l'importance politique qu'a prise l'évaluation, l'anecdote suivante mérite d'être rappelée. A tout changement (tous les six ans) de président au Mexique, il est traditionnel de modifier la plupart des programme sociaux, enjeux importants du débat électoral. Même si un programme social fonctionne bien, le nouveau président cherchant à affirmer l'originalité de son gouvernement a tendance à abolir ou réformer plus ou moins drastiquement les programmes mis en place par ses prédécesseurs. Grâce à son évaluation rigoureuse et très positive, cela n'a pas été le cas de PROGRESA. « Cela fonctionne très bien, pourquoi voulez-vous changer ce programme? » ont dit les électeurs lorsque le Président Fox a assumé le pouvoir. Le résultat est qu'il a décidé d'amplifier et élargir la couverture du programme et d'en changer le nom. De PROGRESA le programme s'est transformé en OPPORTUNIDADES. Mais l'essentiel est resté en place,

[25] Pour des vues formelle et informelle de ces questions voir James Heckman & Carolyn Heinrich & Jeffrey Smith, 2002. "The performance of performance standards", NBER Working Papers 9002, National Bureau of Economic Research, Inc et J. Donohue, Shortchanging the Workforce. The Job Training Partnership Act and the Overselling of Privatized Training, Economic Policy Inst., Washington, DC.

inchangé, ce qui était une première au Mexique. L'aspect politique derrière l'évaluation et l'expérimentation est donc important.

Et puis, l'évaluation de PROGRESA a eu des répercussions considérables sur le reste du monde. On compte aujourd'hui une trentaine de programmes du type "transfert conditionnel" mis en œuvre dans les pays en développement. Parmi ceux-ci, plus des deux tiers font l'objet d'une évaluation expérimentale sur échantillons aléatoires[26].

Je voudrais maintenant dire un mot sur les contraintes à l'application des méthodes expérimentales. Il est assez intéressant de voir que la culture de ces expérimentations est plutôt apparue du côté des Etats-Unis et du Canada. On assiste depuis quelques années à un développement rapide de l'utilisation de ces méthodes d'évaluation dans les économies en développement, sous l'influence d'organismes comme les Banques internationales de développement et aussi d'universitaires bien représentés ici (Esther Duflo, Abhijit Banerjee et d'autres)… Bizarrement, cette approche expérimentale des politiques d'emploi ou d'assistance sociale est moins répandue en Europe. Je lisais il y a peu une étude comparant les programmes d'accompagnement de l'emploi en Europe. Cent trente-sept évaluations étaient comparées dans différents pays européens. Parmi les cent trente-sept, seules neuf évaluations étaient basées sur une approche expérimentale avec échantillonnage aléatoire[27].

Le fait que ces neuf expérimentations soient relativement récentes indique peut-être que nous nous trouvons à un tournant. Je l'espère. A la conférence de Grenoble d'octobre dernier sur "l'expérimentation sociale", organisée par le Haut Commissariat aux Solidarités Actives, j'ai effectivement eu l'occasion de dire que nous étions en train de vivre une révolution culturelle. C'est vrai que la discussion avec le public a montré ensuite que, parmi toutes les expérimentations proposées, l'évaluation dépassait rarement la comparaison avant/après. Mais, tout de même, la nécessité de disposer d'un "contrefactuel" pour évaluer un programme est une idée qui fait son chemin, même si l'on est encore loin d'une utilisation systématique de l'échantillonnage aléatoire.

La principale résistance à la pratique de la méthode expérimentale sur échantillons aléatoires est probablement le principe d'équité. Le fait que, pour un certain temps, certains individus puissent "bénéficier" d'un programme dont sont privés les autres, ou que certains soient "soumis" à un programme qui épargne les autres, choque. Est choquante aussi dans l'opinion l'idée que des être humains puissent être utilisés comme des sortes de cobaye – même si c'est pratique courante dans le domaine pharmaceutique. Cela veut dire que l'on ne peut finalement évaluer de façon expérimentale que des programmes plutôt marginaux qui ne modifient pas trop radicalement l'environnement des individus.

---

[26] On the evalutation of Progresa, see Skoufias, E. and B. McClafferty (2001), Is Progresa working? Summary of the results of an evaluation by IFPRI, Discussion Paper N° 118, IFPRI, Washington, DC. For a more general evaluation of conditional cash transfer programs in the world see the forthcoming, Policy Research Report on Conditional Cash Transfers in World Bank.

[27] See "Study on the effectiveness of ALMPs", Report for DG Employment, Social Affairs, Equal Opportunities, European Commission, 2005; Lead Author: J. Kluve.

L'idée que l'échantillonnage aléatoire est en soi "équitable" dans la mesure où, a priori, tous les individus sont égaux devant la probabilité d'être tiré au sort ne passe pas bien dans les esprits. Comme je le disais tout à l'heure, il s'agit d'une affaire de culture. Il faut probablement laisser le temps faire son œuvre tout en démontrant par la force de l'exemple tout le profit social que l'on peut tirer de l'approche expérimentale. Cela étant, les problèmes d'éthique pure existent et l'on ne saurait passer outre même si la frontière entre ce qui est acceptable et ce qui ne l'est pas est parfois assez floue.

Un cas qui paraît définitivement au-delà de l'acceptable, quoique pour des raisons qu'il serait intéressant d'approfondir, est le suivant. A la Banque mondiale, quelqu'un est venu me voir un jour, pour me proposer une expérimentation dans le domaine de la diminution des risques de transmission du SIDA en Afrique australe. Il s'agissait d'expérimenter un programme de transfert conditionnel, basé sur … la séro-positivité des jeunes filles! Pratiquement, le programme consistait en un transfert monétaire aux familles comportant des jeunes filles ayant passé la puberté. Les jeunes filles seraient testées périodiquement pour le HIV, et les transferts de revenu s'interrompraient dès que le test serait positif. Dur! Cette proposition était clairement au-delà de l'acceptable. Du reste, elle aurait probablement été rejetée par les organismes chargés dans la plupart des pays de vérifier la conformité morale des expérimentations sociales.

