Évaluation des Politiques Publiques : expérimentation randomisée et méthodes quasi-expérimentales

Sylvain Chabé-Ferret(*)

Laura Dupont-Courtade(*)

Nicolas Treich(*)

L'évaluation des politiques publiques s'ancre dans les bonnes pratiques des pays développés. Des avancées récentes dans le domaine de l'analyse statistique et de l'accès aux données microéconomiques ont conduit au développement de méthodes expérimentales et quasi-expérimentales appliquées à l'évaluation des politiques publiques. L'objectif de ces méthodes est d'identifier économétriquement les effets causaux des politiques publiques.

L'évaluation expérimentale d'une mesure (politique publique, dispositif, programme) utilise le même principe que les essais cliniques en médecine : au sein de la population deux groupes sont sélectionnés par tirage aléatoire, l'un bénéficie de la mesure et l'autre non. L'impact de la mesure s'obtient en comparant ex post *le groupe d'agents bénéficiaires au groupe de non-bénéficiaires. En pratique, il existe plusieurs types d'interventions randomisées qui permettent d'adapter l'expérimentation aux caractéristiques de la mesure évaluée. Chaque type d'intervention requiert un processus de mise en œuvre spécifique et des outils d'analyse statistique adéquats. Les méthodes quasi-expérimentales utilisent des données d'observation préexistantes pour estimer l'effet d'une politique publique en tentant de se placer au plus près de conditions expérimentales. Les méthodes quasi-expérimentales sont utiles car elles mobilisent moins de moyens et permettent d'éviter des problèmes éthiques, politiques et comportementaux que peuvent induire une allocation randomisée.*

Dans cet article, nous proposons une introduction accessible et non technique à ces méthodes d'évaluation expérimentale et quasi-expérimentale. Nous présentons les concepts et les intuitions à partir d'exemples numériques simples, complétés par des tableaux et des graphiques, sans recourir à des techniques économétriques avancées. Nous illustrons la discussion avec des exemples concrets de politiques publiques, incluant la politique de revenu de solidarité active (RSA), un projet de construction de barrage, un programme de formation professionnelle, et des mesures agro-environnementales (MAE). Nous discutons systématiquement les biais principaux et les problèmes potentiels associés à chaque méthode.

(*) Toulouse School of Economics, Inra, Université Toulouse Capitole, Toulouse.
Email: sylvain.chabe-ferret@inra.fr

Nous remercions l'éditeur et le rapporteur pour leurs commentaires avisés et leur relecture détaillée des versions précédentes de cet article. Cet article est tiré d'un document de travail de la DG Trésor produit dans la perspective d'un dossier thématique des Cahiers de l'Évaluation sur l'évaluation des politiques publiques. Les auteurs remercient Martine Perbet, rédacteur en chef des Cahiers de l'Évaluation, pour ses nombreux commentaires, ainsi que pour sa contribution à certaines parties du texte. Les auteurs restent néanmoins seuls responsables des erreurs et omissions éventuelles.

La France traverse une crise importante des finances publiques et fait face à une défiance croissante à l'égard des décideurs publics (Algan et Cahuc, 2007 ; Ferracci et Wasmer, 2011). Dans ce contexte, il peut être utile de renforcer la transparence et l'évaluation des décisions publiques.

Des avancées récentes dans le domaine de l'analyse statistique et de l'accès aux données microéconomiques ont conduit au développement de méthodes d'évaluation *ex post* des politiques publiques (Erkel-Rousse, 2014 ; Roux, 2015). L'objectif de ces méthodes est d'identifier économétriquement les effets causaux des politiques publiques. Ces méthodes fondées sur des approches économétriques de forme réduite ont eu un impact massif sur la recherche en économie, principalement en économie du travail, de l'éducation et du développement. Elles sont à l'origine de ce que l'on appelle la « révolution de la crédibilité » en économie (Angrist et Pischke, 2010 ; Imbens, 2010). Au cœur de ces méthodes, on trouve la notion d'expérimentation qui utilise le cadre statistique des tests médicaux. Les méthodes expérimentales sont directement fondées sur le tirage aléatoire de deux groupes au sein d'une même population, l'un bénéficiant du traitement, l'autre non. Les méthodes quasi-expérimentales utilisent des données d'observation préexistantes pour estimer l'effet d'une politique publique en tentant de se placer au plus près de conditions expérimentales.

Les méthodes (quasi-)expérimentales sont depuis longtemps utilisées par les agences de régulation aux États-Unis et par des instances internationales comme la Banque Mondiale. Il n'en est pas de même en France, où le développement de ces méthodes en tant qu'aide à la décision publique est bien plus récent et plus modeste (voir encadré 1). Il remonte à deux mesures phares des politiques sociales mises en œuvre en 2007 qui ont fait l'objet d'une expérimentation : le « revenu de solidarité active » et « l'accompagnement renforcé des demandeurs d'emploi ». La France est encore dans une période d'apprentissage même si les évaluations expérimentales réalisées dans le cadre du Fonds d'Expérimentation pour la Jeunesse (Bérard et Valdenaire, 2014) montrent que ces approches commencent à se diffuser. La progression des expérimentations apparaît limitée par divers facteurs dont, notamment, les problèmes politiques et éthiques que soulèvent, dans notre culture, le tirage aléatoire des échantillons et l'accessibilité aux données individuelles. Les limites politiques sont néanmoins fortement atténuées depuis que la loi constitutionnelle du 28 mars 2003 autorise de déroger au principe d'égalité de traitement à des fins d'expérimentation. Les écueils éthiques ont aussi été réduits par le développement de comités éthiques au sein des universités qui évaluent les plans d'expériences. Toutefois, un autre facteur contribue à la lente progression des méthodes expérimentales et quasi-expérimentales. En effet, ces méthodes reposent sur la connaissance des différentes possibilités de *designs* et sur des techniques statistiques nouvelles et spécifiques auxquelles peu de décideurs publics ont été formés.

Dans cet article, notre ambition est de produire une introduction accessible aux méthodes expérimentales et quasi-expérimentales d'évaluation des politiques publiques. Nous essayons d'expliquer les concepts et les intuitions à partir d'exemples numériques simples, complétés systématiquement par des tableaux et des graphiques, sans recourir à des techniques avancées, ni au jargon économétrique. Nous présentons les différents *designs* possibles de randomisation en expliquant systématiquement les avantages et les inconvénients de chaque *design*. Nous procédons de manière progressive, en donnant au fur et à mesure des références bibliographiques clefs dans la littérature académique, souvent présentées sous la forme d'encadrés, afin que le lecteur intéressé puisse s'y référer de manière indépendante. En ce sens, notre travail va plus loin dans les détails techniques que des ouvrages grands publics sur l'application des méthodes expérimentales (Ferracci et Wasmer, 2011 ; Banerjee et Duflo, 2012), mais reste néanmoins plus accessible qu'une présentation systématique des différentes méthodes économétriques appliquées à l'évaluation *ex post* des politiques publiques disponibles en français (Magnac, 2000 ; Brodaty, Crépon et Fougère, 2007 ; Legendre, 2013 ; Givord, 2014) ou en anglais (Angrist et Pischke, 2014 ; Di Nardo et Lee, 2011 ; Heckman, 2001 ; Imbens et Rubin, 2015 ; Todd, 2007). Behaghel (2006), L'Horty et Petit (2011), les membres du Conseil d'Analyse Économique (2013) et Desplatz et Ferracci (2016) adoptent une approche pédagogique similaire à la nôtre et nous renvoyons le lecteur à ces travaux pour des éléments complémentaires. Dans cet article, nous abordons cependant de manière plus détaillée les différents *designs* d'expériences randomisées.

Le plan de l'article est le suivant. Après un rappel des difficultés rencontrées pour identifier une relation causale entre une politique et un résultat, nous présentons les différents *designs* d'expérimentations permettant d'établir cette relation. Nous présentons ensuite un panorama des méthodes quasi-expérimentales.

Le problème fondamental d'inférence causale et les biais des comparaisons intuitives

L'objectif de l'évaluation *ex post* est de comparer la situation observée en présence de la politique à une situation de référence en l'absence de la politique. La principale difficulté de l'évaluation *ex post* d'une politique publique est la reconstitution de la situation de référence. L'évaluateur fait face à un manque fondamental de données : il est impossible d'observer simultanément la situation en présence de la politique évaluée et la situation en son absence. À un instant t un individu ne peut pas être à la fois bénéficiaire et non bénéficiaire d'une politique. Il est donc impossible d'observer directement l'impact de la politique évaluée sur chaque individu. Ce problème est connu sous le nom de problème fondamental d'inférence causale.

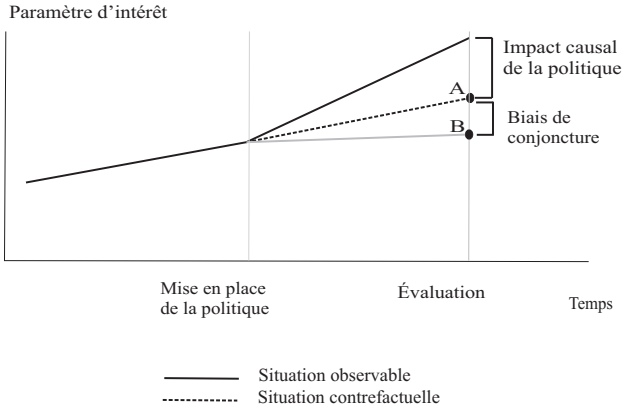
L'évaluateur peut essayer d'approcher la situation contrefactuelle en examinant la situation qui préexistait avant la mise en place de la politique ou celle des individus ne bénéficiant pas de la politique. Malheureusement, ces deux comparaisons intuitives sont généralement biaisées, comme nous allons le voir dans les parties suivantes. Les concepts développés dans ces parties et dans le reste de l'article s'appuieront typiquement sur cinq exemples d'évaluation : la politique du revenu de solidarité active (RSA), une importante réforme des minima sociaux succédant au revenu minimum d'insertion (RMI) en 2009 ; un projet de construction de barrage ; un programme social de formation professionnelle ; des mesures agro-environnementales (MAE) et une campagne de vaccination.

La comparaison avant/après

Prenons l'exemple du RSA. Pour évaluer l'impact du RSA, une première approche (simple) consiste à comparer la « situation avant » et la « situation après » la mise en œuvre de cette politique, c'est-à-dire à comparer le taux d'activité en vigueur en France, lorsque le RMI était encore en place, par rapport au taux d'activité en vigueur aujourd'hui avec le RSA. Cette approche n'apparaît pas satisfaisante car d'autres facteurs que le RSA auraient pu influencer le taux d'activité entre ces deux dates. Dans une période économique difficile, par exemple, où le chômage tend à s'accroître, la comparaison avant/après conduirait à sous-estimer l'impact positif du RSA sur le chômage.

Autre exemple, la construction d'un barrage (voir par exemple l'étude de Duflo et Pande, 2007). On peut s'intéresser à l'impact du barrage sur la productivité agricole en comparant la situation initiale à celle 10 ans après sa construction. Ceci ne prendrait pas en compte les innovations de procédé et les innovations technologiques qui seraient survenues au cours de cette période même en l'absence du barrage (par exemple, la révolution verte). L'impact économique du barrage serait ainsi surestimé. On dit que la comparaison avant/après souffre d'un biais de conjoncture. D'où l'intérêt de simuler une situation (le contrefactuel) correspondant à ce qu'il serait advenu s'il n'y avait pas eu de nouvelle politique. La comparaison de la « situation avec politique » au contrefactuel permet alors d'estimer l'impact réel de cette politique.

Figure 1 : biais de conjoncture



La figure 1 représente l'effet causal d'une politique et le biais de conjoncture potentiel résultant de la comparaison avant/après. Le contrefactuel nécessaire est la situation représentée par le point A alors qu'une comparaison avant/après utiliserait le point B. L'effet estimé est ainsi composé de l'effet réel ainsi que du biais de conjoncture. Dans le cas présent, cela mènerait à une surestimation de l'effet causal. La comparaison avant/après n'est valide que sous la condition que seule la politique étudiée fasse varier le paramètre d'intérêt au cours du temps. Il va sans dire qu'il s'agit d'une hypothèse forte. La politique du RSA à elle seule ne peut expliquer l'entière variation du taux d'activité qui est impacté par de nombreux autres facteurs comme l'activité économique par exemple. De même, la présence d'un barrage n'est pas l'unique source de variation de la productivité agricole : les conditions météorologiques et les technologies de production jouent aussi un rôle par exemple. Il faut donc être très prudent avec ce type d'analyse avant/après, pourtant très commune, qui peut aisément mener à des conclusions erronées.

La comparaison avec/sans

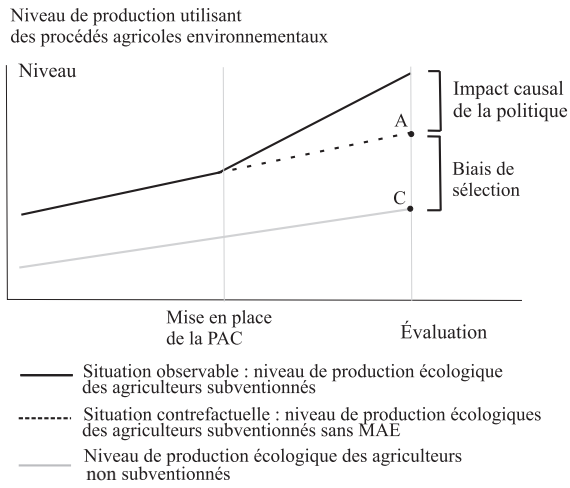
Une méthode alternative pour évaluer une politique publique consiste à comparer un groupe de personnes affectées par la politique à un groupe de personnes non affectées. Pour évaluer le RSA par exemple, on pourrait comparer, au sein de la

population éligible, un groupe ayant recours au RSA et un groupe de non recourants. Pour l'évaluation de l'impact d'un barrage, on pourrait comparer des zones avec barrage à des zones sans barrage. Cependant, pour être comparées, deux populations doivent être comparables. Ce n'est malheureusement généralement pas le cas. En effet, la différence entre un groupe bénéficiaire et un groupe de non bénéficiaires est due, d'une part à l'effet de la politique et d'autre part à la différence initiale entre les deux groupes. La différence initiale entre ces deux groupes est connue sous le nom de biais de sélection, et doit être prise en compte dans l'évaluation.

Ainsi évaluer le RSA en comparant le niveau d'activité des non-recourants et des recourants n'a pas de sens s'ils diffèrent trop en termes de caractéristiques sociales et de comportement face à cette politique. Si les non-recourants ne demandent pas le RSA parce qu'ils estiment être proche de l'emploi, l'effet du RSA sur l'emploi sera sous-estimé. Imaginons maintenant que l'on souhaite comparer deux régions, l'une avec barrage et l'autre sans barrage. Supposons aussi que le barrage a été construit dans une zone rurale pauvre afin de favoriser le développement de cette zone. Une simple comparaison avec/sans conduirait dans ce cas à une sous-estimation de l'impact causal de la construction d'un barrage.

Prenons un autre exemple. L'un des objectifs des mesures agro-environnementales (MAE) de la politique agricole commune, la PAC, est d'encourager l'utilisation de techniques de production agricole écologiques. Des subventions sont versées aux agriculteurs qui emploient ces techniques. Supposons que l'on veuille évaluer l'impact de ces subventions sur le niveau d'utilisation de techniques agricoles écologiques (Chabé-Ferret et Subervie, 2013). On cherche donc la différence entre le niveau d'utilisation lorsque les agriculteurs sont subventionnés et lorsqu'ils ne le sont pas. Cette dernière situation contrefactuelle est représentée par la droite grise figure 2 ci-après. Lors de l'évaluation *ex post* elle atteint le point A. Si l'on compare le niveau d'utilisation des agriculteurs subventionnés et des agriculteurs non subventionnés, qu'advient-il du biais de sélection ? Il est probable qu'un nombre substantiel d'agriculteurs recevant les subventions utilisaient ces méthodes avant la mise en œuvre de la politique et les auraient donc appliquées même en absence de MAE. Utiliser les agriculteurs non subventionnés comme simulation du contrefactuel, représenté par le point C sur la figure 2, surestime donc l'effet causal de ces subventions sur le niveau d'utilisation des techniques de production écologique parce que le comportement des agriculteurs en l'absence des subventions n'est pas bien pris en compte.

Figure 2 : biais de sélection : exemple de la PAC



Un exemple numérique : la formation professionnelle

Nous allons maintenant illustrer les biais de comparaison avant/après et avec/sans en s'appuyant sur un exemple numérique. Imaginons un programme proposant des formations professionnelles aux personnes au chômage afin de les aider à retourner vers l'emploi. Supposons que l'on souhaite évaluer l'impact de ce programme sur le retour à l'emploi. Les individus éligibles au programme sont les individus au chômage au moment où le programme est mis en place. Faisons l'hypothèse simplificatrice que les individus répondant aux critères d'éligibilité du programme appartiennent à deux types définis comme suit :

– les individus de type 1 ont des caractéristiques les rendant plus à même de réclamer une formation. Il s'agit des individus dans une situation économique et sociale précaire, avec un niveau de revenu faible et un niveau de qualification bas. Ces individus auront une probabilité élevée d'accepter un programme de formation professionnelle. Le taux d'activité dans un groupe d'individus de type 1 sans programme s'élève à 30 %. Ils sont représentés par un point noir figure 3. Les individus de type 1 représentent 60 % de la population ;

– les individus de type 2, *a contrario*, sont des individus ayant des caractéristiques les rendant moins à même de réclamer un programme de formation. Il peut s'agir d'individus plus dynamiques, plus proche de l'emploi ou ne ressentant pas le besoin d'être aidé. Ces individus ont une probabilité de trouver un emploi supérieure à celle des individus de type 1. Le taux d'activité au sein de ce groupe d'individus est de 90 % en l'absence du programme. Ils sont signalés par un point gris figure 3. Les individus de type 2 représentent 40 % de la population.

Examinons pourquoi les méthodes d'évaluation avant/après et avec/sans mènent à des résultats biaisés. Le tableau 1 présente cet exemple numériquement et la figure 3 illustre le biais de sélection induit par une évaluation par comparaison avec/sans. Notons que les valeurs données sur ces figures ne sont pas réalistes et ne sont là qu'à titre d'illustration. Procéder à l'évaluation de ce programme par une comparaison avec/sans revient à comparer le niveau d'activité *ex post* d'un groupe de recourants au niveau d'activité d'un groupe d'individus non recourants. Le point clef est que le groupe de traitement et le groupe de contrôle sont formés suite à de l'auto-sélection : les individus choisissent de demander ou non le programme. Ce processus ne garantit pas que la composition des deux groupes soit la même : c'est ce qui constitue le biais de sélection.

Le tableau 1 indique que 80 % des individus de type 1 souscrivent au programme, alors que seulement 10 % des individus de type 2 y ont recours. Le groupe de recourants est ainsi constitué de 92 % d'individus de type 1 alors que le groupe de non recourants est constitué de 75 % d'individus de type 2. Ainsi, le groupe de non recourants comprend davantage d'individus qui ont une plus grande probabilité de retrouver un emploi, et ne forme donc pas une bonne approximation de la situation contrefactuelle. Ce problème de sélection biaise la comparaison avec/sans. Sans correction de ce biais, le programme de formation professionnelle semble impacter négativement le niveau d'activité des bénéficiaires. Plus précisément, avec les données établies tableau 1, le programme semble diminuer de 12 points de pourcentage le taux d'activité moyen alors que l'effet réel est une augmentation du taux d'activité de 28 points de pourcentage. L'effet du type domine l'effet du programme, et ne permet pas de différencier l'impact du type sur le paramètre d'intérêt de l'impact du programme : on dit que le type est un **facteur de confusion de l'effet du programme**.