L'économie politique de l'évaluation des politiques publiques n'est pas évidente non plus. J'ai souvent vu des gouvernements de pays en développement refuser qu'un programme qu'ils étaient sur le point de lancer ou sur lequel ils s'interrogeaient soit évalué par expérimentation, même si celle-ci était complètement subventionnée par un organisme international. La réaction du ministre concerné, du premier ministre ou du président était simplement: « Pourquoi voulez-vous que j'évalue ce programme et que je courre le risque d'avoir à révéler à l'opinion publique qu'il n'a pas atteint ses objectifs? C'est ma carrière politique qui est en jeu. Je préfère ne rien savoir pour ne rien avoir à révéler à l'opinion». Cette réaction est bien compréhensible. Elle montre que le tournant vers la culture de l'évaluation expérimentale doit s'accompagner d'un progrès vers la transparence des gouvernements et la reddition de comptes.

Il y a aussi un problème d'horizon temporel. L'évaluation d'un programme ou d'une expérience pilote peut ne pas être positive mais elle apporte de l'information et peut permettre dans un second temps la mise au point d'un programme plus efficace. Malheureusement cet argument n'est pas très convaincant auprès des dirigeants d'un pays. La plupart du temps, les résultats de l'évaluation ne seront disponibles que lorsque l'équipe en place aura quitté le pouvoir. Cette équipe n'aura plus beaucoup d'intérêt pour ces résultats et le gouvernement en place pourra ne pas être intéressé ou simplement opposé au programme testé. Si l'évaluation est négative, le programme lui-même passera aux oubliettes.

Tous ces aspects sont des aspects importants, qui expliquent qu'il y ait une certaine résistance à la diffusion des techniques d'évaluation expérimentale. Mais il existe par ailleurs des limitations plus objectives à la mise en œuvre de ces techniques. S'il y a des effets de contamination entre les groupes de traitement et de contrôle, alors l'évaluation est

nécessairement biaisée. Il y a aussi la question de l'échelle. Un programme peut donner certains résultats lorsqu'il est appliqué à un petit nombre d'individus ou de localités isolées. Mais d'autres effets, en particulier des effets d'équilibre général, risquent d'apparaître lorsque le programme sera étendu à l'ensemble de la population.

Par exemple, un cas qui est devenu une référence dans le domaine de l'évaluation expérimentale concerne les bons d'éducation expérimentés en Colombie pour remédier à l'engorgement des écoles secondaires publiques et profiter de la capacité excédentaire des écoles privées. Les familles se situant dans le tiers inférieur de la distribution des niveaux de vie se sont vues distribuer des bons leur permettant d'inscrire leurs enfants dans le système privé. Mais comme la capacité de ce dernier était limitée par rapport à la demande potentielle, ces bons furent distribués aléatoirement (loteries), permettant ainsi une expérimentation parfaite des effets d'ouvrir à des familles modestes l'accès à l'enseignement secondaire dans de bonnes conditions. Les analystes purent ainsi mettre en évidence que, dans les familles ayant reçu un bon, les enfants réussissaient mieux leurs examens, redoublaient moins, et restaient plus longtemps dans le système scolaire[28]. Ces conclusions sont importantes. Cependant, elles ne permettent pas de déterminer quels seraient les effets d'augmenter le nombre de bons, ou même, puisque les résultats de l'évaluation sont positifs, d'étendre la distribution de bons à l'ensemble des familles modestes. Que se passerait-il alors? Face à cette augmentation de la demande s'adressant à elles, les écoles privées augmenteraient probablement leurs droits d'inscription ou admettraient plus d'élèves au risque de diminuer la qualité de leur enseignement. Quel serait l'effet global? L'expérimentation naturelle conduite en Colombie ne nous dit pas grand-chose sur cette question. Elle nous révèle seulement des paramètres importants de la réaction micro-économique à la distribution de bons.

La réplicabilité en termes d'échelle est tout aussi importante que la réplicabilité en termes de contexte dont on parlait tout à l'heure.

Je n'ai plus beaucoup de temps. Je voudrais terminer cette présentation en approfondissant ce que j'appellerais les limitations intrinsèques des techniques d'évaluation expérimentale. Je viens de parler des problèmes d'équilibre général. Ils font partie de ces limitations. Il en existe deux types qui ne sont pas à traiter de la même façon. Il y a d'abord le phénomène d'équilibre général local. A l'intérieur de collectivités territoriales dans des pays en développement on crée ou on améliore des infrastructures: électricité, adduction d'eau, assainissement, transport, télécommunication, etc. L'expérimentation et la comparaison à des zones témoin permet de mesurer l'impact de programmes de ce type sur l'emploi, l'assistance scolaire, l'emploi du temps des individus, le mode de décision publique (influencé par la plus grande disponibilité

[28] See J. Angrist, E. Bettinger, E. Bloom, E. King et M. Kremer, "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *American Economic Review*, 92(5), December 2002, 1535-1558. Voir aussi J. Angrist, E. Bettinger, Eric et M. Kremer, "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia", *American Economic Review*, 2006, 96(3), 847-862

des femmes dont la productivité domestique a augmenté), et sur d'autres caractéristiques de la collectivité et de ses habitants. Tous ces effets résultent en fait de mécanismes d'équilibre général provenant de l'interaction entre les individus qui composent la collectivité. Ils sont extraordinairement difficiles à analyser avec des modèles structurels mais la méthode expérimentale permet de les saisir d'un seul coup. Elle ne nous dit pas grand-chose cependant sur la nature de ces effets, la logique selon laquelle ils sont mis en œuvre et la façon dont ils dépendent du contexte, information qui serait nécessaire afin d'optimiser ces programmes.

Mais les effets d'équilibre général peuvent aussi se produire à un niveau plus élevé, celui d'une région ou de la nation, comme dans le cas des bons d'éducation décrit plus haut. Là, l'expérimentation sur des groupes ou des localités représentant une partie infime de la population n'est pas d'un grand secours. Dans l'exemple des bons d'éducation, on a les effets qui ont été signalés, mais on peut aussi penser à d'autres, encore plus délicats à cerner. Par exemple, l'augmentation du nombre de travailleurs éduqués peut provoquer à terme une diminution de leur rémunération relative, ce qui peut dans le futur diminuer l'incitation à la scolarisation[29]. Ce sont des phénomènes qu'il convient de prendre en compte. L'expérimentation ne donne qu'une information sur certains des paramètres pouvant représenter les comportements micro-économiques. Ceux-ci doivent eux-mêmes éventuellement s'intégrer à une modélisation plus complexe rendant compte des effets d'équilibre général de façon plus satisfaisante.