Supposons maintenant que l'on veuille évaluer ce programme de formation professionnelle avec une comparaison avant/après. Cela consiste à comparer le taux d'activité d'un même groupe d'individus éligibles avant et après la mise en place du programme. Le type de biais émergeant ici est un biais de conjoncture : si l'on observe une diminution du taux moyen d'activité par rapport à la situation *ex ante*, comment savoir si cela ne provient pas d'un contexte socio-économique défavorable ou du programme lui-seul ? L'exemple numérique du tableau 1 présente une version pessimiste où le biais de conjoncture est de -0,1. Autrement dit, en l'absence du programme de formation, le taux d'activité aurait chuté de 10 points de pourcentage. Dans notre exemple, la comparaison avant/après sous-estime donc l'effet du programme de formation.

Figure 3 : comparaison avec/sans : exemple d'un programme de formation professionnelle

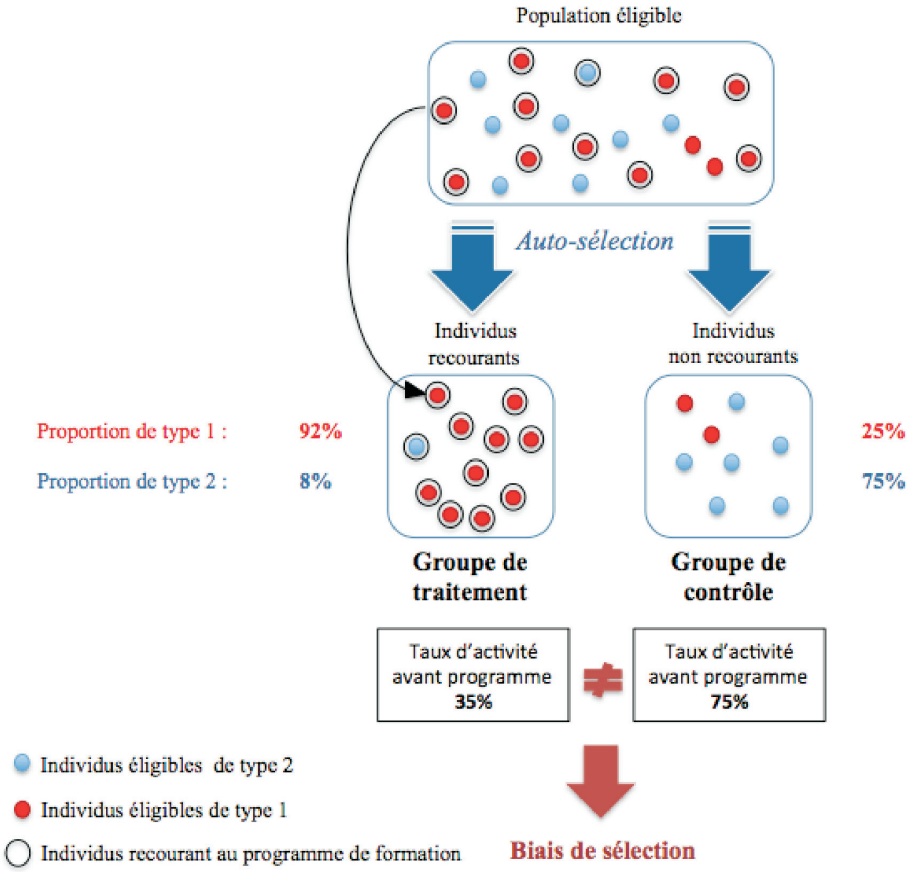


Tableau 1 : comparaison avec / sans et comparaison avant/après – exemple numérique d'un programme de formation professionnelle

	Taux d'activité avant programme ($t = 0$)	Taux d'activité après programme ($t = 1$)		Impact causal moyen du programme	Taux d'auto-sélection
		Traité (Avec formation)	Non-traité (Sans formation)		
Individus de type 1	0,3	0,5	0,2	0,3	0,8
Individus de type 2	0,9	0,9	0,8	0,1	0,1

Impact du contexte socio-économique sur le taux d'activité des individus (biais de conjoncture)	-0,1	Proportion de type 1 dans la population	0,6
		Proportion de type 2 dans la population	0,4

	Comparaison avec/sans			Comparaison avant/après
	Groupe des recourants	Groupe des non recourants		Groupe de recourants
Proportion de type 1 dans le groupe	0,92	0,25		0,92
Proportion de type 2 dans le groupe	0,08	0,75		0,08
Taux d'activité avant randomisation ($t = 0$)	0,35	0,75		0,35
Taux d'activité avec programme ($t = 1$)	0,53	0,80		0,53
Taux d'activité sans programme ($t = 1$)	0,25	0,65		0,25
Impact causal moyen réel sur les recourrants (ICM)	0,28			0,28
Comparaison avec/sans	-0,12		Comparaison avant/après	0,18
Biais de sélection	-0,40		Biais de tendance temporelle	-0,10
Situation utilisée pour simuler le contrefactuel	Groupe des non-recourants en $t = 1$			Traités en $t = 0$

	= Non observable

L'expérimentation randomisée

L'expérimentation randomisée est une méthode d'analyse utilisée originellement dans les sciences biomédicales. Elle est aujourd'hui de plus en plus utilisée en économie et en particulier dans le domaine de l'économie du développement, de l'éducation et du travail.

Définition et randomisation

Définition

Une expérience randomisée consiste à comparer deux groupes formés **aléatoirement** à partir d'un échantillon d'individus : un groupe de traitement, au sein duquel les individus sont sujets à une intervention expérimentale et un groupe de contrôle utilisé comme groupe de référence.

Le rôle de la randomisation

Le principe clé de l'expérimentation randomisée réside dans la *randomisation de l'allocation du traitement*. La distribution aléatoire des individus entre les groupes de contrôle et de traitement permet de garantir, pour un échantillon suffisamment important, que les individus ayant des caractéristiques différentes sont répartis à peu près de façon homogène entre les deux groupes. Autrement dit, si l'on fait l'hypothèse simplificatrice qu'il existe deux types d'individus et qu'on les assigne aux groupes de contrôle et de traitement en tirant à pile ou face, on devrait obtenir environ la même proportion de chaque type dans chaque groupe. Cette méthode permet d'obtenir des groupes de traitement et de contrôle statistiquement similaires et donc comparables. Ainsi, l'expérimentation randomisée assure l'absence de biais de sélection qui, rappelons-le, émerge lorsque

le groupe de contrôle est une mauvaise reconstitution du groupe de traitement dans la situation où ce dernier n'aurait pas été traité.

Si l'on reprend l'exemple du programme de formation professionnelle présenté précédemment, la randomisation garantit les mêmes proportions d'individus de type 1 et de type 2 dans le groupe de traitement et dans le groupe de contrôle, ce qui évite ainsi toute source de confusion entre l'effet du type et l'effet du traitement sur le taux d'activité. Autrement dit, la randomisation est un système d'allocation qui permet aux groupes d'être formés de façon indépendante des caractéristiques des sujets. Ainsi, la différence *ex post* entre les deux groupes provient uniquement du traitement et non d'une différence initiale de caractéristiques entre les deux groupes comme dans une comparaison avec/sans. Cette différence *ex post* représente l'impact causal moyen du traitement (ICM). La figure 4 schématise ce mécanisme.

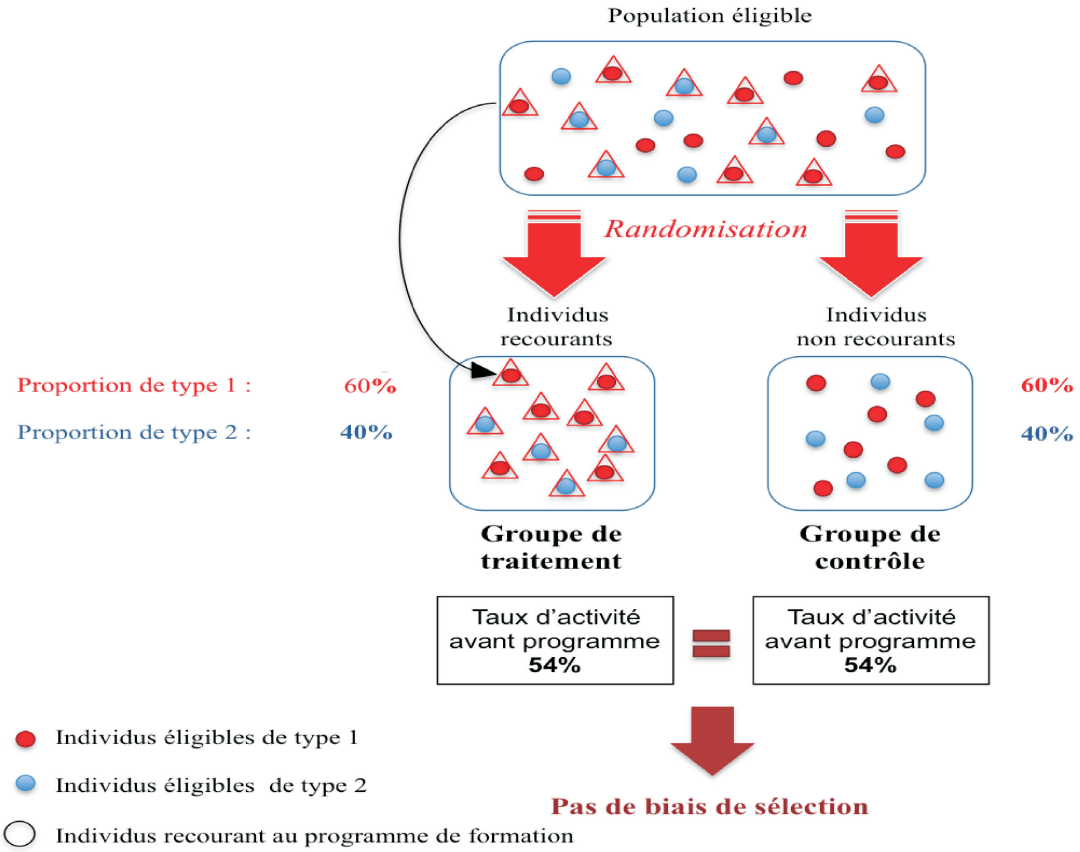
En pratique, il existe plusieurs manières de mettre en œuvre une expérience randomisée. Chacune de ces manières peut être plus ou moins adaptée à un programme donné. Dans ce chapitre, les différents *designs* d'expériences randomisées sont décrits en détail et illustrés par l'exemple du programme de formation professionnelle. La dernière partie de ce chapitre présente les limites majeures associées à l'expérimentation randomisée et quelques solutions.

Design 1 : randomisation d'un traitement

Nous présentons dans un premier temps le *design* de base. Prenons pour cela l'exemple du programme de formation professionnelle et supposons que l'on souhaite l'évaluer avec une expérience randomisée. Les différentes étapes de ce *design* sont illustrées dans la figure 4 et les résultats de cette évaluation sont présentés dans le tableau 2. Les données initiales sont les mêmes que celles utilisées pour construire le tableau 1. Dans un premier temps ($t = 0$), deux groupes sont formés par tirage aléatoire au sein de la population éligible au programme. La randomisation garantit la même proportion d'individus de type 1 et de type 2 dans chacun des groupes. Dans un second temps ($t = 1$), le traitement est donné à l'un des deux groupes : tous les individus le composant reçoivent une formation professionnelle. Ce groupe est le groupe de traitement, l'autre groupe constitue le groupe de contrôle.

Les deux groupes étant constitués des mêmes proportions de chaque type, ils sont similaires à la période $t = 0$ quand personne ne reçoit de traitement. Le niveau d'activité moyen est le même dans les deux groupes, soit 54 %. Ainsi, en $t = 1$, une fois que le groupe de traitement a suivi une formation, le groupe de contrôle reconstitue le groupe de traitement dans la situation hypothétique où le groupe traité n'aurait pas reçu de formation. En effet, en $t = 1$, les deux groupes ont toujours la même composition de types et ne diffèrent que par l'allocation du traitement, ce

Figure 4 : *design* 1 : randomisation d'un traitement, exemple d'un programme de formation professionnelle



Encadré 1 : l'apprentissage de l'expérimentation en France

« Les méthodes expérimentales sont utilisées depuis longtemps dans les sciences dures, en médecine, en agronomie ou même en marketing. Levitt et List (2008), dans leur survol historique, distinguent trois générations d'évaluation aléatoire d'expériences de terrain. La première remonte aux travaux de Neyman et Fisher dans les années 1920 et 1930, où l'évaluation aléatoire est pour la première fois conçue comme un outil permettant d'identifier des effets causals et est appliquée en agronomie. La deuxième génération est celle des expérimentations sociales de grande échelle à partir des années 1960, où l'objet de l'expérimentation n'est plus des terres agricoles mais des groupes de personnes. En référence aux premiers travaux agronomiques, on parle d'essais de terrain (« *Field Trials* ») pour désigner ces méthodes appliquées au social (Burtless, 1995). Plus récemment, un troisième âge de l'expérimentation aurait été ouvert avec un élargissement considérable de leurs domaines d'application (développement, éducation, lutte contre la pauvreté, santé, …), et du nombre et des types de questions traitées.

Les exemples les plus cités d'évaluations aléatoires de grands programmes sociaux viennent tous d'Amérique du Nord : l'expérimentation du New Jersey menée en 1968 pour tester un dispositif d'impôt négatif, suivie de trois autres expérimentations aux États-Unis au début des années 1970 ; le *Self Sufficiency Project* qui est une prime donnée à des bénéficiaires d'aide sociale pour les inciter au retour à l'emploi, expérimentée dans deux provinces canadiennes à partir de 1994 (Nouveau Brunswick et Colombie britannique) ; le programme *Moving to Opportunity*, mis en œuvre entre 1994 et 1998 pour favoriser la mobilité résidentielle des ménages pauvres dans cinq villes des États-Unis (Baltimore, Boston, Chicago, Los Angeles et New York) ; le *Progresa-Oportunidades* qui encourage depuis 1997 la scolarisation des enfants pauvres au Mexique. Ces méthodes sont désormais mises en œuvre dans tous les pays du nord de l'Europe, en Australie et dans de nombreux pays en développement, pour évaluer des programmes dans des domaines très variés (accès à l'emploi, lutte contre la pauvreté, amélioration des pratiques sanitaires, etc.). […].

[La] première évaluation aléatoire de grande taille réalisée en France porte sur une expérimentation qui a eu lieu à l'automne 2007. Elle a été mise en œuvre par le Crest, l'École d'économie de Paris et le Jameel – Poverty Action Lab pour évaluer les effets des opérateurs privés d'accompagnement des demandeurs d'emploi inscrits à l'ANPE (Behaghel, Crépon et Gurgand, 2014). […]. [Martin Hirsch, devenu] Haut-Commissaire aux solidarités actives en juin 2007 et également Haut-Commissaire à la jeunesse en janvier 2009, [est] l'initiateur du revenu de solidarité active (RSA). [Il] va soutenir de façon constante le développement des expérimentations sociales et de leur évaluation. Un premier appel à projet d'expérimentation sociale est lancé en 2007 avec un budget de 6 M€. Il est suivi en 2009 par une série d'appels à projets lancés par le fonds d'expérimentations pour la jeunesse (créé par l'article 25 de la loi généralisant le RSA du 1er décembre 2008) avec un budget total, issu d'un partenariat public-privé, de 150 M€. Plus de 400 projets innovants sont ainsi financés qui prévoient fréquemment, mais pas systématiquement, une évaluation aléatoire.

L'évolution du cadre juridique et institutionnel a joué également un rôle important dans le développement des expérimentations sociales en France. Plusieurs obstacles législatifs et réglementaires ont dû être levés pour rendre possible ce développement qui implique, en pratique, une rupture temporaire du principe d'égalité. Un cadre juridique est donné par la réforme constitutionnelle de décentralisation de 2003 et l'adoption la même année de la loi organique relative à l'expérimentation par les collectivités territoriales. Les expérimentations sociales deviennent possibles dès lors qu'elles ont un objet circonscrit et une durée limitée dans le temps et si elles sont menées en vue d'une généralisation. Elles doivent s'effectuer à l'initiative des collectivités locales et doivent nécessairement faire l'objet d'une évaluation. En pratique, l'expérimentation du revenu de solidarité active (RSA) prévue dans la loi du 21 août 2007 en faveur du travail, de l'emploi et du pouvoir d'achat (« loi Tepa ») va constituer la première expérimentation sociale de grande ampleur en France, même si cette expérimentation n'a finalement pas été évaluée selon une méthode expérimentale. »

Source : L'Horty Y. et Petit P. (2010).

qui assure que le groupe de contrôle a le niveau d'activité que le groupe traité aurait eu s'il n'avait pas reçu de formation professionnelle. Le groupe de contrôle reconstitue donc bien la situation contrefactuelle. L'impact causal du programme sur le taux d'activité est donc mesuré sans biais par la différence de niveau d'activité entre les deux groupes en t=1.

Application numérique

D'après notre exemple numérique présenté dans le tableau 2, la mise en place d'un programme de formation professionnelle augmente, en moyenne, de 22 points de pourcentage le taux d'activité. En effet, le programme augmente respectivement de 30 % et 10 % le taux d'activité des individus de type 1 et de type 2 et le groupe de traitement est composé de 60 % d'individus de type 1 et de 40 % d'individus de type 2. L'impact moyen réel du traitement est donc donné par le calcul 0,6*0,3+0,4*0,1 = 0,22, correspondant à (1) - (3) dans le tableau 2.

Cependant, les impacts causaux moyens du programme sur chaque type, 0,3 et 0,1, ne sont pas observables dans la réalité. Nous avons vu que l'expérimentation randomisée permet d'estimer l'impact moyen du programme en faisant la différence de niveau d'activité entre le groupe de traitement et le groupe de contrôle après la mise en place du traitement. Cette différence correspond à

Tableau 2 : *design* 1 : randomisation de l'ensemble de la population – exemple numérique d'un programme de formation professionnelle

	Taux d'activité avant programme ($t = 0$)	Taux d'activité après programme ($t = 1$)		Impact causal moyen du programme
		Traité (avec formation)	Non-traité (sans formation)	
Individus de type 1	0,3	0,5	0,2	0,3
Individus de type 2	0,9	0,9	0,8	0,1

Impact du contexte socio-économique sur le taux d'activité des individus (biais de conjoncture)	-0,1

Proportion de type 1 dans la population	0,6
Proportion de type 2 dans la population	0,4

	Design 1 : experience randomisee	
	Groupe de traitement (avec formation)	Groupe de contrôle (sans formation)
Proportion de type 1 dans le groupe	0,6	0,60
Proportion de type 2 dans le groupe	0,4	0,40
Taux d'activité avant randomisation ($t = 0$)	0,54	0,54
Taux d'activité avec programme ($t = 1$)	$0,66^{(1)}$	$0,66^{(2)}$
Taux d'activité sans programme ($t = 1$)	$0,44^{(3)}$	$0,44^{(4)}$
Impact causal moyen réel (ICM)	$0,22^{(1)-(3)}$	
Impact causal moyen estimé (ICM estimé)	$0,22^{(1)-(4)}$	
Biais de sélection = IMC - IMC estimé	0,00	
Groupe utilisé pour simuler le contrefactuel	Groupe des non-traités après randomisation (Groupe de contrôle en $t = 1$)	

= Non observable

(1) - (4) dans le tableau 2, soit 0,66 - 0,44 = 0,22. On retrouve exactement l'impact causal moyen réel.