L'effet du temps est aussi une limitation intrinsèque importante, particulièrement lorsque l'on parle d'expérimentations qui ont lieu au niveau local. On a constaté, dans beaucoup de cas, qu'avec le temps, une expérimentation qui avait lieu dans une certaine localité tendait à se transmettre à d'autres localités ou groupes de traitement. Cela n'est pas immédiat. Mais au bout de trois à cinq ans, au moment où, justement, l'on voudrait que l'expérimentation révèle les effets de long terme d'une politique ou d'un programme, il est difficile d'identifier ces effets du fait de cette contamination. C'est un problème important.

Une autre limitation majeure est que l'expérimentation de programmes particuliers ne permet pas d'étudier un nombre important de variétés du programme lui-même. Dans le programme d'impôt négatif que j'ai cité au tout début de cette présentation, la première expérimentation comprenait des taux marginaux d'imposition qui n'étaient pas les mêmes pour différentes personnes. Les auteurs de l'expérimentation avaient donc eu le souci d'expérimenter, non pas sur un programme, mais sur une gamme de programmes. Evidemment, cette gamme est nécessairement limitée si l'on veut une certaine précision des estimations avec une taille raisonnable d'échantillon. Si l'on voulait changer complètement le format du programme, on serait donc bien incapable de tirer de l'information de l'expérimentation. Je parlais de PROGESA tout à l'heure, qui transfère des revenus aux familles pour qu'elles envoient leurs enfants à l'école, avec un peu plus pour les filles que pour les garçons. Ne serait-il pas

[29] Pour un modèle complet de ces effets voir J. Heckman, L. Lochner et C. Taber, "Expalining rising wage inequality: explanations with a dynamic general equilibrium model of labor earnings with heterogeneous agents", Review of Economic Dynamics, 1998, 1(1), pages 1-58, ainsi que des m^mes auteurs,"General-Equilibrium Treatment Effects: A Study of Tuition Policy," American Economic Review, 1998, 88(2), pages 381-86.

intéressant de voir quelle est la sensibilité des résultats à l'importance du transfert ou à l'écart entre filles et garçons ? Si cela n'a pas été intégré dès le début dans l'expérimentation, on ne peut répondre à cette question.

Il faut en effet garder à l'esprit que lorsque l'on évalue une politique, on le fait non seulement pour savoir si cette politique, dans le format initialement décidé, produit les résultats escomptés ou non, mais aussi et surtout pour être capables d'en améliorer le format. Ce qui est demandé aux analystes, c'est essentiellement de suggérer des solutions pour réduire les coûts et augmenter les bénéfices du programme. Il faut avoir cette demande en tête quand on construit une expérimentation. Mais ceci requiert probablement la combinaison d'une approche structurelle et d'une approche strictement expérimentale.

Finalement, il y a l'influence du contexte. J'en ai parlé tout à l'heure avec l'exemple de l'impôt négatif et la comparaison entre le New Jersey, Seattle et Denver. Une politique donne certains résultats dans un endroit donné. Mais ceci ne signifie pas que le même programme dans un autre environnement va nécessairement donner les mêmes résultats. C'est quelque chose qu'il importe absolument de prendre en compte. On ne veut pas dire aux décideurs du pays X: « Regardez, ce programme marche très bien, il a donné d'excellents résultats dans le pays Y ». On veut dire aux décideurs : « Ce programme a été mis en œuvre dans divers pays et dans diverses circonstances. A peu près dans tous les cas, il a généré tels résultats. On peut donc penser que, si vous l'adoptez, vous obtiendrez des résultats comparables.». Cet exercice de synthèse de plusieurs évaluations a pour nom la "méta-évaluation". Elle consiste à comparer des évaluations du même programme dans des contextes différents, et éventuellement avec des formats différents, pour voir si les résultats obtenus sont différents ou au contraire analogues. Elle consiste également à identifier l'origine de ces différences ou analogies[30].

Laissez-moi simplement conclure sur ce que je crois être une nécessaire combinaison des approches d'évaluation. Bien que j'aie intitulé ma présentation « Pour une évaluation des méthodes d'évaluation », je ne vais pas dire qu'une méthode domine les autres. Je crois très honnêtement que les arguments que je viens d'évoquer, la forte demande qui existe dans certains pays et certains organismes de la part des décideurs, de la société elle-même, les instruments qui sont disponibles aujourd'hui appellent à une plus grande utilisation de la méthode expérimentale. L'on ne peut que se réjouir du développement de l'utilisation de cette méthode qui a lieu à l'heure actuelle. Il ne faudrait pas oublier pour autant la démarche inductive, analytique, qui est au total la seule qui permette de comprendre pourquoi on obtient certains résultats à la suite d'une réforme ou d'un nouveau programme et, éventuellement, comment ces résultats pourraient être modifiés en modifiant le format des programmes. Du point de vue du décideur, cette information est capitale.

---

[30] Voir par exemple, K. Ashworth, A. Cebulla, D. Greenberg and R. Walker, Meta-Evaluation: Discovering What Works Best in Welfare Provision Evaluation, 2004, 10(2), 193-216.

Je vous remercie.

(Applaudissements)

Pierre CAHUC
Je crois qu'il y aura quelques questions.

Judith GUERON
Vous avez tiré des conclusions très intéressantes de toutes ces expériences : l'importance du contexte, il faut suivre les gens pendant longtemps, la participation est très inférieure à cent personnes, qu'il y a les *general equilibrium effects.* Vous avez appris cela parce qu'il y avait des expériences. Ce ne sont pas des limitations mais des résultats. Si c'était une mauvaise évaluation, vous n'auriez rien appris et surtout pas ces leçons très importantes. Je trouve que les leçons et même le fait qu'il faille faire une *meta-analysis*, le fait qu'une expérience ne réponde pas à toutes les questions justifient de faire plus d'expériences. Vous avez bien vu les leçons de ces expériences. Si on avait fait autrement, on n'aurait pas fini par avoir toutes ces leçons dont vous avez montré l'importance.