Avantages et inconvénients du design *1*

En permettant au groupe de contrôle de reconstituer le contrefactuel, ce premier *design* d'expérimentation randomisée offre une estimation sans biais de sélection de l'impact causal moyen d'un traitement sur l'ensemble de la population. Plusieurs problèmes sont cependant posés par ce *design*. Il requiert tout d'abord d'obliger les individus du groupe de traitement à accepter le programme, alors que certains individus peuvent éventuellement ne pas le souhaiter. Il prive, d'autre part, les individus du groupe de contrôle d'un programme potentiellement bénéfique. Ce premier *design* soulève donc une question éthique.

Par ailleurs, comme présenté dans le tableau 3, l'effet causal obtenu par ce *design* est l'*effet causal moyen sur toute la population éligible (ICM)*. Or, comme nous venons de l'évoquer, cette population comprend potentiellement des individus qui ne souscriraient pas au programme dans le cadre d'une mise en place non expérimentale. Ainsi, étant donné que certains individus sont considérés comme traités alors qu'ils ne l'auraient pas été dans une version non expérimentale (non obligatoire) du programme, ce que l'on obtient ici est uniquement l'effet du programme sous sa forme expérimentale. Cela ne nous permet donc pas d'inférer l'effet réel du programme. L'effet qui nous intéresse est l'effet d'un programme sur les individus qui y auront recours : l'*impact causal moyen du traitement sur les recourants (ICMR).*

La différence entre l'ICMR et l'ICM provient de l'hétérogénéité des caractéristiques individuelles, autrement dit de l'existence de différents types d'individus. Dans notre exemple le programme impacte plus fortement les individus de type 1 qui sont en plus grande difficulté économique et sociale

et qui, par ailleurs, ont davantage recours au programme. Pour ces deux raisons l'impact causal du programme sur les recourants est d'un plus grand intérêt que l'impact causal moyen du programme. En revanche, si l'on considère une population composée d'individus aux caractéristiques identiques, l'impact causal d'un traitement sur les recourants et l'impact causal de ce même traitement sur l'ensemble de la population sont égaux : si tous les individus réagissent de la même façon à un traitement, la composition des groupes étudiés pour l'évaluation n'importe pas. La littérature économique relative aux expériences randomisées propose des *designs* alternatifs qui offrent des solutions à ces différents problèmes.

Design 2 : randomisation après auto-sélection

Ce *design* d'expérience randomisée réalise la randomisation après avoir laissé le choix aux individus de candidater au programme évitant ainsi toute obligation de participation. Ce deuxième *design* est schématisé dans la figure 5 et illustré numériquement dans le tableau 3.

Restons dans le cadre de l'évaluation d'un programme de formation professionnelle. Ce deuxième type d'expérience randomisée laisse le choix aux individus de candidater au programme de formation ; on parle d'auto-sélection. Deux groupes sont ainsi formés : les recourants et les non-recourants. L'impact du programme sur les recourants est bien ce qui nous intéresse pour évaluer une politique publique de formation professionnelle volontaire étant donné que les non-recourants ne demanderaient pas de formation dans le cas d'une implémentation non expérimentale du programme.

Dans ce *design*, le contrefactuel souhaité est donc un groupe ayant la même composition de types que le groupe de recourants mais ne recevant pas de traitement. Après auto-sélection on obtient un groupe candidat qui comprend davantage d'individus de type 1 que d'individus de type 2. La seconde étape consiste à allouer de façon aléatoire ces individus à un groupe de traitement. Parmi les recourants, seulement une partie recevra véritablement une formation. Le groupe formé des recourants n'ayant pas accès à la formation forme un groupe de contrôle qui permet de rendre observable la situation contrefactuelle nécessaire : des recourants non traités.

De la même façon que dans le *design* 1, les groupes de traitement et de contrôle ont des caractéristiques identiques (mêmes proportions de types : 92 % de type 1 et 8 % de type 2). On peut ainsi procéder à la comparaison de leur taux d'activité (53 % et 25 %). Cette différence de niveau d'activité nous donne l'effet de la formation professionnelle sur le taux d'activité au sein du groupe de recourants qui est égal à 28 points de pourcentage supplémentaires. En effet, cette méthode nous donne l'impact causal moyen du traitement sur les recourants (ICMR) et non l'impact causal moyen (ICM) qui représente la variation du paramètre d'intérêt sur toute la population. On voit, dans le tableau 3, que le taux d'activité sans programme en $t = 1$ du groupe de traitement (25 %), qui représente le contrefactuel, est égal au taux d'activité du groupe de contrôle. Ainsi, cette méthode nous donne une estimation de l'impact du traitement identique à l'impact réel (28 %), le biais de sélection se trouvant ainsi réduit à zéro.

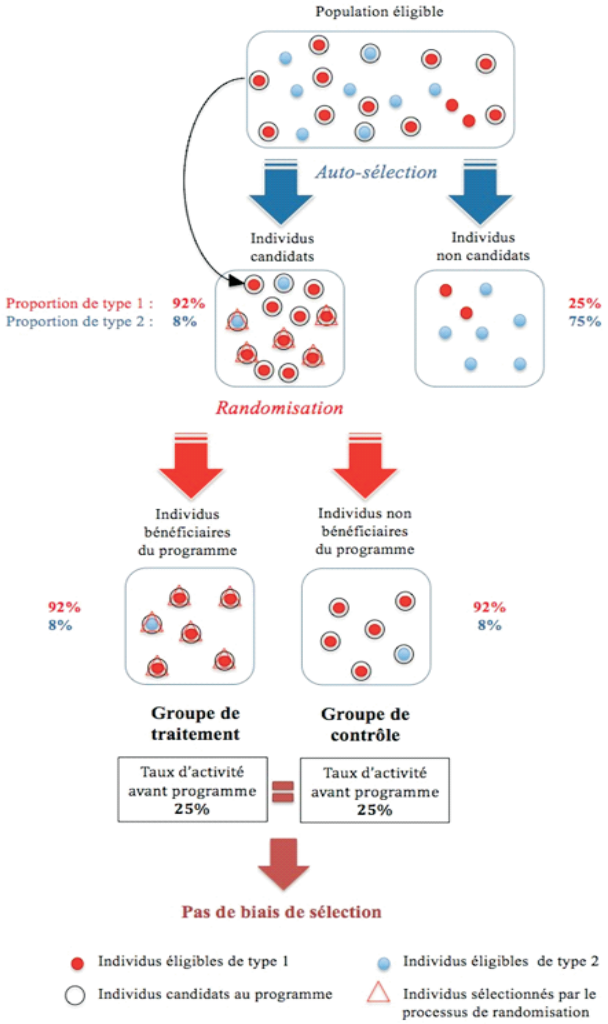
Avantages et inconvénients du *design* 2

Ce type d'expérience randomisée effectue la comparaison après l'auto-sélection et permet ainsi de ne pas imposer de traitement à une population qui

Encadré 2 : estimation de l'impact du programme de formation professionnelle JTPA (Bloom *et alii*, 1997)

Le *Job Training Partnership Act* (JTPA) a été soumis à une évaluation randomisée du type du *design* 2. Il s'agit d'une expérience randomisée proposant un programme de formation professionnelle à 21 000 individus éligibles. Tout d'abord, sont éligibles les individus en situation économique difficile ou les jeunes ayant quitté le système scolaire. Dans un second temps, les individus demandant le programme sont alloués de façon aléatoire à un groupe de traitement (2/3 des recourants) et un groupe de contrôle (1/3 des recourants). La nature aléatoire de cette allocation garantit que la comparaison des deux groupes représente l'impact causal du programme de formation professionnelle sur les recourants. Au sein de la population adulte, la différence de revenus entre ces deux groupes représente l'impact causal du programme. Au sein de la population féminine, cette différence est de 1 176 $ sur 30 mois, soit une augmentation de revenus de 9,6 %. Concernant les hommes, la différence de revenu due au programme s'élève à 978 $ sur 30 mois, soit une augmentation de 5,3 %. Les auteurs trouvent que 32 % des individus ayant abandonné l'école au sein du groupe de traitement ont obtenu un diplôme dans la période de 30 mois suivant l'expérience, contre seulement 20,4 % au sein du groupe de contrôle. Cela signifie que le programme augmente de 11,6 points le pourcentage d'individus obtenant un diplôme dans les 30 mois. Notons que la mise en place de ce programme se fait à travers 16 centres de formation professionnelle sélectionnés sur la base du volontariat. Certains centres ont refusé d'y participer à cause de la nature aléatoire de cette expérience, ce qui peut engendrer un biais de sélection. En effet, la décision de ces centres de participer ou non peut être basée sur des facteurs corrélés avec l'effet du programme.

Figure 5 : *design* 2 : randomisation après auto-sélection, exemple d'un programme de formation professionnelle



ne le souhaite pas. Cette méthode est particulièrement judicieuse dans le cadre de l'évaluation de programmes sociaux. Pour le RSA par exemple, près de 50 % des individus éligibles n'y ont pas recours. Le *design* 1 donnant l'effet moyen sur toute la population éligible ne serait donc pas très pertinent.

Cependant la critique majeure de cette méthode est que, dans un premier temps, elle offre le programme à tous les éligibles puis, une fois que ceux qui veulent y recourir font la démarche de candidature, une certaine proportion est informée qu'en réalité elle n'y a pas accès. Cela pose bien évidemment des problèmes d'ordres éthique et politique. Le *design* 3 présenté ci-après répond à ce problème en proposant le traitement seulement à ceux qui pourront véritablement y avoir accès.

Tableau 3 : *design* 2 : randomisation après auto-sélection – exemple numérique d'un programme de formation professionnelle

	Taux d'activité avant programme ($t = 0$)	Taux d'activité après programme ($t = 1$)		Impact causal moyen du programme	Taux d'auto-sélection
		Traité (avec formation)	Non-traité (sans formation)		
Individus de type 1	0,3	0,5	0,2	0,3	0,8
Individus de type 2	0,9	0,9	0,8	0,1	0,1

Impact du contexte socio-économique sur le taux d'activité des individus (biais de conjoncture)	-0,1

Proportion de type 1 dans la population	0,6
Proportion de type 1 dans la population	0,4

Design 2 : randomisation après auto-selection			
	Groupe des recourants		Groupe des non recourants
Proportion de type 1 après auto-sélection ($t = 0$)	0,92		0,25
Proportion de type 2 après auto-sélection ($t = 0$)	0,08		0,75
Taux d'activité après auto-sélection ($t = 0$)	0,35		0,75
	Groupe de traitement	Groupe de contrôle	
Proportion de type 1 après randomisation ($t = 1$)	0,92	0,92	0,25
Proportion de type 2 après randomisation ($t = 1$)	0,08	0,08	0,75
Taux d'activité avec programme ($t = 1$)	0,53	0,53	0,80
Taux d'activité sans programme ($t = 1$)	0,25	0,25	0,65
Impact causal moyen réel sur les recourrants (ICMR)	0,28		
Impact causal moyen estimé sur les recourrants (ICMR estimé)	0,28		
Biais de sélection	0,00		
Contrefactuel	Individus candidats non traités en $t = 1$		

	= Non observable

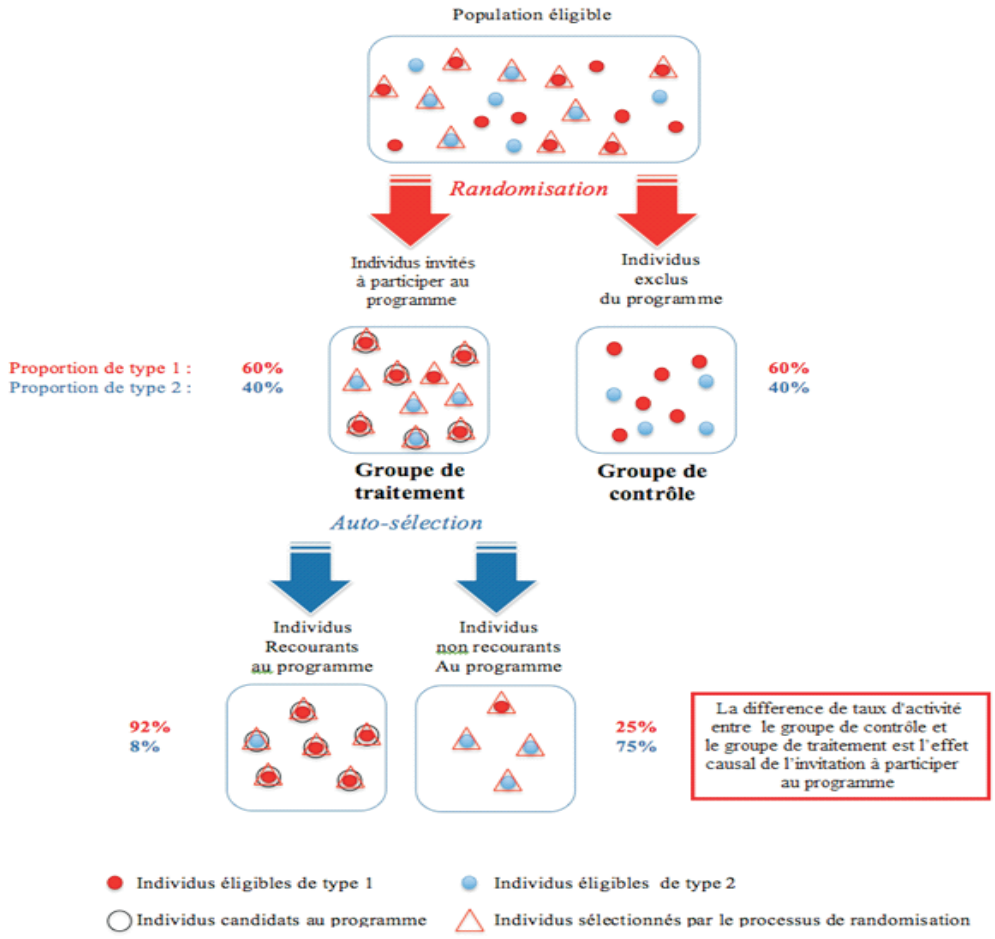
Design 3 : randomisation de l'accès au traitement

Une troisième façon d'utiliser la randomisation consiste à distribuer aléatoirement l'accès au programme au sein de la population éligible. La figure 6 schématise ce troisième *design* et le tableau 4 reprend de nouveau l'exemple numérique du programme de formation professionnelle. Le mécanisme est le suivant. Tout d'abord, deux groupes sont formés aléatoirement à partir d'un échantillon d'individus : un groupe qui aura accès au programme et un groupe qui n'y aura pas accès. Dans un second temps, on informe les individus du groupe ayant accès au programme qu'ils peuvent candidater et on obtient ainsi deux sous-groupes, un groupe de recourants qui entrera dans le programme et un groupe de non-recourants comme présenté figure 6.

Le groupe de contrôle et le groupe de traitement ont donc des caractéristiques identiques (même composition de types : 60 % de type 1 et 40 % de type 2) et ne diffèrent que par l'accès au traitement. Ce *design* d'expérience randomisée permet d'obtenir deux impacts différents d'un programme :

– **l'impact causal moyen de l'accès au traitement (ICMA)**. Dans le cadre d'un programme de formation professionnelle, ce paramètre est donné par la différence de taux d'activité entre le groupe de traitement et le groupe de contrôle. Dans le tableau 4, cela correspond à la différence (4) - (5) c'est-à-dire à 0,59-0,44 = 0,15. Proposer un programme de formation professionnelle à une population augmente ainsi le taux d'activité de 15 points de pourcentage. Cette augmentation de 15 points de

Figure 6 : *design* 3 : randomisation de l'accès au traitement : exemple d'un programme de formation



pourcentage représente l'effet de l'accès au programme et non l'effet du programme. En effet, au sein du groupe de traitement, certains individus ont recours au programme et d'autres non ;

– **l'impact causal moyen sur les recourants (ICMR)**. Ce *design* d'expérience randomisée permet également d'estimer l'ICMR qui constitue le paramètre le plus pertinent. On retrouve ce paramètre en divisant l'impact causal moyen de l'accès au traitement par la proportion de recourants. Pour retrouver l'ICMR, il est donc nécessaire de connaître la proportion de recourants. Par ailleurs, cet estimateur est valide sous l'hypothèse que le simple fait de proposer un programme de formation n'impacte pas directement le taux d'activité. Autrement dit, l'accès au traitement ne doit pas avoir d'effet direct sur le résultat d'intérêt. Dans notre exemple numérique, le programme de formation professionnelle augmente en moyenne de 28 points de pourcentage le taux d'activité de ceux qui y auront recours, soit 0,15/0,52=0,28, l'ICMA divisé par la proportion de recourants.

Avantages et inconvénients du design *3*

Cette méthode permet de ne pas obliger les individus à entrer dans un programme. Dans notre exemple il s'agit de ne pas forcer les individus à suivre une formation alors qu'ils ne le souhaitent pas. Un problème persistant est que le traitement n'est pas accessible au groupe de contrôle. Cependant, par les problèmes éthiques qu'elle résout, cette méthode est plus favorable que le *design* 2 qui propose un traitement à des individus auxquels il est ensuite interdit de participer une fois qu'ils ont accepté. Dans ce *design* les individus invités à participer sont libres de décider et ne reçoivent pas d'informations contradictoires. Le *design* suivant se propose de résoudre le problème de l'accès limité au traitement à une partie de la population.

Design 4 : randomisation d'un encouragement

Il existe une façon alternative d'utiliser la randomisation qui consiste à distribuer aléatoirement un **encouragement** à une population et à donner l'accès au traitement à toute la population. Un encouragement représente une

Tableau 4 : *design* 3 : randomisation de l'accès au traitement – exemple numérique d'un programme de formation professionnelle

	Taux d'activité avant programme ($t = 0$)	Taux d'activité après programme ($t = 1$)		Impact causal moyen du programme	Taux d'auto-sélection
		Traité (avec formation)	Non-traité (sans formation)		
Individus de type 1	0,3	0,5	0,2	0,3	0,8
Individus de type 2	0,9	0,9	0,8	0,1	0,1

Impact du contexte socio-économique sur le taux d'activité des individus	-0,1

Proportionde type 1 dans la population	0,6
Proportionde type 2 dans la population	0,4

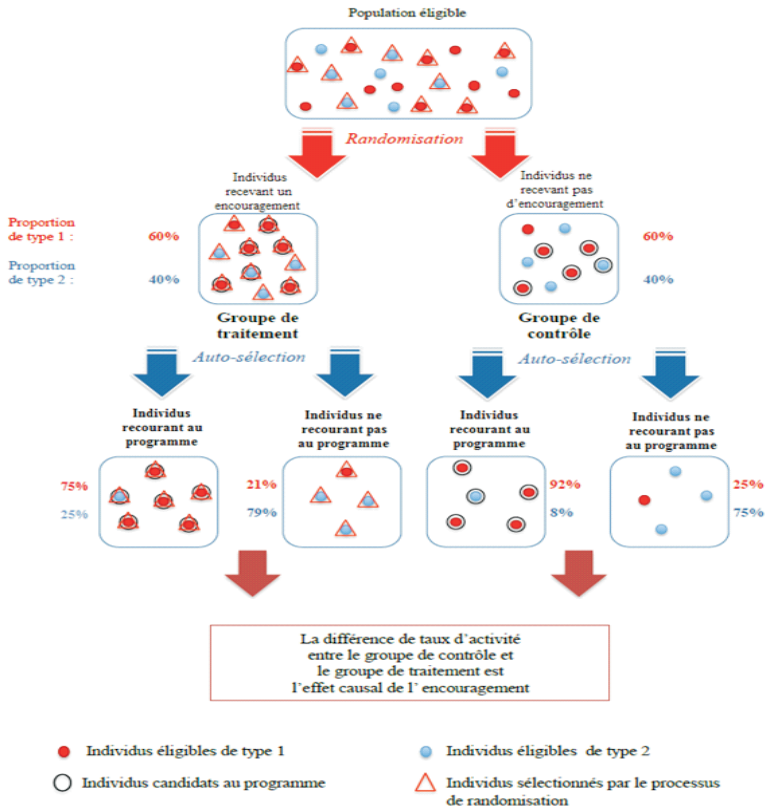
Desing 3 : randomisation de l'accès au programme			
	Groupe de traitement (accès libre)		Groupe de contrôle (accès bloqué)
Proportion de type 1 après randomisation ($t = 0$)	0,60		0,60
Proportion de type 2 après randomisation ($t = 0$)	0,40		0,40
Taux d'activité après randomisation ($t = 0$)	0,54		0,54
	Groupe de recourants	Groupe de non recourants	
Proportion de type 1 après auto-sélection ($t = 1$)	0,92	0,25	0,6
Proportion de type 2 après auto-sélection ($t = 1$)	0,08	0,75	0,4
Taux d'activité avec programme ($t = 1$)	0,53(1)	0,80	0,66
Taux d'activité sans programme ($t = 1$)	0,25(2)	0,65	0,44
Proportion de recourrants	0,52(3)		0,00
Taux d'activité moyen observé	0,59(4)		0,44(5)
Impact moyen de l'offre du traitement (ICMA) (4) - (5)	0,15(6)		
Impact causal moyen réel sur les recourrants (ICMR réel) (1) - (2)		0,28	
Impact causal moyen estimé sur les recourrants (ICMR estimé) (6) / (3)		0,28	
Contrefactuel	Individus candidats non traités en $t = 1$		

= Non observable

incitation à entrer dans le traitement évalué : envoyer un courrier d'information sur une politique sociale, organiser des réunions d'information ou encore donner une incitation financière ou matérielle (Banerjee et Duflo, 2009). Cet encouragement doit être choisi de sorte à n'influer sur le résultat d'intérêt (par exemple, le taux d'activité) qu'à travers la participation au programme. Autrement dit, l'encouragement ne doit pas avoir d'effet direct sur le résultat d'intérêt. Ce *design* de randomisation est schématisé par la figure 7.

Dans le cas d'un programme de formation professionnelle, comme le présente numériquement le tableau 5, un encouragement peut être un courrier postal informatif (condition, organisation, fiche d'inscription, etc.) ou encore une suggestion par un conseiller de Pôle Emploi. Cet encouragement est alloué de façon aléatoire au sein d'une population ce qui permet d'obtenir deux groupes : un groupe de traitement ayant accès au programme et recevant un encouragement et un groupe de contrôle ayant accès au programme mais ne recevant pas d'encouragement. La randomisation garantit que ces

Figure 7 : *design* 4 : randomisation d'un encouragement : exemple d'un programme de formation professionnelle



deux groupes comprennent les mêmes proportions d'individus de type 1 et d'individus de type 2 (60 % et 40 %) et ne diffèrent donc que par l'allocation de l'encouragement. Dans un second temps les individus de chaque groupe choisissent d'entrer ou non dans le programme de formation. Ce design d'expérience randomisée permet d'obtenir deux impacts différents d'un programme :

– **l'impact causal moyen de l'encouragement (ICME)**. Le groupe de contrôle et le groupe de traitement ne diffèrent que par la réception d'un encouragement. Ainsi la différence *ex post* du paramètre d'intérêt entre les deux groupes donne l'impact causal de l'encouragement. Dans le cadre de notre exemple, envoyer un courrier informatif sur un programme de formation à une population augmentera le taux d'activité moyen de 3 points de pourcentage ;

– **l'impact causal moyen du traitement sur les *compliers* (ICMC)**. Les *compliers* sont les individus qui n'ont pas recours au programme en l'absence d'encouragement mais qui y ont recours lorsqu'ils reçoivent un encouragement. Sous l'hypothèse de l'absence de *defiers*, c'est-à-dire d'individus qui décident de ne pas recourir au programme après avoir reçu l'encouragement, mais qui auraient recouru au programme en l'absence de l'encouragement, l'écart de niveau du taux d'activité, entre le groupe de contrôle et le groupe de traitement est dû uniquement aux compliers. On peut obtenir l'impact causal du traitement sur les *compliers* en divisant l'impact

Encadré 3 : randomisation d'un suivi renforcé par l'ANPE

Behaghel, Crépon et Gurgand (2014) présentent les résultats d'une large expérience randomisée réalisée en France, et organisée comme suit. Les individus éligibles pour cette expérience sont les demandeurs d'emploi. Trois groupes ont été créés : (1) un groupe de contrôle constitué d'individus ayant un suivi classique de l'ANPE, (2) un premier groupe de traitement ayant droit à un suivi complémentaire géré par un programme public et (3) un second groupe de traitement ayant un suivi complémentaire géré par un programme privé. La randomisation a eu lieu lors du premier rendez-vous à l'ANPE durant lequel un employé, grâce à une application informatique, détermine aléatoirement le groupe auquel le demandeur d'emploi était assigné. Un demandeur d'emploi assigné au suivi renforcé est libre de le refuser. Un demandeur d'emploi assigné au suivi classique peut demander de bénéficier d'un suivi renforcé. Cette expérimentation adopte donc bien le *design* 4. Néanmoins, seule une infime partie (entre 1 % et 3,8 %) des demandeurs d'emploi assignés au suivi classique choisit de bénéficier du suivi renforcé. En pratique, cette expérimentation s'apparente donc au *design* 3 : la population de *compliers* constitue quasiment l'ensemble de la population des recourants. Les auteurs trouvent ainsi qu'un suivi plus important améliore l'accès à l'emploi : par exemple, un suivi complémentaire augmente de 4 à 9 points de pourcentage le taux d'emploi après six mois. Les impacts sont significativement différents pour les traitements (1) et (2). Le programme public est plus efficace.

Tableau 5 : *design* 4 : randomisation d'un encouragement – exemple numérique d'un programme de formation professionnelle

	Taux d'activité avant programme ($t = 0$)	Taux d'activité après programme ($t = 1$)		Impact causal moyen du programme	Proportion de *switchers*	Taux d'auto-sélection
		Traité (avec formation)	Non-traité (sans formation)			
Individus de type 1	0,3	0,5	0,2	0,3	0,1	0,8
Individus de type 2	0,9	0,9	0,8	0,1	0,35	0,1

Impact du contexte socio-économique sur le taux d'activité des individus	-0,1

Proportion de type 1 dans la population	0,6
Proportion de type 1 dans la population	0,4

Design 4 : randomisation d'un encouragement				
	Groupe de traitement (*réception d'un encouragement*)		Groupe de contrôle (*pas d'encouragement*)	
Proportion de type 1 après randomisation ($t = 0$)	0,60		0,60	
Proportion de type 2 après randomisation ($t = 0$)	0,40		0,40	
Taux d'activité après randomisation ($t = 0$)	0,54		0,54	
	Groupe de recourants	Groupe de non-recourants	Groupe de recourants	Groupe de non-recourants
Proportion de type 1 après auto-sélection ($t = 1$)	0,75	0,21	0,92	0,25
Proportion de type 2 après auto-sélection ($t = 1$)	0,25	0,79	0,08	0,75
Proportion de recourants	0,72		0,52	
	0,62			
Taux d'activité avec programme ($t = 1$)	0,60	0,81	0,53	0,80
Taux d'activité sans programme ($t = 1$)	0,35	0,67	0,25	0,65
Taux d'activité moyen	0,62		0,59	
Impact moyen de l'encouragement	0,03			
Impact moyen du programme sur les *switchers*	0,15			
Impact causal moyen réel sur les recourants (ICMR réel)	0,25		0,28	

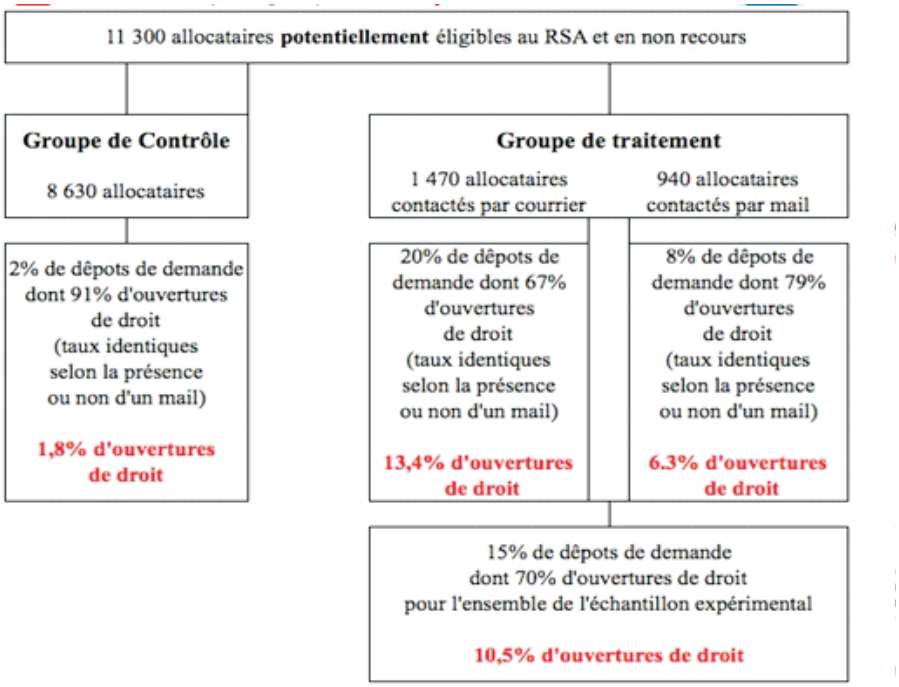
= Non observable

causal moyen de l'encouragement (ICME) par la proportion de *compliers*. L'estimation de ce paramètre suppose donc de connaître la proportion de *compliers* dans la population (Angrist, Imbens et Rubin, 1996).

Dans notre exemple, participer au programme de formation professionnelle augmente de 15 points le taux d'activité moyen des *compliers*, c'est-à-dire l'ICME (0,03) divisé par la proportion de *compliers* (*i.e.*, la proportion de *compliers* est obtenue par le calcul suivant : 0,1*0,6 + 0,35*0,4). Il a été choisi arbitrairement que les individus de type 2 comportaient davantage de *compliers* (35 %) que les individus de type 1 (10 %). En ajustant par les proportions de chaque type on obtient 20 % de *compliers* dans la population. Dans la réalité ces chiffres ne sont pas directement observables ; mais la proportion de compliers peut se retrouver en calculant la différence de proportion de recourants entre le groupe de contrôle et le groupe de traitement (*i.e.*, 0,72 - 0,52 = 0,20).

Encadré 4 : randomisation d'un encouragement en Gironde (Okbani, 2013)

Le pourcentage d'individus ne recourant pas au RSA étant très élevé, la Caf de la Gironde a mené une expérience afin d'évaluer l'impact de différents encouragements (emails, SMS et courriers postaux informatifs) sur le taux de participation au RSA. *In fine* trois groupes ont été formés : un groupe de traitement constitué d'individus recevant un encouragement, mail ou courrier, et un groupe de contrôle. Ces groupes ont été construits par tirage aléatoire. Plus précisément, les expérimentateurs ont procédé à un tirage aléatoire par rapport à plusieurs variables de stratification : situation familiale, estimation des montants des droits au RSA et points d'accueil physiques de rattachement. Le groupe de traitement fut scindé en deux groupes. Le premier groupe était constitué des individus ayant communiqué leur adresse électronique à la Caf et dont l'encouragement est donc l'envoi d'un courrier électronique. Le second groupe comprenait les individus, qui n'ayant pas communiqué d'adresse électronique, ont reçu un courrier postal. Les résultats suggèrent que la différence de taux d'ouverture de droit entre le groupe de contrôle et le groupe de traitement est très grande. L'envoi d'un encouragement augmente de 8,7 points de pourcentage le taux d'ouverture de droit. Le taux de recours au sein du groupe de contrôle est identique que les individus aient déclaré ou non une adresse électronique. Cela suggère que la différence de taux de recours entre les deux groupes de traitement est due à la différence d'encouragement et non à une différence de comportement entre les deux groupes face au recours au RSA, c'est-à-dire un problème de sélection.



Source : Rapport intermédiare 2010 du Comité national d'évaluation du RSA. *La documentation française*, Paris, Janvier 2011, Annexe 2

Dans le tableau 5, l'impact causal du traitement sur les recourants (ICMR réel) est renseigné alors qu'il n'est pas observable dans la réalité. On observe que le programme de formation a un impact moyen plus fort sur les recourants du groupe sans encouragement que sur ceux du groupe avec encouragement (0,28 *versus* 0,25) ainsi qu'un impact plus fort sur les recourants que sur les *compliers*. Ceci est dû au fait que l'encouragement, dans notre exemple, affecte davantage les individus de type 2, qui, rappelons-le, sont moins affectés par le programme. Le groupe de *compliers* comprend donc davantage de types 2 ce qui réduit l'impact du programme sur le taux d'activité.

Avantages et inconvénients du design *4*

Cette méthode a l'avantage de ne pas contraindre l'accès au traitement. Elle consiste à évaluer l'impact d'un encouragement puis à isoler l'effet du programme sur les *compliers*. La limite majeure de ce *design* est qu'il permet d'évaluer l'effet d'un programme seulement sur la population limitée des *compliers* et non sur les recourants. Cependant, ce *design* est adapté aux politiques qui visent à augmenter le taux de participation à un programme donné et qui ciblent précisément les éventuels *compliers*. Dans le cas du RSA, cette méthode apparaît particulièrement pertinente. En effet, en 2011, 36 % des individus éligibles au RSA socle (équivalent du RMI) et 68 % des éligibles au RSA activité (revenus complémentaires versés aux travailleurs pauvres) n'ont pas demandé à en bénéficier (Comité national d'évaluation du RSA, 2011).

Encadré 5 : PROGRESA
(*Programa Nacional de Education, Salud y Alimentacion*)

PROGRESA est un programme social de lutte contre la pauvreté mis en place en 1997 par le gouvernement mexicain et dont l'évaluation a pris la forme d'une expérience randomisée. L'intervention visait tout d'abord les zones rurales et plus tard a été étendu aux zones semi-rurales puis urbaines. En 1999 PROGRESA touchait 24 000 ménages et représentait 40 % du budget de lutte contre la pauvreté du gouvernement. En 2012, il touchait 5,8 millions de ménages et est aujourd'hui mis en place à grande échelle sous l'appellation *Oportunidades*. Le but de PROGRESA était de développer le capital humain du Mexique en donnant aux individus en dessous d'un certain seuil de pauvreté, représentant 2/3 de la population, des aides matérielles et financières afin de réduire le niveau de pauvreté et de les inciter à investir dans l'éducation, la santé et la nutrition. PROGRESA proposait ce que l'on appelle des transferts monétaires conditionnels. Plus précisément, afin de promouvoir l'éducation, des transferts financiers ont été attribués aux mères éligibles, conditionnellement au fait que leurs enfants assistent à 85 % des jours d'école de l'année. Les filles ayant une probabilité plus élevée que les garçons de quitter l'école, les subventions étaient plus élevées si l'enfant concerné était une fille. Le programme PROGRESA proposait également des interventions préventives pour l'hygiène et la santé en fournissant des suppléments nutritionnels aux jeunes enfants et aux femmes enceintes, lesquels avaient également droit à des subventions supplémentaires pour de la nourriture. Lors de l'évaluation de ce programme, la randomisation s'est effectuée au niveau des localités et non au niveau des ménages, ce qui aurait pu engendrer un biais de diffusion. Tout d'abord, 506 localités ont été choisies aléatoirement pour participer à l'expérience, soient 24 077 ménages. PROGRESA étant un programme visant les zones les plus pauvres, le tirage aléatoire a été exécuté sur la base d'un système de pondération favorisant les localités les plus pauvres. Sur les 506 localités sélectionnées, deux groupes furent formés aléatoirement. Le groupe de contrôle comptait 186 localités alors que les 320 localités qui eurent accès au programme formaient le groupe de traitement, soit une probabilité de 60 % de recevoir le programme. Par souci d'équité, alors que le groupe de traitement eut accès au programme en mai 1998, le groupe de contrôle y eut accès un peu plus d'un an plus tard en 1999. Ce programme, étudié par de nombreux économistes, semble avoir eu des impacts substantiels sur le bien-être et le capital humain des familles pauvres. La présence des enfants à l'école a augmenté de façon significative et particulièrement pour les filles avec une augmentation de 14 points de pourcentage de la présence à l'école, soit 0,7 année d'éducation supplémentaire. Les résultats indiquent par ailleurs que cette hausse de niveau d'éducation a augmenté de 8 % le revenu futur des enfants touchés par PROGRESA. Le programme semble également avoir des impacts bénéfiques sur la santé : la probabilité de tomber malade a été réduite de 12 points de pourcentage pour les enfants et de 19 points pour les adultes. Les ménages traités semblent aussi mieux s'alimenter.

Design 5 : randomisation de l'ordre d'allocation du traitement

Il existe une autre forme de randomisation permettant de ne pas restreindre l'accès au traitement et qui permet cependant l'estimation d'impacts causaux plus pertinents. Il s'agit des **expériences randomisées progressives**. Elles attribuent le traitement successivement à tous les groupes en randomisant l'ordre d'allocation du traitement.

Un exemple de randomisation progressive est le programme social mexicain PROGRESA qui verse des subventions aux mères pauvres dans le but d'augmenter le niveau d'éducation des enfants. L'encadré 5 décrit plus en détail ce programme. Une distorsion induite par ce *design* est que les individus sachant qu'ils auront droit au traitement peuvent être incités à modifier leurs comportements. Cela altère la validité du groupe de contrôle. On parle de **biais d'anticipation**. D'autre part, bien que cette méthode permette à toute la population de recevoir le traitement, elle peut causer des problèmes quant à l'évaluation des effets de long terme du programme. En effet une distribution trop rapide du traitement au groupe de contrôle peut faire obstacle à l'identification de l'effet de la politique au sein du groupe de traitement qui prendra un certain temps avant d'être observable. Ainsi le traitement ne doit pas être distribué trop rapidement entre les différents groupes afin que l'effet du traitement ait le temps de se matérialiser et qu'il puisse être évalué (Duflo, Glennerster et Kremer, 2008).

Problèmes et biais potentiels des expériences randomisées

Cette partie résume tout d'abord les problèmes éthiques soulevés par les différents *design*s d'expérimentation puis elle passe en revue les différentes menaces à la validité d'une l'expérimentation évoquées dans la littérature économique ainsi que les solutions existantes pour les contourner. On distingue ici deux types de validité d'une expérience randomisée : la validité interne et la validité externe :

– **validité interne** : une expérience randomisée a une bonne validité interne si l'impact estimé représente bien l'effet causal réel du traitement sur le paramètre d'intérêt au sein de la population expérimentale ;

– **validité externe** : une expérience randomisée a une bonne validité externe si l'impact estimé représente bien l'effet causal réel du traitement sur le paramètre d'intérêt au sein de la population visée par la

politique, et donc si l'impact peut être raisonnablement généralisé à d'autres populations que celles de l'expérience.