François BOURGUIGNON
Je suis tout à fait d'accord. Il y a probablement une ambiguïté dans la façon dont j'ai utilisé le mot "limitation". Il ne s'agit pas d'une limitation dans le sens accumulation de la connaissance. Vous avez raison, la disponibilité d'un nombre croissant d'expérimentations bien faites permet d'augmenter notre stock de connaissances sur les phénomènes que l'on cherche à étudier. Je parle de "limitation" dans la façon font on peut interpréter les résultats d'une expérimentation particulière. Il faut se garder de prendre ces résultats au pied de la lettre, en concluant simplement "c'est bien" ou "ce n'est pas bien", "on applique" ou "on n'applique pas", etc. Il faut aller plus loin que cela. C'est dans ce sens que je dis qu'il y a une limitation à la méthode expérimentale. Mais je conviens que c'est peut-être plus dans l'utilisation qui peut en être faite que dans son principe même. Encore une fois, vous avez raison d'insister sur l'accumulation de connaissances. Cela est capital.

Jeffrey KLING
Talking about the Voucher program Josh Angrist had evaluated, there is, I don't think, any sense in what you could have actual evidence regarding what it would be like to have a nationwide policy until you had a nationwide policy; so - in terms of generating some evidence that would tell you which direction you would want to go in - it seems like the best, first thing, as far as I could tell, would be to do an experiment on a moderate scale that would tell you in which direction you might want to go, and then you might have some uncertainty about what the ultimate, large-scale impact would be. But that's still a very important part of figuring out what to do; so I was wondering if you would agree with that, or what your thinking is?

François BOURGUIGNON

I completely agree with this way of proceeding. Almost by definition, it is impossible to do randomized evaluation at the national level taking so as to include all macro general equilibrium effects. Thus, we have to start from partial experiments, which will tell us about how individual agents may modify their behaviour as a reaction to a given program, an information that should permit to derive key parameters in the behavior of some agents. Then what we have to do is try to incorporate those parameters into a representation of the whole economic system, or at least the minimal representation that you need in order to take into account all feedback effects. And really what I would cite as an example of this kind of approach would be the work done by James Heckman and other people recently where they are looking at training programs, but they try to take into account the fact that by generating more trainees you might be modifying the structure of wages and you might reduce the number of candidates for training.

These feedback effects may be something really important, and we cannot simply ignore them. It's not always the case—in some cases, you can do it because the program you're evaluating has little impact at the macro level. Reflecting on this at the time of scaling up an experimentation is absolutely crucial. Having said this, I agree with you that a partial equilibrium evaluation of an small-scale experimentation is providing a huge amount of information about what would happen at full scale.

Pierre CAHUC

Merci. Il est l'heure d'écouter Esther Duflo, qui va nous présenter le programme J-PAL.

Retour sommaire

Présentation du J-PAL Europe

Esther DUFLO, MIT

Merci beaucoup. Je voudrais encore une fois remercier Antoine Magnier, Béatrice Sédillot, Dominique, Julie et la DARES en général d'organiser cette conférence. Ces deux jours ont vraiment été stimulants, intéressants et encourageants. Je voudrais aussi remercier les intervenants d'avoir fait le trajet et vous tous d'être ici. C'est vraiment une nouvelle occasion de remercier la DARES de nous donner l'occasion aujourd'hui, de présenter officiellement, de lancer devant vous J-PAL Europe, qui est un centre qui appartient à l'Ecole d'économie de Paris – ce qui me permet de remercier aussi François Bourguignon – qui a eu la générosité d'accueillir J-PAL Europe comme un centre de recherches qui en fait administrativement et physiquement partie.

J-PAL Europe est une filiale ou un centre affilié à J-PAL Monde, qui est lui-même un centre de recherche du MIT. Il existe trois J-PAL dans le monde : J-PAL à Cambridge, au MIT, J-PAL Europe et J-PAL Asie du Sud en Inde. J-PAL signifie Poverty Action Lab, laboratoire d'action de lutte contre la pauvreté, avec le J pour Jameel, qui est notre sponsor. Je voulais commencer par laisser la parole à Abhijit Banerjee pour décrire les objectifs de J-PAL. Ensuite, je vais décrire les actions et en particulier celles de J-PAL Europe jusqu'à maintenant et dans le futur.

Abhijit BANERJEE, MIT

Thank you. It's been wonderful sitting through these two days of talks. Sometimes the French has been a little bit beyond me, but the translations have been excellent, so I actually understood most of what happened. The talks were great; and the fact that, even at the end of two days, we have so many of us here is also encouraging. I first wanted to say "thank you" to all of you for joining us. For us, it is an incredibly important occasion. Esther asked me to take us back a little bit in time, to when we founded "J-PAL World," as she calls it (which we used to call "J-PAL"), and to "what made us do it?" Maybe that's helpful in thinking about what J-PAL Europe might achieve in the next few years.

I apologize for speaking in English. You don't want to hear my French.

Starting sometime in the mid-90's, there were a bunch of developing countries with randomizing evaluations being done by development economists (often people who were at Harvard or MIT). There was a group of us who were doing them, and the results have been interesting. There's been lots of debate on whether this is a technique that could be used for other things as well. There's a huge world criticism that this is just one new thing, and that it's not going to go anywhere, but things seem to be going well. We seem to be doing more and more projects.

I think there was a critical time when we realized what we were doing. We were researchers. We weren't really thinking about "impact" in the sense that policy people often think of it. We were thinking about our papers getting published in journals. They were getting published in journals, and we were happy. And it was going fine, but there was a point where I think we

were, in a sense, invited by MIT (and this was very prescient of them) to start something that would somehow be a bit bigger than our own research. We were basically invited and given a little bit of money by MIT, who said, "Look, you guys are doing research. Do you think you want to create something that will be more than the individual research that you're doing? That's what forced us to think, "What are we missing?" In retrospect, it shows—as academics should think about this issue—that we hadn't thought about it.

There were two basic activities where we were clearly not doing nearly as much as we should have been. The first few years of randomizing evaluations (in the developing country context, where there are various institutional constraints that are sometimes quite distinct) had institutional advantages. Weak states often have the advantage that, for example, you can do things "under the radar screens;" so there was a fair amount—I don't know that it was all "disadvantage." One of the things we realized in the process is that we had evolved a body of, I would say, "tricks" for carrying out evaluations in contexts where you might imagine that, given all the constraints, nothing could be done. I think we had learned, in the process of doing these evaluations, that you could "do this design and combine it with that." There was one level that was just "how to create a design that's politically acceptable?" It was very much a range of options. There are various constituencies, and each may feel that a certain type of design may be more desirable. How do you balance those sorts of questions about evaluation practice?