Problèmes éthiques et politiques

L'expérimentation randomisée implique une rupture du principe d'égalité des citoyens devant la loi. Des individus sélectionnés aléatoirement vont bénéficier d'une politique publique, alors que les individus placés dans le groupe de contrôle ne vont pas avoir cette opportunité. La loi constitutionnelle du 28 mars 2003 a ouvert la possibilité de déroger au principe d'égalité de traitement dans les lois et les règlements (art. 37-1) et pour les collectivités territoriales (art. 72, alinéa 4) à des fins d'expérimentation. Même lorsque le cadre juridique existe, il semble difficile de donner l'accès à une politique potentiellement bénéfique à une partie restreinte de la population. En pratique, l'expérimentation randomisée semble plus facile à faire accepter dans deux situations : lorsque les bénéfices du programme évalué par rapport à la situation de référence sont incertains (opérateurs de placement privés ou public) et lorsque l'accès à la politique est limité par nature.

L'expérimentation randomisée pose aussi la question épineuse du consentement des participants. D'une part, il n'est pas toujours possible ou souhaitable d'obtenir un tel consentement. D'autre part, l'obtention d'un consentement peut biaiser l'estimation de l'effet de la politique, la population donnant son aval à l'expérimentation n'étant généralement pas représentative de la population bénéficiaire (Behaghel, Crépon et Le Barbanchon, 2015 ; Sianiesi, 2017). Les écueils éthiques liés aux expérimentations randomisées ont entrainé le développement d'institutions spécifiques comme par exemple les comités éthiques au sein des universités qui évaluent si les projets expérimentaux respectent les bonnes pratiques.

Menaces à la validité interne

Attrition : l'attrition est le phénomène selon lequel des individus quittent l'expérience avant qu'elle ne soit terminée. Des individus peuvent décider de ne plus recevoir le traitement ou de ne pas répondre aux questionnaires réalisés. Lors d'expériences se déroulant sur plusieurs années il arrive que les chercheurs perdent la trace de certains individus, par exemple à cause de déménagements. Cela constitue un problème majeur car l'attrition n'est pas aléatoire : les individus quittant l'expérience ont en général des caractéristiques spécifiques, ce qui peut fausser la randomisation initiale et donc biaiser l'estimation de l'effet du traitement.

Solution : les solutions disponibles concernent typiquement l'organisation même de l'expérience : maintenir un contact régulier avec les sujets, avoir le soutien de la ville ou du gouvernement pour encadrer l'expérience ou pour l'accès à des données administratives. Des méthodes statistiques permettent également de tenir compte de l'attrition, par exemple en estimant les bornes minimale et maximale de l'impact d'un programme (Behaghel, Crépon, Gurgand et Le Barbanchon, 2015).

Biais de diffusion : un biais de diffusion émerge lorsque des individus assignés au groupe de contrôle sont affectés, positivement ou négativement, par le traitement. Il s'agit d'un problème majeur pouvant biaiser l'estimation de l'impact causal d'un traitement parce que la structure du groupe de contrôle est modifiée et dès lors ne constitue plus un bon contrefactuel.

Exemple 1 *: un exemple de biais de diffusion est celui émergeant dans les études évaluant l'impact d'une campagne de vaccination. La vaccination implique de larges externalités positives : un individu vacciné ne peut pas contaminer d'autres individus. Le traitement peut donc améliorer la santé des individus du groupe de contrôle de façon indirecte qui, autrement auraient pu être potentiellement contaminés par les individus du groupe de traitement.*

Exemple 2 *: un programme de formation impacte le marché du travail. Un individu retrouve un emploi grâce au programme mais au détriment d'un autre travailleur qui en l'absence du programme aurait pu occuper cet emploi. L'effet net d'un programme de formation à grande échelle sur le marché du travail devient incertain. Ce point renvoi à la notion familière en économie d'équilibre général.*

Solution : les groupes de contrôle et les groupes de traitement peuvent être espacés géographiquement pour éviter les effets de diffusion. Il s'agit d'allouer le traitement non pas au niveau individuel au sein d'une même ville mais dans différentes zones où il n'y a pas d'interaction affectant le paramètre d'intérêt, comme deux marchés distincts par exemple.

Ainsi, pour évaluer l'effet d'un programme de vaccination dans des écoles, la randomisation peut s'établir au niveau des écoles et non au niveau des élèves (Miguel et Kremer, 2001). Une étude (Crépon *et alii*, 2013) analysant l'impact d'un programme de formation professionnelle offert à des jeunes demandeurs d'emploi qualifiés en France, évoque le jeu des « chaises musicales » pour illustrer le problème de diffusion. Cet article se propose d'évaluer cet effet de diffusion en allouant un programme de formation à des proportions aléatoires de demandeurs d'emplois dans des zones géographiques distinctes. Les auteurs évaluent tout d'abord l'effet du programme sur les individus traités et trouvent que les individus étant au chômage au début de l'étude ont une probabilité de 11 % supérieure de retrouver un CDD et de 4 % supérieure de retrouver un CDI que les individus au chômage non traités. Ils comparent ensuite les travailleurs non

traités des zones expérimentales aux travailleurs non traités des zones non expérimentales. Ils trouvent que les individus non traités des zones expérimentales ont une probabilité de moindre de 2,1 points de pourcentage de trouver un emploi stable et que l'effet total du programme est nul.

Biais de substitution : il est fréquent, lors d'une expérience, que les individus se retrouvant dans le groupe de contrôle recherchent un substitut au programme évalué alors qu'ils ne l'auraient pas recherché en l'absence d'expérience. Ce phénomène distord le niveau du paramètre d'intérêt dans le groupe de contrôle, qui, dès lors, ne constitue plus un bon groupe de référence.

Exemple 1 : *des individus non assignés au groupe de traitement d'un programme de formation, en apprenant l'existence de ce programme, peuvent rechercher des programmes alternatifs qu'ils n'auraient pas demandé en absence d'une expérience randomisée. Ce comportement mènerait à une sous-estimation du programme évalué à cause d'un groupe de contrôle dont le niveau moyen d'activité augmenterait en raison de nouveaux bénéficiaires de formation. Heckman, LaLonde et Smith (1999) reportent que jusqu'à 40 % des membres de groupes de contrôle ont accès à des programmes de substitution.*

Exemple 2 : *l'évaluation d'un programme de vaccination par expérimentation randomisée peut inciter les parents dont les enfants ne sont pas assignés au groupe de traitement à vacciner leurs enfants en passant par un autre organisme, parce qu'ils ont pu se rendre compte du rôle essentiel de la vaccination ou par un effet de mimétisme par exemple. Ces comportements augmenteraient le niveau de vaccination dans le groupe de contrôle et ainsi l'impact causal du traitement serait sous-estimé.*

Solution : on peut procéder à une expérience randomisée progressive. Les individus, sachant qu'ils auront accès au traitement plus tard, ont moins d'incitation immédiate à rechercher un substitut.

Effet de l'expérimentateur : il s'agit de l'effet produit par les expérimentateurs sur les sujets. Une expérience randomisée passe naturellement par la rencontre de l'expérimentateur avec les sujets. Cette rencontre peut affecter les comportements des sujets. Rosenthal explique par exemple que l'expérimentateur peut influencer les sujets tant par une communication verbale que non verbale, c'est-à-dire par la gestuelle, le ton de voix ou encore les expressions du visage qui traduiraient son attente vis-à-vis des comportements attendus (Rosenthal, 1966).

Solution : l'idée est de réfléchir sur une mise en œuvre qui permet aux expérimentateurs d'être le plus neutre possible.

Durée limitée d'une expérience : il est possible que certains effets ne se matérialisent qu'après une certaine période de temps. Afin de pouvoir estimer correctement un impact causal avec une expérience randomisée, il est dans ce cas nécessaire que l'expérimentation dure suffisamment longtemps. Par exemple, une politique de subvention sur l'achat de biens durables comme dans le domaine de l'immobilier ne peut pas être évaluée sur une période très courte puisque ces décisions demandent du temps aux sujets. Par ailleurs une expérience se déroulant sur une période trop courte peut distordre les comportements individuels. Sachant que le traitement sera de courte durée, des individus peuvent renoncer au traitement ou peuvent avoir des comportements différents vis-à-vis des paramètres étudiés, comme faire plus ou moins d'efforts.

Solution : la solution évidente est de faire durer l'expérimentation suffisamment longtemps, au moins sur une partie de l'échantillon afin de pouvoir mesurer l'effet spécifique de la durée de l'expérimentation.

Effet d'Hawthorne et effet de John Henry : le fait qu'un individu sache qu'il fait partie d'une expérience suffit pour modifier son comportement. Lorsqu'un individu du groupe de traitement modifie son comportement, on parle d'effet d'Hawthorne. Cette dénomination provient d'une usine, la *Hawthorne Works*, dans laquelle des chercheurs faisant des expériences sur la productivité du travail se sont aperçus que quelles que soient les conditions de travail testées, la productivité des travailleurs augmentait systématiquement. Cela peut être lié à un effet psychologique relatif à l'estime de soi, ou à la recherche d'une forme de reconnaissance. Dans ce cas, l'effet estimé sera donc l'impact causal réel de la politique plus l'effet expérimental, ce qui altère l'évaluation de la politique.

On parle d'effet de John Henry lorsque les individus du groupe de contrôle changent de comportement. Les individus du groupe de contrôle peuvent par exemple se sentir frustrés ou offensés de se voir refuser la politique ce qui peut affecter leurs décisions. Cet effet peut aller dans différentes directions : les individus peuvent se trouver davantage déterminés ou bien au contraire peuvent avoir un comportement de révolte les incitant à faire un moindre effort au regard des paramètres étudiés. Cet effet biaise ainsi l'estimation de l'impact causal de la politique. L'appellation de ce dernier effet provient du personnage folklorique américain John Henry, un ouvrier du XIX^{ème} siècle, employé pour le travail épuisant de la construction d'un chemin de fer. Lorsque l'innovation du marteau pilon a mené à remplacer la main-d'œuvre, révolution dans le

monde métallurgique, John Henry, afin de sauver les ouvriers, a fait le pari avec son employeur qu'il pouvait être aussi efficace qu'un marteau pilon. Il gagna son pari à l'issue duquel il décéda et devint ainsi le héros de la lutte contre le progrès technique. Un cadre expérimental peut causer une distorsion de comportement du groupe de contrôle et se traduit le plus souvent par un excès de motivation, d'où le nom d'effet de John Henry.

Solution : une solution est de récolter des données sur une plus longue période afin d'effacer ces effets qui ont tendance à se dissiper dans le temps. Une autre solution peut également se trouver dans le *design* même de l'expérience. Par exemple Ashraf, Karlan et Yin (2006) ont analysé l'impact d'un système d'épargne en utilisant un *design* d'expérience randomisée comprenant trois groupes : un groupe de traitement recevant des visites pour les encourager à épargner *via* le système évalué, un groupe de contrôle et un troisième groupe d'individus recevant le même type de visites d'encouragement que le groupe de traitement mais se limitant aux anciens produits d'épargne, c'est-à-dire des individus auxquels n'est pas proposé le nouveau système d'épargne. De façon générale une solution souvent utilisée est la réalisation d'un *design* placebo en plus de l'intervention étudiée. Il s'agit de répliquer au sein d'un groupe l'intervention expérimentale à l'identique excepté l'attribution du traitement. Cette méthode permet de dégager l'effet dû à la démarche expérimentale de l'effet réel produit par le traitement.

Menaces à la validité externe

Indépendance environnementale : les résultats d'une expérience peuvent-ils être généralisés et applicables hors de la population expérimentale ? L'indépendance environnementale d'une expérience dépend du *design* de l'expérience, de l'échantillon sur lequel est étudié le programme ou encore des spécificités du programme testé. Le *design* d'une expérience est souvent exécuté avec beaucoup d'application alors que dans un mode routinier, la mise en place d'une politique publique peut être moins élaborée. Dans ce cas, on s'attend à un problème d'indépendance environnementale. Allcott (2015) montre ainsi qu'un programme d'incitations aux économies d'électricité évalué expérimentalement a eu des effets plus élevés dans les premières zones où il a été mis en œuvre. Le biais de diffusion peut accentuer le problème d'indépendance environnementale lorsque l'on passe par exemple d'une expérimentation à petite échelle à la mise en œuvre d'une politique à grande échelle qui peut engendrer des problèmes d'équilibre général (Duflo *et alii*, 2008).

Solution : il est nécessaire d'expérimenter des programmes réalistes qui peuvent être applicables et généralisables. Concernant le problème de sélection de l'échantillon testé, une solution est de sélectionner aléatoirement des zones expérimentales au sein desquelles sont ensuite randomisés des groupes de traitement et de contrôle. Une autre solution est de réaliser d'autres expériences dans des contextes différents en s'appuyant sur la théorie économique pour inférer sur les situations non testées. Cette dernière solution résout également le problème de spécificité de la politique évaluée. Par ailleurs, l'économie structurelle, qui se base sur la modélisation des comportements individuels, est un moyen d'analyser les mécanismes sous-jacents d'une intervention économique (Duflo *et alii*, 2008).

Acceptation partielle et biais de randomisation : le problème d'acceptation partielle est relatif au refus de la part des individus assignés au groupe de traitement de recevoir le traitement. Si ce refus n'est pas aléatoire, il génère un phénomène de sélection qui biaise l'estimation. Le biais de randomisation est un cas particulier d'acceptation partielle qui provient de la sélection engendrée par la randomisation (Heckman, 1992). Ce biais de sélection peut provenir de problèmes éthiques qui poussent des individus à refuser de participer à une expérience (ou des organismes ou des villes à ne pas mettre en place le programme). La randomisation peut aussi être perçue comme un signal de mauvaise qualité. De plus, participer à une expérience randomisée en faisant partie du groupe de contrôle signifie ne pas bénéficier du traitement mais participer aux questionnaires de l'étude ce qui demande du temps et peut constituer une autre source de refus. Aussi, il a été montré que des individus refusent davantage de participer à une expérience randomisée qu'à une expérience non randomisée. À titre d'exemple Kramer et Shapiro ont trouvé, dans le cadre du test d'un médicament, que le taux de refus pour une expérience non randomisée était de 4 % et qu'il grimpait à 94 % pour une expérience randomisée pour les mêmes individus (Kramer et Shapiro, 1984). Sianesi (2017) montre que les individus refusant de participer à l'évaluation randomisée d'un programme d'aide au retour à l'emploi sont moins proches de l'emploi que ceux qui acceptent, alors que ceux qui n'ont pas été informés de l'existence du programme par leurs conseillers sont plus proches de l'emploi. Au bilan, l'effet expérimental surestime sans doute l'effet réel du programme.

Solution : l'acceptation partielle peut survenir lorsque l'on force les individus à entrer dans le programme. C'est le cas par exemple lors d'une expérience randomisée classique présentée par le *design* 1. On a vu qu'un moyen d'éviter cette sélection peut être la randomisation au niveau des recourants (*design* 2), la randomisation de l'accès au traitement (*design* 3) ou bien la randomisation d'un système d'encouragement (*design* 4). Le *design* 2 est particulièrement sujet au biais de randomisation parce qu'il est basé sur un processus d'auto-sélection qui peut être influencé par la réaction des individus

face à une expérience randomisée. Procéder à une expérience progressive (*design* 5) peut permettre d'atténuer ce biais.

Les méthodes quasi-expérimentales

L'analyse *ex post* consiste à évaluer un programme après sa mise en œuvre. Comme nous l'avons vu précédemment, il s'agit de comparer la situation dans laquelle la politique est mise en œuvre à la situation contrefactuelle dans laquelle elle ne l'aurait pas été. Cette partie présente des démarches *ex post* alternatives à l'expérimentation randomisée. Ces méthodes, appelées **quasi-expérimentales**, permettent d'évaluer une politique ou un programme qui est alloué de façon non aléatoire au sein d'une population. On distingue les *expériences naturelles* et les méthodes observationnelles (Dominici, Greenstone et Sunstein, 2015).

Expériences naturelles

Les expériences naturelles exploitent des situations réalisées sans intervention expérimentale, d'où le terme « naturel ». Cependant, ces expériences possèdent certaines caractéristiques permettant de les analyser comme si l'allocation du traitement était aléatoire (Angrist et Krueger, 2001). On distingue la méthode des variables instrumentales, la méthode de régression par discontinuité et la méthode de double différence.

Variables instrumentales

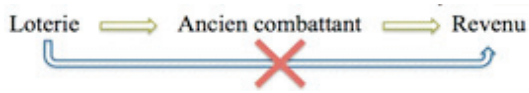
Définition : une variable instrumentale est une variable qui n'impacte pas directement la variable d'intérêt mais qui affecte la participation au programme évalué.

Dans le cadre d'une évaluation, l'utilisation d'une variable instrumentale consiste à estimer l'impact d'un programme sur une variable d'intérêt par l'intermédiaire d'une variable indépendante de ce paramètre. L'indépendance entre le paramètre d'intérêt et la variable instrumentale, ou *l'instrument*, permet d'isoler l'effet net du programme. Dans la partie précédente, le *design* 4 d'expérience randomisée, randomisation d'un encouragement, utilise précisément la nature instrumentale de l'encouragement pour identifier l'effet d'un programme sur un paramètre. L'encouragement augmente le taux de participation d'un programme, qui modifie à son tour le paramètre d'intérêt, mais n'affecte pas directement le résultat. On peut ainsi identifier l'impact causal du traitement sur les *compliers*, ceux qui recourent au programme lorsqu'ils reçoivent un encouragement mais qui n'y recourent pas en l'absence d'encouragement. Pour reprendre de nouveau l'exemple du programme de formation professionnelle, estimer l'impact d'un encouragement sur le taux d'activité permet d'évaluer l'impact du programme de formation au sein de la population participant au programme grâce à l'encouragement. Une variable constitue ainsi un instrument valide si elle n'est pas directement corrélée à la variable d'intérêt, à savoir le taux d'activité, mais qu'elle l'affecte uniquement par l'intermédiaire du taux de participation :

Ainsi, envoyer un courrier informatif à propos d'un programme de formation professionnelle augmente le nombre de recourants ; la conséquence est que le taux de participation est plus élevé, si bien que le

Encadré 6 : les anciens combattants ont-ils des revenus plus faibles ? (Angrist, 1990)

Un débat courant au sein des discussions autour des politiques militaires et sociales est de savoir si les vétérans sont suffisamment compensés pour leur service. Angrist a réalisé une étude afin d'évaluer l'impact de la guerre du Vietnam sur le revenu futur des vétérans. Une comparaison vétéran/non-vétéran serait sujette à un biais de sélection. En effet, il est possible que les individus ayant moins d'opportunités s'engagent plus dans l'armée. Ainsi les vétérans auraient des caractéristiques influençant à la fois leur décision d'entrer dans l'armée mais également leurs revenus futurs. Afin de répondre à cette question, Angrist utilise une expérience naturelle : à l'époque de la guerre du Vietnam les individus étaient appelés à faire leur service militaire sur la base d'un tirage au sort. Un numéro était attribué à chaque personne éligible, les hommes âgés de 19 à 26 ans, par un système de loterie. N'étaient appelés au combat que les individus ayant les numéros les plus faibles. Ce tirage se faisant de façon aléatoire, la sélection est indépendante du revenu futur. On a donc un bon instrument :



Ainsi la comparaison entre le revenu des hommes ayant reçu un numéro de loterie élevé et le revenu de ceux ayant reçu un numéro bas permet d'isoler l'effet causal d'être un ancien combattant sur le revenu. L'auteur trouve que les individus enrôlés lors de la guerre du Vietnam perçoivent un salaire annuel de 15 % inférieur aux non-vétérans. L'auteur relève cependant la possibilité d'un effet direct entre la loterie et le salaire. En effet, dans certains cas les individus appelés au combat étaient autorisés à repousser leur entrée dans l'armée pour terminer un cycle d'étude. L'attribution d'un numéro faible augmentant le risque d'aller au Vietnam, certains individus ont pu être incités à poursuivre leur cursus pour éviter d'entrer dans l'armée. Leur niveau d'éducation, et *a fortiori* leur salaire, se trouvant ainsi plus élevé.

Encadré 7 : l'impact de l'instruction obligatoire sur le revenu (Angrist et Krueger, 1991)

Quel est l'impact de l'éducation sur le niveau de revenu ? Angrist et Krueger apportent une réponse en exploitant une expérience naturelle afin d'estimer l'effet causal des politiques d'instruction obligatoire sur le revenu futur des élèves. Supposons que la scolarisation soit obligatoire à partir de 6 ans. Étant donné que les enfants naissent à différents mois de l'année, tous ne commencent pas l'école exactement au même âge. En effet, pour les enfants célébrant leur 6ème anniversaire durant l'année civile courante, ceux étant nés en janvier commencent à l'âge de 6 ans et 8 mois alors que les enfants nés en décembre commencent leur scolarité à 5 ans et 8 mois. Ainsi, si l'école est obligatoire jusqu'à l'âge de 16 ans (comme dans de nombreux pays), les individus qui quittent le système scolaire à ce moment-là n'auront pas tous la même durée d'éducation en fonction de leur mois de naissance. Les enfants nés en début d'année atteignent l'âge légal de sortie de l'école plus tôt dans leur cursus éducatif et donc acquièrent un niveau d'éducation inférieur. La variable instrumentale utilisée dans cette étude est le trimestre de l'année durant lequel un individu est né : janvier-mars, avril-juin, juillet-septembre, et octobre-décembre. Être né à une période de l'année ou une autre est *a priori* indépendant du revenu futur : cela n'impacte pas le milieu socio-économique ou les capacités intellectuelles, et constitue donc une sorte d'aléa. Cependant, la période de naissance affecte le revenu à travers la durée d'éducation :

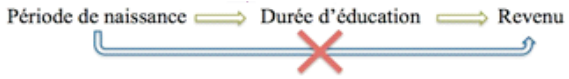
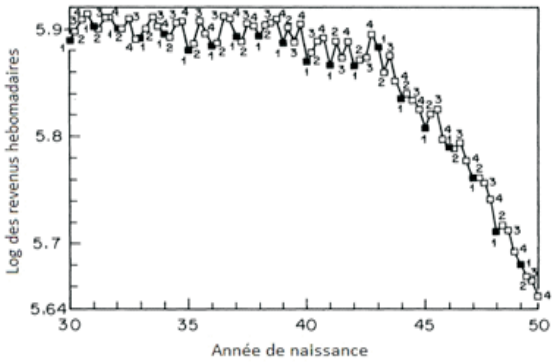


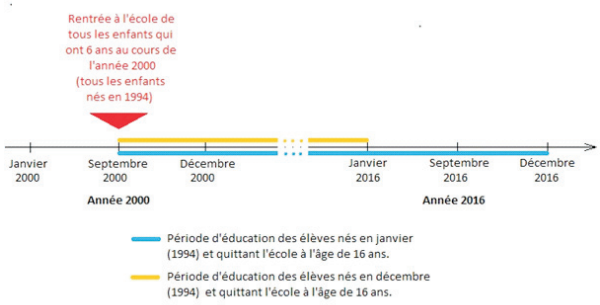
Figure 8 : salaire moyen des hommes nés entre 1930 et 1949 en fonction de leur trimestre de naissance



Source : *Quartely Journal of Economics*, vol. 106, Issue 4, (Nov. 1991), pp. 979-1014.

Ainsi, si on observe une différence significative de revenu entre les individus nés dans le premier quart et ceux nés dans le dernier quart de l'année, cette différence peut être attribuée à la durée d'éducation. De la sorte, cette variable instrumentale permet d'analyser l'impact causal de lois d'instruction obligatoire sur le revenu. Les auteurs trouvent que les individus nés au premier trimestre de l'année ont en moyenne un revenu inférieur à ceux qui sont nés plus tard comme on peut l'observer sur la figure 8. Les élèves nés en fin d'année et qui sont donc obligés de suivre une éducation plus longue à cause de la législation ont un salaire moyen plus élevé. Les auteurs précisent que cette étude est pertinente pour la population d'élèves quittant l'école tôt. En effet, il est moins évident que l'on observe ce même phénomène pour les élèves poursuivant des études supérieures. Grenet (2010) étudie l'impact du trimestre de naissance sur le niveau d'éducation et la situation professionnelle en France. Contrairement à l'article de Angrist et Krueger, cet article suggère que les individus nés en fin d'année et commençant ainsi leur éducation plus tôt sont pénalisés dans leur éducation puis dans leur carrière notamment à cause de leur retard originel de maturité intellectuelle. L'écart entre les individus nés en janvier et ceux nés en décembre semble s'atténuer dans le temps mais les résultats indiquent que les individus nés en fin d'année ont tout de même une probabilité plus grande de redoubler une classe. Grenet trouve également que les hommes nés en fin d'année ont une probabilité plus grande d'avoir un revenu plus faible ou d'être au chômage.

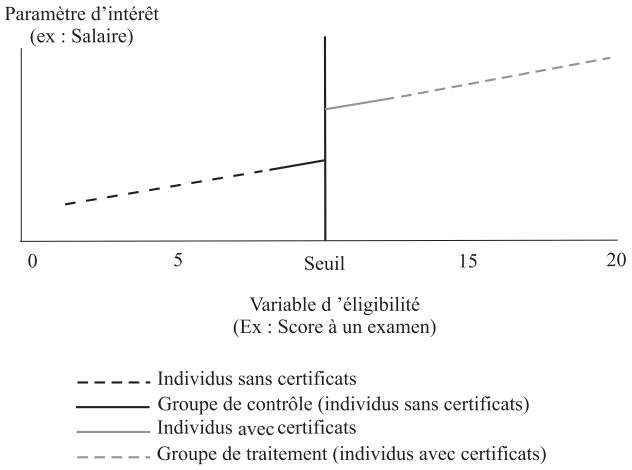
Figure 9 : durée d'éducation selon la période de naissance



programme affectera davantage le taux d'activité, sans que ce courrier n'ait d'impact direct sur le taux d'activité :

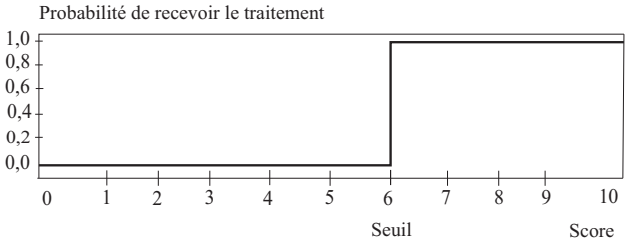
Imaginons maintenant qu'un encouragement, disons l'envoi d'un courrier électronique, impacte directement la probabilité d'avoir un emploi, et *a fortiori* le taux d'activité. Ce phénomène se produirait si l'on imagine par exemple que les courriers incitent les individus à une recherche d'emploi plus active. Dans cette configuration l'encouragement est donc corrélé au taux d'activité indépendamment du programme de formation. Il ne constitue pas un bon instrument puisque l'impact du programme de formation sur les *compliers* surestimerait l'effet de la formation.

Figure 10 : la régression par discontinuité

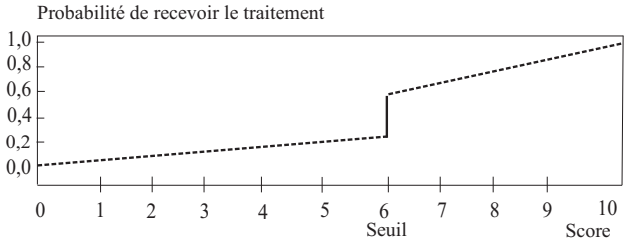


Figures 11 : règle d'éligibilité flexible et règle d'éligibilité stricte

1. Règle règle d'éligibilité stricte



2. Règle d'éligibilité flexible



Régression par la discontinuité

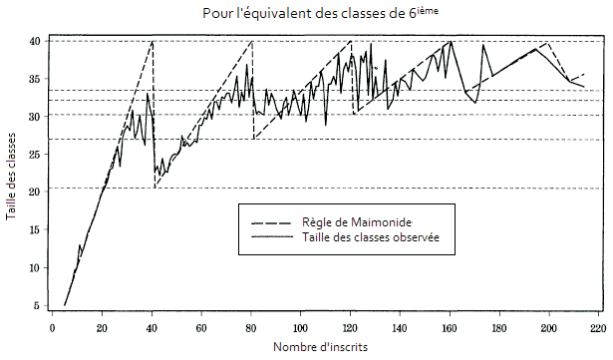
Cette méthode consiste à identifier une situation où l'allocation dépend d'**une règle de sélection relative au seuil d'un certain facteur**, par exemple un programme social disponible à partir d'un certain seuil de revenu ou une bourse accessible à partir d'un certain score à un examen.

Définition : cette méthode utilise le fait que lorsqu'un programme est distribué par rapport à un seuil d'éligibilité, une discontinuité est créée dans l'allocation du programme. Les individus juste

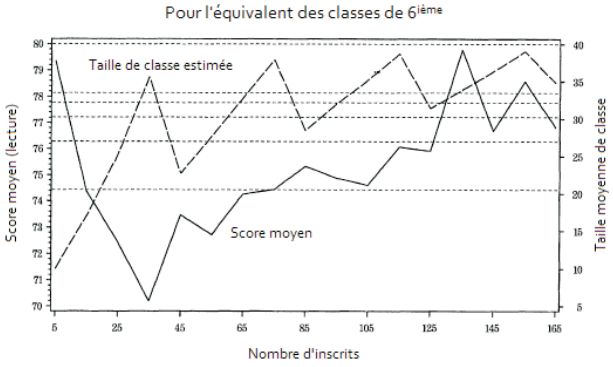
Encadré 8 : l'impact de la taille des classes sur les résultats scolaires (Angrist et Lavy, 1999)

La taille des classes est un sujet courant dans les débats politiques autour de la qualité de l'enseignement. Angrist et Lavy s'interrogent sur l'impact de la taille des classes sur la réussite scolaire. Comparer les résultats scolaires des élèves entre des classes de petite et de grande taille serait sujet à un biais de sélection : les classes de petite taille accueillent généralement des élèves plus en difficulté. La comparaison des résultats des élèves entre petites et grandes classes sous-estime l'impact causal de la taille des classes sur le niveau scolaire. Ce biais est en général suffisamment élevé pour faire apparaître une corrélation positive entre taille des classes et résultats des élèves. Celle-ci est bien entendu fallacieuse : augmenter la taille des classes n'améliore pas les performances des élèves. Pour identifier l'effet causal, Angrist et Lavy, utilisent la règle de Maïmonide, qui est utilisée en Israël pour déterminer la taille des classes. Cette règle fixe le nombre maximal d'élèves par classe à 40. Elle induit une discontinuité entre le nombre d'élèves inscrits dans une école et la taille des classes à chaque multiple de 40 (voir figure 12). La règle de Maïmonide constitue une source de variation aléatoire de la taille des classes. En effet, il est peu probable que cette règle affecte les résultats scolaires par un autre biais que la taille des classes. La figure ci-après illustre clairement l'impact causal de la taille des classes sur le niveau moyen des scores. En exploitant cette discontinuité, Angrist et Lavy trouvent que réduire la taille des classes améliore la réussite des élèves.

Figure 12 : relation entre nombre d'inscrits, résultats scolaires et taille des classes



Source: The Quartely Journal of Economics, Vol. 114, No. 2 (May, 1999), 533-575.

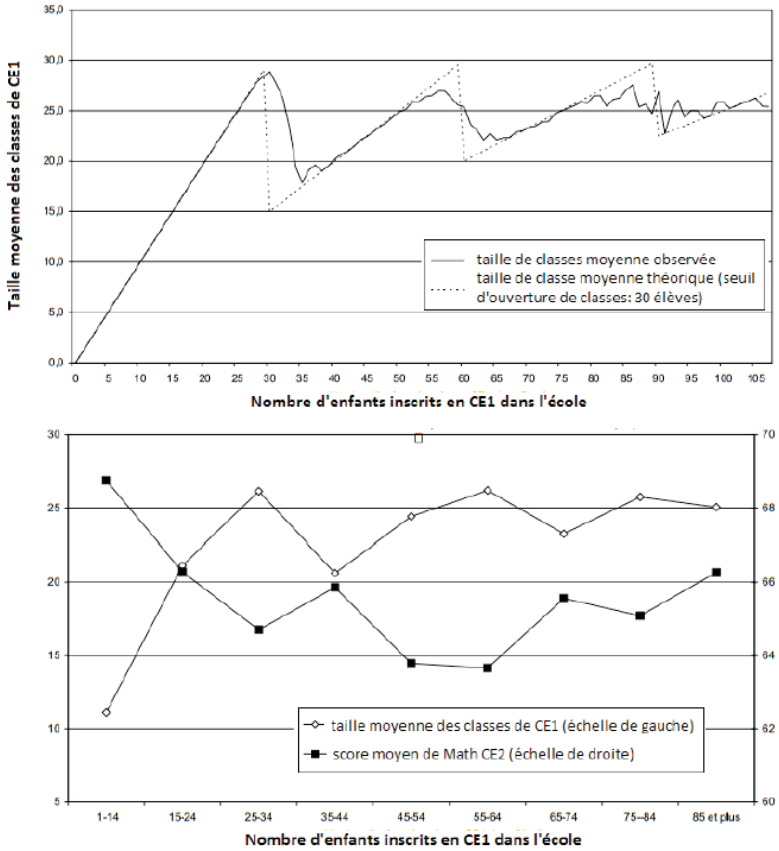


Source: The Quartely Journal of Economics, Vol. 114, No. 2 (May, 1999), 533-575.

Encadré 9 : l'impact de la taille des classes sur les résultats scolaires en France (Piketty et Valdenaire, 2006)

Une étude sur l'impact de la taille des classes sur la réussite scolaire a également été réalisée en France par Piketty et Valdenaire. Les auteurs ont utilisé des données sur des classes de primaire et du secondaire. Leurs résultats montrent que la taille des classes a un impact très fort sur la réussite scolaire dans les écoles primaires. En effet, ils trouvent qu'une réduction d'effectif d'un élève dans une classe de CE1 augmente de 0,7 point la note obtenue par les enfants défavorisés aux premiers tests de CE2. Ci-dessous, le premier graphique représente la discontinuité de la taille des classes à chaque multiple de 30. Le second graphique illustre la relation négative entre taille des classes et réussite scolaire. Piketty et Valdenaire discutent également de la politique des ZEP (zones d'éducation prioritaire), dont les classes ont en moyenne 20,9 élèves contre 22,8 dans les écoles non ZEP. Leurs résultats suggèrent qu'une diminution plus grande de l'effectif de ces classes diminuerait substantiellement les inégalités de réussite scolaire. Concernant l'éducation secondaire, l'impact de la taille des classes est moins important mais va dans la même direction.

Figures 13 : relation entre nombre d'inscrits, résultats scolaires et taille des classes



Source: Les dossiers - Enseignement scolaire 173 - 2006 -L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français - Estimations à partir du panel primaire 1997 et du panel secondaire 1995 - Thomas Piketty (EHESS), Mathieu Valdenaire (EHESS).

en-dessous du seuil et ceux juste au-dessus ont naturellement des caractéristiques similaires alors que seuls les derniers ont accès au traitement. Cette discontinuité permet de comparer l'effet net d'un programme sur les traités proche du seuil d'éligibilité en utilisant comme groupe de contrôle les individus non traités mais aussi proches de ce seuil (*cf.* figure 10).

Exemple 1 : un étudiant arrivant 121ème à un concours où seulement 120 personnes sont admises a probablement un niveau de connaissance équivalent à celui arrivé 120ème et étant admis. Comparer les individus autour de la 120ème place permettrait de distinguer l'effet moyen de ce concours, sur le niveau de revenu par exemple, pour les individus près du seuil. En 1960, les auteurs Thistlethwaite et Campbell ont analysé l'impact de certificats au mérite sur la performance future des étudiants bénéficiaires. Ils ont utilisé le fait que l'attribution de ces certificats était basée sur le score d'un examen. Les étudiants juste en-dessous du seuil d'attribution constituaient donc un groupe de contrôle valide pour les étudiants bénéficiaires proche du seuil (Thistlethwaite et Campbell, 1960).

On peut remarquer que cette méthode peut également être utilisée lorsque que la règle d'éligibilité est flexible, c'est-à-dire lorsque que la probabilité d'être traité n'est pas nulle pour des individus sous le seuil. La figure 11 illustre cette différence. Un concours d'entrée pour une formation

Encadré 10 : estimation de la valeur économique de l'air non pollué et impact des politiques de régulation (Chay et Greenstone, 2005)

Chay et Greenstone exploitent la structure des *Clean Air Act Amendements* (CAAAs) de 1970 afin d'estimer la valorisation de l'air non pollué par la méthode de régression par discontinuité. Les CAAAs, mis en œuvre au niveau régional, visent à diminuer le niveau de pollution. Si le niveau de pollution d'une région dépasse un seuil de concentration annuel s'élevant à 75$\mu g/m^3$, cette région considérée comme polluée est alors soumise à une règlementation plus stricte. Cette expérience naturelle attribue un statut aux régions, *polluée* et *non polluée*. Les auteurs peuvent ainsi analyser la différence de prix de l'immobilier autour du seuil. La différence de prix de l'immobilier entre les régions juste en dessous et juste au-dessus du seuil peut être interprétée comme l'impact du label « régions polluées ». Ils s'intéressent également à la variation de pollution autour du seuil afin d'inférer sur l'efficacité de la politique de régulation. Ces discontinuités induites par les CAAAs représentent l'impact du signal « région polluée » sur le prix de l'immobilier et le niveau de pollution. Il est en effet peu probable que cette politique impacte ces paramètres par un autre mécanisme. On observe pour cette raison des régions désignées comme polluées à gauche de la droite verticale matérialisant le seuil annuel.

La figure 14 montre que, au niveau du seuil, les régions polluées connaissent une réduction de la pollution substantiellement plus élevée que les régions classées comme non polluées en 1975. Par ailleurs, la figure 15 illustre l'impact sur dix ans du statut « pollué » sur les prix de l'immobilier. On y voit une relation nette : les régions classées comme polluées connaissent une augmentation du prix de l'immobilier plus importante. Les auteurs concluent qu'une diminution de 1$\mu g/m^3$ de la concentration de particules polluées induit une augmentation de la valeur immobilière allant de 0,20 % à 0,35 %. L'étude montre donc que la politique du CAAAs a permis une baisse du niveau de pollution qui a été reflétée par une hausse des prix de l'immobilier.