There is also a set of things we had learned about how to articulate both the need for evaluation and what it's supposed to achieve—how to persuade people that they would get something out of it. I think that was very important, too. I suspect most of you who are here today feel that evaluation is important, whether randomized or not. But in fact, in the world—even now—you would hear people say, "Well, we know the answer to this question. We know what this education policy does. Why bother with evaluations?" There was a range of learnings we had of how to articulate why there is a need to do evaluations. I think that, because we were doing it every day—talking to organizations (and not just governments, private organizations, and N.G.O.'s, but all of those we were talking to) and explaining to them why we wanted to run an evaluation—what we hadn't recognized is that all that constituted a body of knowledge, a set of communication skills that were actually useful. And they were useful enough both to codify and to offer to people as a way of making it possible for them to do evaluations.

That was one thing we hadn't recognized: that even the elementary, relatively low-tech things we were doing were not obvious to people. One of the things I think we learned was that it was possible to codify these things and turn them into a teaching program; so that we could bring people in who were interested, but who were maybe afraid of "how do you do it in this context? In the country where I live, the villages are very far away, and therefore how do you do a survey in a context where people have to travel long distances?" There were little tricks like that (like "How do you do these things in efficient ways?") that we had figured out, which were useful things for somebody who's trying to do it. "When my minister asks me, 'Why are you doing this evaluation this way,' how do I respond to it? When the Press asks

me, 'Why do you want to do an evaluation,' how do I respond to it?" Answering all of those questions was a body of knowledge, and we came to articulate that body of knowledge.

I think one of the things we do these days is that we run lots and lots of courses (five-day courses for people who want to do random evaluations). This year we will be running at least four (and possibly more) of these courses in different parts of the world. So there will be one in Boston, one in Paris, one in India, one in Indonesia, and, I think, several more. And these are ways in which people can come in and learn many of these tricks of how to do the evaluation.

Where I think there was a more obvious need was that you don't get very much credit in the policy community for results published in a journal. That seems trivial, except that I hadn't thought about it very much. It was "The result is out there, so why aren't people tapping it?" And the reason is that results need to be interpreted, translated, stated in simple terms, and described in terms of the underlying conditions. "This is where the experiment was done, this is why this experiment is relevant for you, this is how other experiments that have been done were different, this is (what François said, very importantly) the 'meta-knowledge' we have." All of those things—putting them together in different forms—was another activity we weren't doing. Why were we not doing them? Because, as an academic, once you get your paper published, you don't really have an incentive to do more. We were "done." One realization we had, which was very important, was that that whole piece of it—of just propagating the knowledge, codifying it, and making it accessible was important.

The third and last thing we increasingly realized in the process of setting up J-PAL was that a large part of doing experiments is about feeling confident, and a large part about feeling confident is being able to talk to somebody who says, "Well, I did this experiment, and it's actually not so hard. This is how you do it. Here are a few designs." There is now, in a lot of places (I see it in France, but also in a lot of other places), an interest in doing experiments. The constraint is often that it looks completely alien. "What do you mean, 'randomized?' Is it okay if I just pick these villages? They look random. Why don't I randomly pick a few?" When people say "randomly," they often don't mean "randomized." Almost always, they don't mean "randomized." They mean "choose in an unsystematic way," which is very different from randomized. There are a lot of very basic things that make people comfortable. "No, no—that's not how you randomize. This is how you randomize. But how can I do it? Well, there's a state program that does it." It's very basic things like that, but when you talk to people and you say all those things, it goes from looking like something that people from Mars do to something that you can do yourself, and that comforts.

Part of what we do as an institution is to be the place for people to come and say, "We have an interesting program, and we've been told that we should do randomized evaluation, but we don't actually know what that means. That sounds really scary. How do we do it?" What we often do is talk to them, and after two hours, the demystification has happened. Then, what is left is just negotiation of a process. That is an enormous amount of value you can generate very quickly. In that sort of way, we are available. I will say more about specifically what we do and in exactly what form we are available. We are available as a resource to the world—

partly to provide these shared learnings we have, partly to propagate results (both of projects we have done and projects other people have done), which are reliable experiments, where the results are interesting to codify and to propagate, and to provide this "hand-holding"—which I think is very important in getting people to do randomized evaluations. In the process of the last few years, the J-PAL World has become reasonably successful on delivering all of these things, and we are optimistically hoping that J-PAL Europe will play an equivalent role in Europe. To elaborate on that, I'm going to throw it back to Esther.

Esther DUFLO

J-PAL est donc un réseau de chercheurs à trois pieds : un au MIT, un à l'Ecole d'économie de Paris, un à IFMR à Madras en Inde. Au total, ce réseau comprend une trentaine de chercheurs. Il facilite les évaluations expérimentales, toujours en partenariat avec des acteurs locaux, qui peuvent être des gouvernements, des ONG, des territoires locaux, un peu toutes les combinaisons dont on a parlé plus tôt cet après-midi. On a une soixantaine de projets dans vingt-deux pays. On s'occupe aussi de formation, comme l'a dit Abhijit, avec des cours de cinq jours destinés à des acteurs, souvent des responsables d'ONG, de gouvernements et des personnes de différentes administrations, ainsi que des étudiants – chacun pour un tiers environ. On s'occupe enfin de dissémination des résultats, à la fois sous forme de préparation de matériel et de communication directe avec différentes personnes, dans différents ministères, à la fois dans les pays riches et dans les pays pauvres.

Pour l'Europe, l'impulsion est venue d'une combinaison de souhaits de travailler avec l'Agence française de développement, la DARES et l'Ecole d'économie de Paris. Au démarrage, l'Agence française du développement était désireuse d'essayer une évaluation aléatoire d'un de ses projets. On travaille avec eux au Maroc, avec Bruno Crépon et William Pariente. Cela a été notre premier projet basé vraiment en France, avec l'influence essentielle de Bruno. Cela a été suivi par de plus en plus de travail dont vous avez entendu parler par Bruno – ANPE/Unedic, jeunes diplômés, RMI, les efforts avec Martin Hirsch, avant même qu'il ne soit Haut Commissaire, et avec la Ville de Grenoble. Tout cela a fini par créer une masse critique qui rendait nécessaire de commencer J-PAL Europe.