Figure 14 : variation de la pollution entre 1970 et 1980 selon le statut et le niveau de pollution

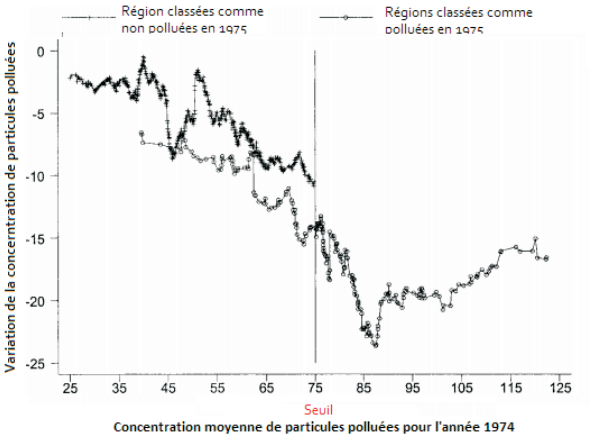


Figure 15 : variation des prix de l'immobilier entre 1970 et 1980 selon le statut et le niveau de pollution

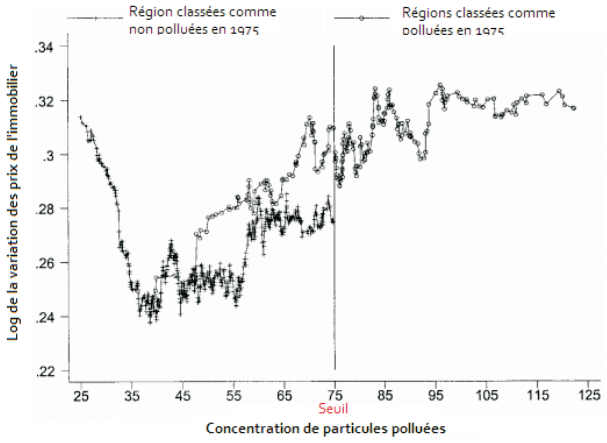
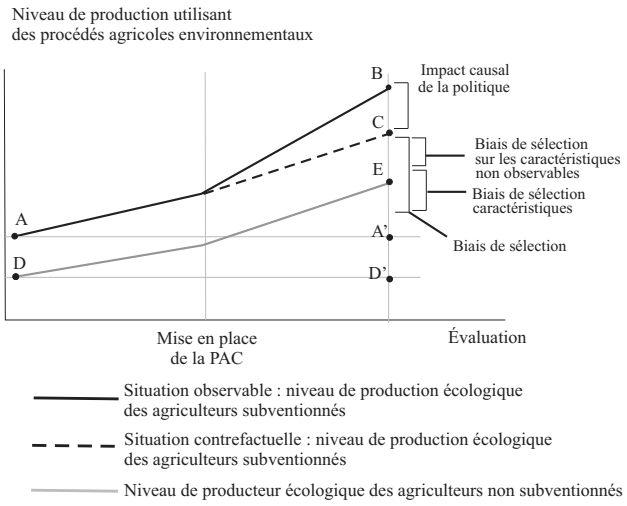


Figure 16 : décomposition du biais de sélection



peut par exemple comprendre une règle d'éligibilité flexible si l'on considère que parmi les 120 premiers étudiants (si tel est le seuil), certains n'accepteront pas de suivre la formation concernée permettant ainsi à des étudiants ayant un rang inférieur à 120 d'y accéder. Dans cette situation, la probabilité de traitement est supérieure à 0 pour les individus au-dessous du seuil et est inférieure à 1 pour ceux qui sont au-dessus.

Double différence

La méthode de double différence exploite le fait qu'une partie de la population a vu son exposition à la politique évaluée changer dans le temps. La double différence compare l'évolution du groupe dont l'exposition a changé à celle des groupes dont l'exposition n'a pas changé. La méthode de double différence est souvent appliquée lorsqu'une expérience naturelle a exposé des individus ou des

Encadré 11 : l'impact d'une hausse du salaire minimum sur le taux d'emploi et les prix (Card et Krueger, 1994)

Comment les entreprises réagissent-elles à une hausse du salaire minimum ? Il s'agit d'une question socio-économique importante et c'est pour cela que Card et Krueger proposent une nouvelle réponse. En effet, intuitivement une hausse du salaire minimum devrait inciter les entreprises à embaucher moins ou à répercuter cette hausse de coût sur les prix. Les auteurs utilisent une double différence pour analyser l'effet d'une augmentation du salaire horaire minimum sur le taux d'emploi et les prix sur le marché des *fast-foods* au New Jersey et en Pennsylvanie. Ils parviennent à récolter des données sur l'emploi et les prix avant l'augmentation et 8 mois après l'augmentation pour quasiment 100 % des *fast-foods*. L'augmentation du salaire minimum légal n'ayant lieu qu'au New Jersey, le groupe de comparaison est constitué des *fast-foods* de Pennsylvanie. Les auteurs comparent la variation du niveau d'emploi, de salaire et des prix des restaurants du New Jersey avant et après la hausse du salaire minimum à la variation observée dans les restaurants du groupe de contrôle. Utilisant la technique de la double différence, ils trouvent que la hausse de salaire ne réduit pas l'emploi mais au contraire l'augmente. Les auteurs établissent également une comparaison de ces paramètres entre les restaurants du New Jersey qui payaient leurs employés déjà au-dessus du salaire minimum de la nouvelle législation à ceux qui rémunéraient leurs employés en dessous de ce nouveau salaire minimum. Les auteurs trouvent que l'augmentation du taux d'emploi est presque aussi grande au sein des restaurants payant déjà de hauts salaires qu'au sein de ceux payant en dessous du nouveau salaire minimum. Concernant l'évolution des prix, les prix dans le New Jersey ont augmenté significativement par rapport à niveau des prix en Pennsylvanie. Les auteurs ne trouvent, cependant, pas de différence significative de prix entre les restaurants payant à l'origine de hauts salaires et ceux offrant des salaires plus bas.

Encadré 12 : l'impact des écoles sur le niveau d'éducation et le revenu (Duflo, 2001)

Duflo étudie l'impact d'un important programme de construction d'école en Indonésie sur le revenu ainsi que le niveau d'éducation des élèves. Cette étude renvoie aux questions fréquentes de l'économie du développement concernant le rôle de l'investissement et des infrastructures dans l'éducation. Le programme étudié est le Sekolah Dasar INPRES lancé en 1973 et qui a permis la construction de 61 000 écoles primaires. Duflo utilise une double différence et calcule la différence de niveau d'éducation et de revenu entre les individus qui ont bénéficié du programme (âgés de 2 à 6 ans en 1973) et les individus trop âgés pour bénéficier de la construction d'écoles primaires (âgés de 12 à 17 ans) qui forment le groupe de contrôle. Cette différence est établie pour deux types de zones : les zones où peu ou aucune école n'a été construite et les zones fortement touchées par le programme. Il s'agit donc bien d'une double différence : Duflo compare la variation de niveau d'éducation entre cohortes et entre régions. L'hypothèse de tendance temporelle parallèle nécessaire à la réalisation d'une double différence se traduit de la façon suivante : en l'absence de programme, la variation de niveau de revenu et d'éducation entre cohortes n'aurait pas été significativement différente entre les zones fortement exposées et les zones faiblement exposées au programme. Sous cette hypothèse la différence entre cohortes de chaque zone fournit une estimation de l'impact causal du programme. Pour tester la validité de cette hypothèse, Duflo compare l'évolution des cohortes 12 à 17 et 17 à 22 entre les deux types de régions. Comme aucune de ces deux cohortes n'est concernée par le programme, l'évolution de leur niveau d'éducation et de revenu devrait suivre la même tendance. C'est bien le cas, ce qui renforce considérablement la crédibilité de l'estimateur de double différence. On parle de test placebo, puisqu'il consiste à estimer un effet là où aucun traitement n'a été appliqué. Les résultats de l'estimation de l'effet suggèrent que le programme de construction a entraîné une hausse du nombre d'années d'éducation de 0,12 à 0,19, ainsi qu'une augmentation de revenu de 1,5 % à 2,7 %.

régions à une politique dont ils ne bénéficiaient pas auparavant. La double différence est aussi applicable à l'analyse d'expérimentations randomisées ou en l'absence d'expérience naturelle.

Nous présentons maintenant cette méthode de façon plus précise à travers l'exemple des mesures agro-environnementales de la PAC. Nous expliquons ici la double différence graphiquement à l'aide de la figure 16. La réalisation d'une double différence nécessite des données sur le paramètre d'intérêt sur deux périodes, une période pré-programme et une période post-programme, et ce, pour un groupe traité et un groupe non traité. Dans notre exemple cela correspond à des données sur le taux d'utilisation des procédés agricoles environnementaux pour les agriculteurs subventionnés et les agriculteurs non subventionnés pour une période antérieure à la mise en place des subventions et une période postérieure. L'impact de la politique recherché par cette méthode est l'**impact causal de la politique sur les recourants**, représenté par le segment [*BC*] figure 16. Cette méthode s'appuie sur l'hypothèse de tendances temporelles parallèles équivalente à une égalité du biais de conjoncture.

Hypothèse de tendances temporelles parallèles ou égalité du biais de conjoncture : la double différence s'appuie sur l'hypothèse selon laquelle l'évolution de l'économie a le même impact sur le groupe sujet à la politique évaluée et le groupe non traité. Ceci signifie que le biais de conjoncture est le même dans les deux groupes. Sur la figure 16, cela se traduit par un écart constant entre (*AC*) et (*DE*), qui sont donc parallèles.

L'impact causal sur les recourants est estimé par la différence entre la comparaison avant/après du groupe traité ([*A'B*] ou (*B - A*)) et la comparaison avant/après du groupe non traité ([*ED'*] ou (*E - D*)). Le second terme de cette différence est utilisé comme estimation du biais de conjoncture. L'impact

causal est estimé par la soustraction à une comparaison avant/après de l'estimation du biais de conjoncture. Or, par hypothèse, le biais de conjoncture est identique dans les deux groupes : on a [*A'C*] = [*ED'*]. Autrement dit, on obtient une estimation non biaisée de l'impact causal. Plus formellement, la double différence se traduit par l'expression suivante : (*B* - *A*) - (*E* - *D*), ce qui correspond en terme de longueur à : [*A'B*] - [*D'E*]. Or étant donné que les droites (*AC*) et (*DE*) sont parallèles, [*A'C*] = [*D'E*]. Ainsi, en remplaçant dans la première égalité on obtient : [*A'B*] - [*D'E*] = [*A'B*] - [*A'C*] = [*BC*], ce qui correspond exactement à l'impact causal recherché. On peut aussi montrer que la double-différence consiste à corriger la comparaison avec / sans ([*EB*]) par la différence qui existait entre bénéficiaires et non-bénéficiaires avant la mise en place de la politique ([*A'D'*]). En effet, il est toujours vérifié que [*A'B*] - [*D'E*] = [*EB*] - [*D'A'*]. La double-différence utilise donc la différence entre bénéficiaires et non-bénéficiaires avant la mise en place de la politique comme une estimation du biais de sélection. Elle est non biaisée si le biais de sélection est constant dans le temps : [*D'A'*] = [*CE*]. Sous cette hypothèse, on a bien en effet [EB] - [D'A'] = [EB] - [CE] = [BC]. Il est à noter finalement que l'hypothèse de biais de sélection constant dans le temps est équivalente à l'hypothèse de biais de conjoncture identique pour les bénéficiaires et les non-bénéficiaires. En effet, [*CA'*] - [*ED'*] = [*CE*] + [*EA'*] - ([*EA'*] + [*A'D'*]) = [*CE*] - [*D'A'*].

Encadré 13 : estimation de l'impact des mesures agro-environnementales de la PAC (Chabé-Ferret et Subervie, 2013)

Nous reprenons ici l'exemple des mesures agro-environnementales (MAE) de la PAC mentionnées précédemment. Les MAE consistent à rémunérer les agriculteurs qui adoptent des pratiques agricoles environnementales (diversités des cultures, agriculture biologique, etc.). La figure 18 illustre le problème de biais de sélection qui peut intervenir dans l'évaluation des MAE. En effet, comme nous l'avons mentionné, il est probable qu'un nombre substantiel d'agriculteurs recevant les subventions utilisait déjà ces pratiques et les aurait donc tout de même appliquées en absence de MAE. Ainsi utiliser le groupe des non subventionnés comme contrefactuel surestimerait l'effet causal des MAE sur le niveau d'utilisation des pratiques environnementales. Les auteurs estiment l'impact causal des MAE par la méthode du *matching*. Le contrefactuel utilisé est le groupe d'individus représentés par la courbe gris clair figure 18 tracée à partir de données sur des caractéristiques observables de ces individus. Les auteurs trouvent que les MAE entraînent une augmentation des pratiques environnementales se traduisant par une utilisation sur 11,24 hectares de cultures supplémentaires alors qu'une simple comparaison avec/sans produit un résultat de 16,12 hectares. Cette estimation ne prend pas en compte les caractéristiques non observables des agriculteurs. Pour ajuster leur estimation, Chabé-Ferret et Subervie utilisent la méthode de double différence que nous avons présenté précédemment. Ils trouvent finalement un impact causal de 10,66 ha.

Méthodes observationnelles

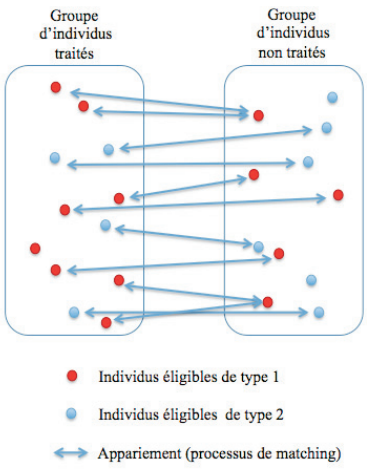
En l'absence d'expérience randomisée ou d'expérience naturelle exploitable, l'évaluateur peut utiliser une méthode observationnelle. Les méthodes observationnelles essaient de prendre en compte les déterminants observés du biais de sélection. L'approche observationnelle la plus simple est d'introduire les déterminants observés du biais de sélection dans un modèle linéaire estimé par la méthode des moindres carrés ordinaires. Une approche plus sophistiquée est basée sur la méthode du *matching*. Dans cette partie, nous discutons brièvement la méthode du *matching* puis exposons les enjeux auxquels font face les méthodes observationnelles.

Le matching

Définition : la méthode de *matching*, ou méthode d'appariement, utilise des données non-expérimentales et estime l'impact causal d'un programme en comparant des individus traités à des individus non traités qui possèdent des **caractéristiques observées similaires**.

Autrement dit, en utilisant des données sur des individus traités et non traités, la méthode de *matching* distingue les individus traités et les individus non traités ayant les mêmes caractéristiques (revenus, sexe, situation sociale, nationalité, etc.) puis établit une comparaison entre les individus similaires. Dans une population composée d'individus de type 1 et de type 2 on peut comparer les individus traités de type 1 aux individus non traités de type 1 si les types sont observables. La figure 17 illustre de façon simplifiée ce mécanisme. Puisque selon le type de procédure de *matching* utilisée, différents individus traités peuvent être appareillés à un même individu non traité ayant des caractéristiques similaires, on observe sur la figure 17 plusieurs flèches allant vers un même individu non traité.

Figure 17 : principe du *matching*



Cette méthode repose sur plusieurs hypothèses :

– **Hypothèse 1** : des individus ayant des caractéristiques similaires ont potentiellement la même valeur des paramètres étudiés (exemple : taux d'activité) en l'absence du programme. Ainsi, si l'on prend l'exemple du programme de formation professionnelle, s'ils avaient été bénéficiaires, les individus non traités auraient le même taux d'activité que des individus traités ayant les mêmes caractéristiques (revenu, éducation,...). De même des individus traités, s'ils n'avaient pas reçu le programme, auraient le même taux d'activité que des individus non traités ayant les mêmes caractéristiques.

– **Hypothèse 2** : pour chaque caractéristique, il existe à la fois des individus traités et des individus non traités. Cela garantit que chaque individu traité ait un « jumeau » non traité à qui il peut être comparé.

Cette seconde hypothèse suggère que **les différences de caractéristiques entre individus sont observables**. En effet on peut dire d'individus qu'ils sont similaires uniquement sur la base de leurs caractéristiques observables. La méthode d'appariement résout le problème de sélection au regard des caractéristiques observables. Toutefois, le problème de biais de sélection concernant les caractéristiques non observables ne l'est pas, ce qui constitue la faiblesse majeure de cette méthode d'évaluation. Dans le cas du programme de formation, la motivation et les facultés intellectuelles ne sont pas facilement observables bien qu'elles constituent des facteurs impactant le taux d'activité. La figure 18 présente la décomposition du biais de sélection entre caractéristiques observables et non observables à travers l'exemple des mesures agro-environnementales de la PAC présenté précédemment. Pour rappel, ces mesures consistent à verser des subventions aux agriculteurs qui utilisent des technologies de production écologiques.

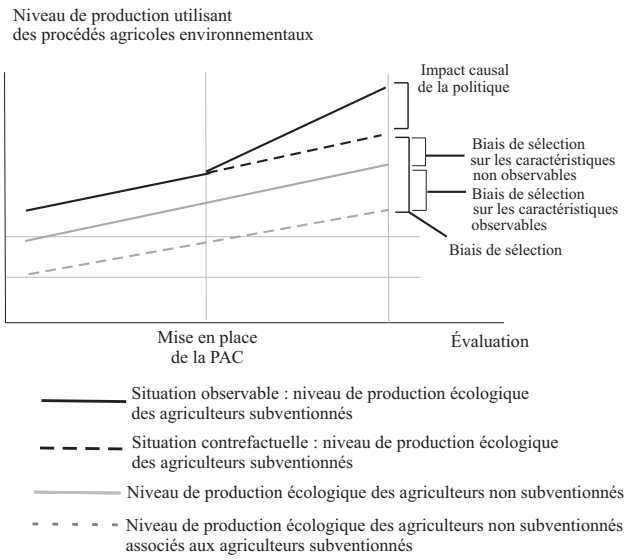
– **Hypothèse 3** : comme pour une majorité de méthodes d'évaluation, il est supposé que l'effet d'un programme n'impacte pas d'autres individus que ceux visés par le traitement, autrement dit qu'il n'y a pas d'effet de diffusion. Si une politique a un impact à la fois sur les individus traités et non traités, comparer des individus traités et non traités similaires ne permet pas d'estimer l'impact causal réel de cette politique.

L'hypothèse 1 sous-entend donc qu'au sein d'un groupe d'individus identiques il existe une source de variation qui permet d'avoir des individus traités et des individus non traités. L'hypothèse 2 ajoute indirectement que cette source de variation peut être considérée comme étant aléatoire. La méthode d'appariement utilise donc des individus non traités comme approximation de l'état contrefactuel d'individus traités ayant les mêmes caractéristiques en faisant l'hypothèse qu'un mécanisme a alloué aléatoirement le traitement parmi les individus similaires. La complexité de cette méthode réside dans la technique utilisée pour associer les individus traités à des individus non traités similaires. Il existe différents procédés pour réaliser cet appariement que nous ne décrivons pas plus ici.

Méthodes quasi-expérimentales *v.s.* expériences randomisées

Les quasi-expériences, qui utilisent des données *ex post*, ont l'avantage de ne pas soulever les problèmes éthiques et politiques induits par la randomisation. Il est cependant important de se demander dans quelle mesure elles parviennent à reproduire l'estimation d'une expérience randomisée, c'est-à-dire une estimation non biaisée. Depuis LaLonde (1986), une série d'études ont comparé les résultats des méthodes observationnelles à ceux des expériences randomisées dans le contexte de programmes de formation (par exemple, Bléhaut et Rathelot, 2014). Les résultats de cette littérature sont encore à compiler mais ils semblent indiquer que les méthodes observationnelles captent une grande partie du bais de sélection lorsqu'elles utilisent des variables de contrôle riches comme l'évolution récente des revenus salariaux, mais qu'elles souffrent toujours généralement d'un biais négatif (Chabé-Ferret, 2014). D'autres travaux tentent de mieux comprendre les sources du biais des méthodes observationnelles. Chabé-Ferret (2015) montre que les méthodes observationnelles ne parviennent généralement pas à capturer l'ensemble des déterminants qui font que les salariés les plus éloignés de l'emploi participent aux programmes de formation.

Figure 18 : décomposition du biais de sélection



Discussion

Les outils traditionnels d'évaluation des politiques publiques reposent largement sur des méthodes *ex ante*, de type analyse coût bénéfice (ACB). De nombreux guides méthodologiques sur ces méthodes ont été produits par l'administration française, notamment dans le secteur des transports et des infrastructures (Boiteux, 2001 ; Quinet, 2013). Quand les politiques publiques ont des impacts importants sur l'économie, ces méthodes ont souvent recours à des modèles numériques d'équilibre général ou de croissance à long terme, comme cela a été le cas pour l'étude des politiques climatiques par exemple. Dans tous les cas, ces méthodes reposent sur des hypothèses sur les effets attendus d'une politique sur l'économie, en particulier sur des modèles décrivant le comportement des agents économiques (entreprises, consommateurs) et le fonctionnement des marchés et des institutions.

En pratique, cette évaluation *ex ante* est compliquée. Ceci est particulièrement vrai pour les politiques sociales dont les effets attendus dépendent crucialement du facteur humain. Les réactions des individus qui bénéficient de ces politiques (aide au retour à l'emploi, incitation à la scolarisation, réponse aux subventions, etc.) sont difficiles à anticiper car elles varient en fonction de caractéristiques propres à chacun ; d'où l'intérêt d'expérimenter les mesures envisagées en les appliquant d'abord à des groupes restreints avant de décider, au vu des résultats, d'en élargir (ou non) l'application. L'expérimentation doit donc être vue comme un « *processus d'apprentissage en continu* » (Banerjee et Duflo, 2009). Elle joue un rôle tout au long du cycle de vie de la politique étudiée. Les informations issues de la phase expérimentale alimentent les hypothèses des études de type ACB qui cherchent à identifier les mesures les plus efficaces en termes d'effets sur le bien-être social. Ensuite, une fois celles-ci mises en œuvre, les systèmes d'observation *ex post* renseignent sur d'éventuels effets causaux pervers ou inattendus. Ainsi, l'enjeu est finalement d'articuler l'ensemble des outils d'évaluation pour qu'ils servent au mieux la décision publique dans le cadre d'un système d'évaluation révisé en permanence.

Cette vision de l'évaluation comme un processus continue amène naturellement à reconsidérer les domaines d'application respectifs des évaluations *ex ante* et *ex post*. Comme nous l'avons dit précédemment, l'évaluation *ex post* a été beaucoup utilisée dans les domaines du travail, de l'éducation et du développement, c'est-à-dire dans des domaines où la compréhension fine des comportements est déterminante. D'un autre côté, l'évaluation *ex ante* a été largement appliquée dans des domaines comme les transports, l'énergie et l'environnement. Dans ces domaines, les politiques peuvent avoir des effets macroéconomiques importants sur le (très) long terme, et l'intérêt des approches *ex post* apparaît moins évident. Considérons par exemple la politique climatique. En caricaturant, on voit mal comment simuler les impacts à long terme d'une planète avec ou sans politique climatique. Pourtant, des études *ex post* récentes nous informent sur le lien entre politique climatique et émissions de gaz à effet de serre (Aichele et Felbermayr, 2015), ou sur le lien entre la température et le dommage climatique (Deschenes et Greenstone, 2011). Ainsi, ces études peuvent être pertinentes pour mieux appréhender *ex ante* les impacts à long terme de la politique climatique. Des arguments similaires peuvent être avancés concernant l'évaluation des politiques macroéconomiques (Fuchs-Schuendeln et Hassan, 2015). Enfin, l'utilisation des méthodes (quasi-)expérimentales n'est pas limitée aux seules politiques publiques. Des acteurs privés comme les entreprises de l'Internet utilisent régulièrement ces méthodes pour évaluer l'effet de leurs décisions (Kohavi *et alii*, 2009). De manière symétrique, les méthodes informatiques comme le « machine *learning* » viennent enrichir la panoplie des méthodes quasi-expérimentales (Athey, 2015).

À ce stade, il convient d'ajouter que l'approche empirique en économie, et *a fortiori* l'évaluation des politiques publiques, ne se limite pas aux méthodes expérimentales et quasi-expérimentales. Les méthodes d'économétrie structurelle qui permettent de prédire *ex ante*, à partir de modèles du comportement des agents, les conséquences d'interventions publiques, sont aussi très utiles. Les modèles structurels bénéficient aussi d'une révolution de la crédibilité, leurs prédictions étant de plus en plus confrontées à la réalité, que ce soit en comparant les prédictions *ex ante* aux observations *ex post* (McFadden, 2001 ; D'haultfoeuille, Durrmeyer et Février, 2011), ou les prédictions basées sur un modèle estimé sur les données du groupe de contrôle d'une expérimentation randomisée à ce qui est observé dans le groupe de traitement (Todd et Wolpin, 2006).

Reste que l'évolution des méthodes d'évaluation des politiques publiques reflète, selon nous, la conversion progressive de la science économique en une véritable science empirique. Les prédictions théoriques sont testées systématiquement en ayant recours à une batterie d'outils dont la validité fait consensus parmi les chercheurs. Bien entendu, les résultats de ces méthodes ne suffisent pas seuls à valider des propositions théoriques. C'est la multiplication de résultats similaires, la possibilité de les reproduire indépendamment, les méta-analyses, qui cristallisent peu à peu le savoir acquis. En ce sens, si nous sommes d'accord avec Cahuc et Zylberberg (2016) sur la transformation de la science économique en une science empirique, nous pensons

que ces auteurs accordent trop de poids à des études isolées et au prestige de la revue qui les publie. Le consensus sur les résultats scientifiques ne peut venir que de l'accumulation progressive de résultats convergents générée par l'ensemble de la communauté scientifique.

Finalement, les méthodes que nous avons présentées ici participent aussi à l'évolution du rôle de l'économiste qui conseille, assiste et évalue la décision publique. Longtemps l'économiste a surtout orienté la décision publique en proposant des grands principes robustes et rigoureux pour évaluer la décision publique. Avec la révolution de la crédibilité, et notamment avec les méthodes expérimentales, l'économiste s'intéresse aux détails de la mise en œuvre des politiques publiques. Il travaille en amont avec les décideurs pour concevoir le *design* de l'expérimentation et la mise en œuvre des évaluations d'alternatives politiques faisables et crédibles, entre lesquelles le décideur hésite réellement. Après l'avènement de l'économiste-ingénieur, nous assistons aujourd'hui à l'émergence de l'économiste-plombier (Duflo, 2017).

Bibliographie

Aichele R. et Felbermayr G. (2015). "Kyoto and Carbon Leakage: An Empirical Analysis of the Carbon Content of Bilateral Trade", *Review of Economic and Statistics*, vol. 97, n° 1, pp. 104-115.

Algan Y. et Cahuc P. (2007). *La Société de Défiance*, Rue d'Ulm.

Allcott H. (2015). "Site Selection Bias in Program Evaluation", *Quarterly Journal of Economics*, vol. 130, n° 3, pp. 1117-1165.

Angrist J.D. (1990). "Lifetime Earnings and the Vietnam era Draft: Evidence from Social Security Administrative Records", *American Economic Review*, vol. 80, n° 3, pp. 313-336.

Angrist J.D., Imbens G.W. et Rubin D.B. (1996). "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association*, vol. 91, n° 434, pp. 444-455.

Angrist J.D. et Krueger A.B. (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?", *Quarterly Journal of Economics*, vol. 106, n° 4, pp. 979-1014.

Angrist J.D. et Krueger A.B. (2001). "Instrumental Variables and the Search for Identification: from Supply and Demand to Natural Experiments", *Journal of Economic Perspectives*, vol. 15, n° 4, pp. 69-85.

Angrist J.D. et Lavy V. (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement", *Quarterly Journal of Economics*, vol. 114, n° 2, pp. 533-575.

Angrist J.D. et Pischke J.-S. (2010). "The Credibility Revolution in Empirical Economics: How Better Research is Taking the Con Out of Econometrics", *Journal of Economic Literature*, vol. 24, n° 2, pp. 3-30.

Angrist J.D. et Pischke J.-S. (2014). *Mastering Metrics, Princeton University Press.*

Athey S. (2015). "Machine Learning and Causal Inference for Policy Evaluation", *KDD'15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 5-6.

Ashraf N., Karlan D. et Yin W. (2006). "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines", *The Quarterly Journal of Economics*, vol. 121, n° 2, pp. 635-672.

Banerjee A.V. et Duflo E. (2009). "L'approche expérimentale en économie du développement", *Revue d'économie politique*, vol. 119, n° 5, pp. 691-726.

Banerjee A.V. et Duflo E. (2012). *Repenser la Pauvreté*, Seuil.

Banerjee A.V., Duflo E., Glennerster R. et Kothari D. (2010). "Improving Immunization Coverage in Rural India: Clustered Randomized Controlled Evaluation of Immunization Campaigns with and Without Incentives", *BMJ*, vol. 340, 5 pages.

Behaghel L. (2006). *Lire l'Économétrie*, La Découverte.

Behagel L., Crépon B. et Gurgand M. (2013). "Robustness of the Encouragement Design in a Two-Treatment Randomized Control Trial", *IZA Discussion*, 7447.

Behaghel L., Crépon B. et Gurgand M. (2014). "Private and Public Provision of Counseling to Job-Seekers: Evidence from a Large Controlled Experiment", *American Economic Journal*: *Applied Economics*, vol. 6, n° 4, pp. 152-174.

Behaghel L., Crépon B., Gurgand M. et Le Barbanchon T. (2015). "Please Call Again: Correcting Non-Response Bias in Treatment Effect Models", *Review of Economics and Statistics*, vol. 97, n° 5, pp. 1070-1080.

Behaghel L., Crépon B. et Le Barbanchon T. (2015). "Unintended Effects of Anonymous Resumes", *American Economic Journal: Applied Economics*, vol. 7, n° 3, pp. 1-27.

Bérard J. et Valdenaire M. (2014). *De l'éducation à l'insertion : 10 résultats du fonds d'expérimentation pour la jeunesse*, La Documentation Française.

Bléhaut M. et Rathelot R. (2014). "Expérimentation contrôlée contre appariement : le cas d'un dispositif d'accompagnement de jeunes diplômés demandeurs d'emploi", *Économie et Prévision*, n° 204-205, pp. 163-181.

Bloom H.S., Orr Larry L., Bell S.H., Cave G., Doolittle F., Lin W. et Bos J.M. (1997). "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study", *Journal of Human Resources*, vol. 32, n° 3, pp. 549-576.

Boiteux M. (2001). *Transports : Choix des Investissements et Coût des Nuisances*, Commissariat Général du Plan, La Documentation française.

Brodaty T., Crépon B. et Fougère D. (2007). "Les méthodes micro-économétriques d'évaluation et leurs applications aux politiques actives de l'emploi", *Économie et Prévision*, n° 177, pp. 93-118.

Burtless G. (1995). "The Case for Randomized Field Trials in Economic and Policy Research", *Journal of Economic Perspectives*, vol. 9, n° 2, pp. 63-84.

Cahuc P. et Zylberberg A. (2016). *Le Négationnisme Économique et Comment s'en Débarrasser*, Flammarion.

Card D. et Krueger A.B. (1994). "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania", *American Economic Review*, vol. 84, n° 4, pp. 772-793.

Chabé-Ferret S. (2014). "Commentaire sur l'article de Marianne Bléhaut et Roland Rathelot, "Expérimentation contrôlée contre appariement", *Économie et Prévision*, n° 204-205, pp. 183-191.

Chabé-Ferret S. (2015). "Analysis of the Bias of Matching and Difference-in-Difference Under Alternative Earnings and Selection Processes", *Journal of Econometrics*, vol. 185, n° 1, pp. 110-123.

Chabé-Ferret S. et Subervie J. (2013). "How Much Green for the Buck? Estimating Additional and Windfall Effects of French Agro-Environmental Schemes by DID-Matching", vol. 65, n° 1, pp. 12-27.

Chay K.Y. et Greenstone M. (2005). "Does Air Quality Matter? Evidence from the Housing Market", *Journal of Political Economy*, vol. 113, n° 2, pp. 376-424.

Comité national d'évaluation du RSA, rapport final (2011). Technical report, Revenu de Solidarité Active.

Crépon B., Duflo E., Gurgand M., Rathelot R. et Zamora P. (2013). "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment", *Quarterly Journal of Economics*, vol. 128, n° 2, pp. 531-580.

Deschenes O. et Greenstone M. (2011). "Climate Change, Mortality, and Adaptation: Evidence from Annual Fluctuations in Weather in the U.S.", *American Economic Journal: Applied Economics*, vol. 3, n° 4, pp. 152-185.

Desplatz R. et Ferracci M. (2016). "Comment évaluer l'impact des politiques publiques ? Un guide à l''usage des décideurs et des praticiens", France Stratégie.

D'haultfoeuille X., Durrmeyer I. et Février P. (2011). "Le coût du bonus/malus écologique : que pouvait-on prédire ? ", *Revue économique*, 62 pages.

Di Nardo J. et Lee D. (2011). "Program Evaluation and Research Designs", in *Handbook of Labor Economics*, 4a, Elsevier.

Dominici F., Greenstone M. et Sunstein C.R. (2015). "Particulate Matter Matters", *Science,* vol. 344, n° 6181, pp. 257-259.

Duflo E. (2001). "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment", *American Economic Review*, vol. 91, n° 4, pp. 795-813.

Duflo E. (2017). "The Economist As Plumber", *Richard Ely Lecture*, mimeo, MIT.

Duflo E., Glennerster R. et Kremer M. (2008). "Using Randomization in Development Economics Research: A Toolkit", in volume 4 *of Handbook of Development Economics*, chap. 61, pp. 3895-3962. Elsevier.

Duflo E. et Pande R. (2007). "Dams", *Quarterly Journal of Economics*, vol. 122, n° 2, pp. 601-646.

Erkel-Rousse H. (2014). "Méthodes d'évaluation des politiques publiques : introduction générale", *Économie et Prévision*, n° 204-205, pp. I-XII.

Ferracci M. et Wasmer E. (2011). *État Moderne, État Efficace*, Odile Jacob.

Fuchs-Schuendeln N. et Hassan T.A. (2015). "Natural Experiments in Macroeconomics", *NBER Technical Working Papers* 21228.

Givord P. (2014). "Méthodes économétriques pour l'évaluation des politiques publiques", *Économie et Prévision*, n° 204-205, pp. 1-28.

Grenet J. (2010). "Academic Performance, Educational Trajectories and the Persistence of Date of Birth Effects. Evidence from France", mimeo, Paris School of Economics.

Heckman J.J. (1992). "Randomization and Social Policy Evaluation", dans *Evaluating Welfare and Training Programs*, édité par C. F. Manski et I. Garfinkel, Harvard University Press, pp. 201-230.

Heckman J.J., Ichimura H., Smith J.A. et Todd P.E. (1998). "Characterizing Selection Bias Using Experimental Data", *Econometrica*, vol. 66, n° 5, pp. 1017-1099.

Heckman J.J., LaLonde, R.J. et Smith J.A. (1999). "The Economics and Econometrics of Active Labor Market Programs", in vol. 3 of du *Handbook of Labor Economics*, chap. 31, pp. 1865-2097, Elsevier, North Holland.

Heckman, J.J. (2001). "Micro Data, Heterogeneity and the Evaluation of Public Policy: Nobel Lecture", *Journal of Political Economy*, vol. 109, n° 4, pp. 673-748.

Imbens G.W. (2010). "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)", *Journal of Economic Literature*, vol. 48, n° 2, pp. 399-423.

Imbens G.W. et Rubin D.B. (2015). *Causal Inference for Statistics, Social and Biomedical Sciences*, Cambridge University Press.

Kohavi R., Longbotham R., Sommerfeld D. et Henne R.M. (2009). "Controlled Experiments on the Web: Survey and Practical Guide", *Data Mining and Knowledge Discovery*, vol. 18, n° 1, pp. 140-181.

Kramer M.S. et Shapiro S.H. (1984). "Scientific Challenges in the Application of Randomized Trials", *JAMA*, vol. 252, n° 19, pp. 2739-2745.

LaLonde R.J. (1986). "Evaluating the Econometric Evaluation of Training Programs with Experimental Data", *American Economic Review,* vol. 76, n° 4, pp. 604-620.

Legendre F. (2013). "Une introduction à la micro-économétrie de l'évaluation", *Revue Française d'Économie*, vol. 28, n° 1, pp. 9-41.

Les membres du Conseil d'Analyse Économique (2013). *Note n°1 du Conseil d'Analyse Économique*, 12 pages.

Levitt S. et List J. (2008). "Field Experiments in Economics: The Past, the Present, and the Future," mimeo.

L'Horty Y. et Petit P. (2011). "Évaluation aléatoire et expérimentations sociales", *Revue Française d'Économie*, vol. XXVI, pp. 13-48.

Magnac T. (2000). "L'apport de la microéconométrie à l'évaluation des politiques publiques", *Cahiers d'Économie et Sociologie Rurales*, n° 54, pp. 89-113.

McFadden D. (2001). "Economic Choices", *American Economic Review*, vol. 91, n° 3, pp. 351-378.

Miguel E. et Kremer M. (2001). "Worms: Education and Health Externalities in Kenya", *NBER Working Papers* 8481, September.

Okbani N. (2013). "Le non recours au RSA activité : étude auprès des allocataires de la CAF de la gironde", *Technical report*, CAF de la Gironde.

Piketty T. et Valdenaire M. (2006). "L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français", *Les dossiers - Enseignement scolaires*, n° 173.

Quinet E. (2013). *Évaluation Socio-Économique des Investissements Publics*, Commissariat Général à la Stratégie et à la Prospective, La Documentation Française.

Rosenthal R. (1966). "Experiment Effects in Behavioral Research", *Appeton-Century-Crofts*, 464 pages.

Roux S. (2015). "Approches structurelles et non structurelles en micro-économétrie de l'évaluation des politiques publiques", *Revue Française d'Économie*, vol. 30, n° 1, pp. 13-65.

Sianesi B. (2017). "Evidence of Randomisation Bias in a Large-Scale Social Experiment: The Case of ERA", *Journal of Econometrics*, vol. 198, n° 1, pp. 41-64.

Thistlethwaite D.L. et Campbell D.T. (1960). "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment", *Journal of Educational Psychology*, vol. 51, n° 6, pp. 309-317.

Todd P.E. (2007). "Evaluating Social Programs with Endogenous Program Placement and Selection of the Treaded", in *Handbook of Development Economics*, vol. 4, Elsevier.

Todd P.E. et Wolpin K.I. (2006). "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility," *American Economic Review*, vol. 96, n° 5, pp. 1384-1417.