J-PAL Europe a vocation à servir d'intermédiaire à des chercheurs désireux de travailler sur l'évaluation des programmes et des implémenteurs, qu'ils soient des gouvernements étrangers, locaux ou des acteurs de terrain désireux d'évaluer une de leurs politiques. Elisabeth Beasley est la chef de J-PAL Europe, William Pariente est en post-doc et Dylan les aide sur tous les sujets. C'était l'équipe initiale, mais ils sont maintenant sept employés, dix à partir de septembre. Ils ne sont jamais tout seuls sur un projet mais servent d'intermédiaire et de support général aux chercheurs qui s'intéressent à travailler sur des évaluations. On a beaucoup parlé, en particulier cet après-midi, d'un besoin de communication entre les chercheurs et les décideurs. C'est une des fonctions de J-PAL Europe.

Concrètement, aujourd'hui, J-PAL Europe est un réseau de dix chercheurs, dont cinq sont en France, la plupart étant là aujourd'hui : Luc Behagel, Bruno Crépon, Marc Gurgand, Eric Maurin et Philippe Zamora. Cinq ne sont pas en France : Jakob Svensson était là hier,

plusieurs chercheurs à Londres. Un ensemble de projets sont en France. On en a vu un certain nombre. On a été extrêmement enthousiastes de l'appel d'offres du Haut Commissariat et on y a répondu avec cinq projets, dont j'espère qu'ils satisfaisaient un certain nombre des critères qui ont été proposés par Eric. Ils ont été sélectionnés tous les cinq, même s'il a ajouté que ce n'était pas nécessairement une condition suffisante. Nous avons travaillé sur ces projets avec des partenaires très différents : la mission locale de Tulle, le rectorat de Créteil, l'école de la deuxième chance, l'Adie, qui sont à la fois des ONG, des partenaires locaux, un rectorat. Sur ces projets, on essaie d'apporter le maximum de support possible à la fois aux chercheurs, aux administrations ou aux ONG. On a utilisé notre réseau pour recruter des assistants de recherche formidables, avec à présent une personne par projet à plein temps, pour s'assurer que la communication ait lieu et se fasse bien. Il y a aussi tout un ensemble de projets, une dizaine, dans les pays en développement – je vous en épargne la liste – à nouveau en partenariat avec des ONG, des gouvernements. On travaille en particulier de très près avec le royaume du Maroc et avec des partenaires privés, tels que Veolia, avec lequel on travaille aussi au Maroc, donc, différentes institutions.

J-PAL Europe s'occupe aussi de diffuser au maximum la connaissance. Tous les documents que l'on a produits jusqu'à maintenant, dont Abhijit a parlé, sont en anglais. Or une grande partie du monde est francophone. Un des objectifs est donc de traduire tous les documents en français, pour qu'ils puissent être diffusés de manière plus large. Un autre objectif – et c'est vraiment un avantage d'être situé en Europe – est que l'Europe est plus près de beaucoup de pays européens et aussi de beaucoup de pays en développement. On veut donc être capables d'aller voir, d'aller rencontrer les gens. On l'a déjà beaucoup fait. On a travaillé avec l'OCDE, l'UNESCO, SEAGAP, les ministères que j'ai cités en France, le *World Economic Forum* de Davos.

Lundi commence la première session de formation organisée en France, en français, avec quarante-trois participants, dont dix-neuf venant de diverses administrations françaises, des étudiants ou des représentants d'ONG. Les autres viennent de pays francophones en développement. C'est un programme de cinq jours qui sera sûrement répété. Cela permet vraiment d'aller dans les détails, comment cela se passe, pourquoi on fait les choses comme cela. J'espère que cela va répondre à la demande de formation qui a été exprimée ici plus tôt dans l'après-midi. On serait ravis de faire cela à nouveau, s'il existe une demande.

Nos plans pour le futur. Un plan important, que l'on avait déjà et que la réflexion de ces deux jours m'a encore plus convaincue de mettre en place : d'une part, de servir de plate-forme, de support un peu plus vaste qu'uniquement pour les chercheurs affiliés à J-PAL Europe. Si quelqu'un a une question sur le déséquilibre entre l'énergie qui était derrière le Grenelle de l'insertion et le niveau de compétence de ses différents acteurs, on espère pouvoir potentiellement, non seulement par les cours, mais aussi en rencontrant les gens, les aider à comprendre pourquoi faire les choses de cette manière plutôt que d'une autre, etc.

Nous voulons également établir un comité d'éthique, à l'Ecole d'économie de Paris, avec des chercheurs, différentes personnalités pas seulement issues de cette école, afin de s'assurer que

tous ses protocoles satisfont aux standards éthiques internationalement acceptés. Ce sont des plans à venir. Je vous remercie de votre attention.

Pierre CAHUC
Comme d'habitude, s'il y a des questions dans la salle, elles sont les bienvenues.

Marie-Hélène CARLAT, direction départementale du Travail d'Indre-et-Loire
J'ai compris que pour cette méthode randomisée, il y avait besoin d'échantillons de taille importante. Y a-t-il une taille critique ? Est-ce que vous pouvez essaimer auprès d'associations et d'organisations qui atteindraient dans leurs objectifs cette taille critique ? Mais est-ce que vous pouvez aussi faire de la sensibilisation auprès d'associations qui travaillent sur des expérimentations qui sont en dessous de cette taille, pour diffuser plutôt un esprit qu'une méthode ?

Esther DUFLO
La réponse à la première question, c'est qu'il y a une taille critique pour chaque projet, mais pas une taille critique absolue. Cela dépend vraiment de l'effet attendu du programme, de la méthode de mise en place. On peut vraiment aller de deux cents personnes à cent mille personnes, selon la manière dont les choses sont mises en place. Il y a donc une taille critique par projet.

La réponse à la deuxième question, c'est oui sur les deux plans. Il m'est arrivé un certain nombre de fois, et je pense que c'est une conversation utile, de dire à de toutes petites organisations, qui viennent me voir pour me dire : « Nous, on aimerait bien évaluer nos projets. On a une école, qu'est-ce que vous en pensez ? », « Je ne pense pas que cela va être possible avec une école. Assurez-vous que vous avez au moins une évaluation de processus, pour être sûrs que vous mettez en place votre programme de la manière la plus optimale possible. L'effet du programme lui-même, vous ne serez pas capable de l'évaluer vous-mêmes. Mais, peut-être, ce que vous pouvez faire, c'est de vous assurer de choisir un programme qui a été prouvé par d'autres comme étant un programme efficace. Dans ce cas, vous pouvez espérer que, vous aussi, si vous faites bien le programme, cela sera efficace ». On a ce genre de conversation assez souvent.

Annie FOUQUET, ancienne directrice de la DARES, vice-présidente de la Société française d'évaluation
Merci. Je suis très impressionnée par tout ce qui vient d'être dit. Je voudrais profiter de cette énergie qui monte sur l'évaluation pour faire une petite page de publicité pour le colloque de Strasbourg sur l'évaluation des politiques publiques en Europe, que la Société française d'évaluation a monté en partenariat avec la société allemande, la DeGEval, qui va se dérouler lors de la première semaine de la présidence française de l'Union européenne au Parlement européen à Strasbourg. Un atelier sur l'évaluation d'impact est notamment prévu. Je vous invite tous à continuer cette conversation à Strasbourg, les 3 et 4 juillet prochain. http://www.sfe.asso.fr.

Pierre CAHUC

Merci beaucoup. S'il n'y a pas d'autres questions, nous allons pouvoir passer à la conclusion de cette journée. J'accueille Thomas Fatome, directeur de cabinet de Monsieur Wauquiez, et Antoine Magnier.

Retour sommaire

Conclusion

Antoine MAGNIER, directeur de la DARES

Nous approchons effectivement de la fin de cette journée et je suis très heureux d'accueillir Thomas Fatome, qui va clore ces débats. Thomas Fatome dirige le cabinet de Laurent Wauquiez et celui de Christine Lagarde sur les sujets de l'emploi. Il va évidemment nous apporter la voix de nos ministres sur ces sujets.

Avant de lui laisser la parole, je souhaiterais néanmoins remercier l'ensemble des intervenants et notamment ceux qui nous sont venus de l'étranger, qui ont accepté de nous donner un peu de leur disponibilité et de leur temps pour nous éclairer sur ces sujets importants. Je souhaiterais également remercier à nouveau Esther Duflo et Bruno Crépon, qui nous ont apporté un appui décisif pour l'organisation de ce colloque, Dominique Goux et son équipe de la mission de l'animation de la recherche, qui ont organisé ces deux journées. Au cours de ces deux journées, nous avons eu clairement des échanges très riches, à la fois sur les expérimentations menées à l'étranger, et sur celles qui sont menées depuis peu et qui se mettent en place en France. Nous en sortons, me semble-t-il, avec une meilleure idée des potentialités qu'apporte ce type d'approche.

J'ai évidemment noté, comme beaucoup d'entre vous, l'enthousiasme des intervenants, mais j'ai noté aussi que c'était un enthousiasme raisonné. A l'issue de ces deux journées, il me semble que l'on ressort avec une meilleure idée sur les limites et les difficultés de mise en œuvre que posent ces méthodes d'expérimentation contrôlée. Il me semble également que la priorité pour ceux d'entre nous qui travaillent sur ces sujets en France – chercheurs, représentants des administrations centrales ou locales – est aujourd'hui double : elle est d'abord de mener à bien les expérimentations engagées ou qui se mettent en place ; elle est aussi d'en extraire les résultats et de communiquer publiquement sur ces résultats, à la fois avec beaucoup de vigueur, mais aussi avec beaucoup de rigueur scientifique. Ces deux conditions me paraissent être nécessaires pour conforter dans la durée l'intérêt des décideurs de politiques publiques, qui a émergé, sur la période récente, en faveur de ce type d'approche.

A l'issue de ces deux journées et pour terminer, je souhaiterais également former deux vœux : le premier, c'est que ces deux journées aient pu susciter des vocations parmi les jeunes chercheurs qui étaient là, qui sont là aujourd'hui. Le deuxième, c'est que les échanges que nous avons pu avoir au cours de ces deux journées entre participants et intervenants puissent se poursuivre. C'est une condition importante pour que nous puissions continuer à progresser sur ces sujets en France.

Voilà ce que je souhaitais vous dire, avant de laisser le mot de la fin à Thomas Fatome. Merci.

Retour sommaire

Thomas FATOME, directeur de cabinet de M. Laurent WAUQUIEZ et de Mme Christine LAGARDE

Merci beaucoup, Antoine. Bonjour à tous. Je voudrais tout d'abord excuser Christine Lagarde et Laurent Wauquiez, qui auraient aimé être présents aujourd'hui pour conclure ce colloque, mais qui ont été retenus par d'autres obligations. Ils m'ont demandé d'essayer de vous transmettre un certain nombre de messages qui leur tiennent à cœur à tous les deux.

D'abord, vous l'avez sans doute dit dans ces journées, la France a accumulé un certain retard dans ce domaine de l'expérimentation et de l'évaluation. Ce n'est pas forcément sa culture historique et c'est pour cela que le colloque qui se tenait aujourd'hui sur ce thème était très important. Les intervenants présents qui sont parmi les plus grands spécialistes français et internationaux de cette question montrent bien la nécessité de ce colloque. Un autre enjeu, comme l'a dit Antoine Magnier tout à l'heure, consiste à sensibiliser aussi l'ensemble des administrations et des décideurs de la sphère publique et sociale, sur l'intérêt de ce type d'approche. Là encore, il y a sûrement un travail d'acculturation et de conviction à faire pour pousser cette modalité de l'expérimentation. Vous avez illustré, aujourd'hui, les domaines du champ social qui peuvent recourir à cette démarche avec le plus grand bénéfice, en particulier en matière de formation professionnelle, de logement, d'indemnisation du chômage ou de prestations sociales. Nous avons aujourd'hui beaucoup de résultats dans ces domaines, à partir de la comparaison avec les exemples étrangers.

Si j'essaie de revenir sur les principaux enseignements de la journée, vous avez redit l'importance du caractère scientifique et rigoureux de l'approche expérimentale. Si l'on veut pouvoir en tirer quelque chose en termes de décision politique, le premier critère est cette approche scientifique et rigoureuse incontestable. Vous avez démontré à quel point les questions d'ordre méthodologique, également juridique et éthique sont centrales dans la réalisation d'une expérimentation de qualité. L'enjeu est bien d'évaluer, de la manière la plus rigoureuse possible, la plus-value des mesures de politique publique : tel dispositif a-t-il un effet bénéfique pour ses bénéficiaires ? Que se serait-il passé si telle ou telle politique n'avait pas été mise en œuvre ? Toute la difficulté réside bien dans la définition et l'identification de cette situation un peu contrefactuelle, qui n'est pas facilement observable, en pratique, sauf si on s'y prend à l'avance pour bien l'identifier. Cette question est d'autant plus importante pour le politique qu'elle peut représenter évidemment un impact financier, une économie ou une dépense pour les finances publiques et cela est particulièrement important dans le contexte actuel de notre pays. L'évaluation de l'efficacité des dispositifs à petite échelle peut permettre de renoncer à étendre des dispositifs jugés inefficaces même si, là aussi, nous avons de mauvaises habitudes dans notre pays : nous avons parfois du mal à renoncer à des dispositifs, même quand nous avons des doutes sur leur pertinence.

De nombreux exemples étrangers montrent pourtant l'utilité de cette approche, comme les pays anglo-saxons, qui ont une longue pratique en la matière. Certains programmes sont de renommée bien connue : le *Self Sufficiency Project* au Canada, qui a démontré l'efficacité des aides financières au retour à l'emploi pour les bénéficiaires de minima sociaux et l'utilité de combiner ces aides à un accompagnement personnalisé ; le programme *GAIN* en Californie,

qui a démontré l'efficacité des programmes de recherche d'emploi accompagnée pour les bénéficiaires du *Welfare*. Et puis, les pays du nord de l'Europe, la Suède ou les Pays-Bas, ont mené des expérimentations, eux aussi, qui montraient l'utilité de certaines politiques de suivi renforcé, en particulier des personnes en congé maladie, pour éviter qu'elles ne tombent dans l'inactivité.

Nous assistons à un développement récent de ces pratiques, de ces expérimentations, en France. Il se fait avec un certain retard par rapport aux pays anglo-saxons. Une des principales raisons est sans doute le manque de cadre juridique. Cela a été rappelé ce matin, on n'a pas attendu complètement la loi constitutionnelle de 2003, puisque la loi sur le RMI, à la fin des années 80 – début des années 90, comportait déjà en elle-même cette logique d'expérimentation et d'évaluation. Mais, il est vrai qu'avec le cadre de la loi constitutionnelle de 2003, on a des fondements clairs à l'expérimentation législative, avec d'une part la possibilité d'expérimentations dans le cadre du transfert de compétences aux collectivités locales, d'autre part la possibilité d'introduire, même dans la loi, des dispositions à caractère expérimental, qui peuvent déroger au principe d'égalité cher à notre pays.

Ces évolutions juridiques nous donnent maintenant un cadre et peuvent favoriser le développement d'expérimentations. C'est bien le cas. Depuis fin 2006, vous en parliez à l'instant, l'Unédic, l'ANPE, la DARES ont mis en place une évaluation de grande ampleur et extrêmement importante sur l'évaluation des effets des politiques d'accompagnement renforcé par les opérateurs privés et l'ANPE. Nous disposerons très prochainement des résultats. C'est évidemment une avancée à deux titres : d'une part, parce que le domaine de l'accompagnement renforcé, je parle ici vraiment pour Laurent Wauquiez et Christine Lagarde, est un domaine central pour moderniser nos politiques du service public de l'emploi, passer d'une logique de traitement de masse à une logique individualisée. Et aussi parce que dans la méthode, l'association d'équipes de recherche spécialisées dans l'évaluation, comme le CREST ou l'Ecole d'économie de Paris, a permis de créer les conditions d'une estimation rigoureuse au plan scientifique. L'on voit bien que sur un sujet pareil, qui est un sujet très polémique, très sensible – il y avait encore dans *Les Echos* de ce matin, une allusion à la perspective que le futur opérateur délègue à des sous-traitants plusieurs centaines de milliers de chômeurs pour le suivi – la qualité des intervenants et celle de la méthodologie sont indispensables si l'on veut avoir un discours pacifié avec les partenaires sociaux et si l'on veut aussi que les ministres puissent prendre des décisions dans de bonnes conditions. Nous avons de nombreux chercheurs de très haut niveau dans notre pays. Il y en a un certain nombre ici aujourd'hui. Ces expérimentations leur donnent l'occasion d'éclairer très directement la décision en matière de politique publique.

De son côté, cela a été dit aussi également, le Haut Commissariat aux Solidarités actives contre la pauvreté, avec l'appui de la DIIESES, a lancé des appels à projets pour l'année 2008 sur le thème de l'expérimentation sociale. Ces enseignements pourront nourrir la réflexion collective sur les moyens de lutter contre la pauvreté. On est encore en pleine actualité, avec le Grenelle de l'insertion qui va se clôturer dans les tout prochains jours. Trente-sept projets, comme vous le savez, ont été retenus dans divers domaines d'intervention, tels que l'insertion professionnelle des jeunes, notamment ceux issus des quartiers sensibles, l'éducation avec

l'évaluation des dispositifs d'accompagnement éducatif renforcé mis en œuvre par l'Education nationale ou la réduction des inégalités de santé à travers la lutte contre l'obésité, sujet également très important.

Dans les réformes à venir, il existe de nombreux domaines dans lesquels l'expérimentation pourra se mettre en place. La constitution du nouvel opérateur, je l'évoquais ; ses relations aussi avec les acteurs de service public de l'emploi, les missions locales, Cap Emploi, maisons de l'emploi. Il y a des choses à expérimenter, des regroupements territoriaux… On pourrait sans doute tester des choses et voir quelle plus-value on en tire sur une base territoriale. De la même façon, la réforme de la formation professionnelle, qui sera menée au deuxième semestre, peut être aussi un terrain fertile pour tester des dispositifs ou des modes d'organisation entre les acteurs, ils sont nombreux.

En conclusion, les ministres voulaient vraiment encourager l'effort actuellement mené, aussi, dans de nombreux programmes d'enseignement supérieur, au niveau des masters, dans le domaine des politiques publiques, pour enseigner plus largement cette approche de l'expérimentation et de l'évaluation. Un certain nombre d'acteurs sont ici, qui participent au développement de ces programmes d'enseignement.

Il me reste à remercier tous les intervenants de ce colloque au nom des ministres, qui ont fait partager leur expérience de ces méthodes expérimentales et contribuent à ce que l'on saisisse mieux l'enjeu de ces méthodes et leur portée. Je voudrais aussi remercier, en votre nom à tous, l'implication de la DARES, d'Antoine Magnier et de toute son équipe dans l'organisation de ces journées, et puis, plus généralement, l'ensemble du travail que fait la DARES, au quotidien, au service des politiques portées par les deux ministres. Christine Lagarde et Laurent Wauquiez ont souhaité que ce soit souligné. Merci beaucoup et bonne continuation à tous pour vos travaux.

Retour sommaire