CAHIERS

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

DOCUMENTS DE TRAVAIL DE LA DG TRÉSOR

Numéro 2015/01 – Février 2015

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

Laura DUPONT-COURTADE

et

Sylvain CHABÉ-FERRET
Nicolas TREICH
Martine PERBET

TRÉSOR
DIRECTION GÉNÉRALE

ÉVALUATION DES POLITIQUES PUBLIQUES : EXPÉRIMENTATION RANDOMISÉE ET MÉTHODES QUASI-EXPÉRIMENTALES

Laura DUPONT-COURTADE*

et

Sylvain CHABÉ-FERRET,
Nicolas TREICH, Martine PERBET*

Ce document de travail n'engage que ses auteurs. L'objet de sa diffusion est de stimuler le débat et d'appeler commentaires et critiques

* **Laura DUPONT-COURTADE** Toulouse School of Economics - Master in Economics - Public Policy and Development (PP&D), Sous la direction de : Sylvain Chabé-Ferret (France)
lauradupontcourtade@gmail.com

* **Sylvain CHABÉ-FERRET** Toulouse School of Economics - LERNA-INRA
Sylvain.chabe-ferret@.tse-fr.eu

* **Nicolas TREICH** Toulouse School of Economics - LERNA-INRA
nicolas.treich@tse-fr.eu

***** **Martine PERBET** est Rédactrice en chef des "*Cahiers de l'évaluation*" en poste à la Direction Générale du Trésor du Ministère des Finances et des comptes Publics et du Ministère de l'Économie, de l'Industrie et du Numérique (France)
martine.perbet@dgtresor.gouv.fr (+33-1-44-87-18-45)

TRÉSOR DIRECTION GÉNÉRALE

Remerciements

Ce document a été réalisé dans le cadre d'un stage de fin de Master[1]. Effectué en partenariat avec le ministère de l'Économie et des Finances (Direction Générale du Trésor) et l'INRA[2], la rédaction de ce document a été encadrée par Sylvain Chabé-Ferret et Nicolas Treich.

Ce stage a été initié par Martine Perbet dans la perspective de la publication d'un nouveau numéro des *Cahiers de l'évaluation*. Cette revue publie des dossiers thématiques sur l'évaluation économique des politiques publiques[3].

Durant ce stage, certaines personnes m'ont beaucoup aidée et je tiens à les remercier ici.

Je tiens, dans un premier temps, à remercier Sylvain Chabé-Ferret. Enseignant la matière *Program Evaluation Methods* dans le master que j'ai suivi, il a su, grâce à la grande qualité de ses cours, étayer mon goût pour les problématiques d'évaluation et m'a ainsi donné l'envie de candidater pour ce stage. Je le remercie de m'avoir donné l'opportunité d'écrire ce document. Je le remercie pour son soutien tout le long de ce stage, pour la clarté de ses explications et pour ses remarques toujours pertinentes. J'adresse ensuite mes remerciements à Nicolas Treich, qui m'a également permis de réaliser ce stage, pour avoir su me guider dans ce monde inconnu qu'était pour moi celui de l'évaluation *ex-ante*. Je le remercie pour la grande qualité de ses conseils.

Par ailleurs, ce document ne serait rien sans Martine Perbet, rédactrice en chef des Cahiers de l'évaluation, qui est à l'origine de ce stage. Je la remercie pour toutes ses contributions à ce document.

À l'occasion de ce document, je tiens à remercier également mon responsable de master Jean-Paul Azam. Je le remercie pour ses cours passionnants d'économie du développement, vecteurs de mon orientation, sans lesquels je n'aurais pas réalisé ce document.

Mes remerciements vont également à toute l'équipe du LERNA pour leur accueil et leur disponibilité. Je remercie enfin toutes les personnes m'ayant encouragée au cours de ce travail, avec une attention particulière à Florimond Bourdeaux, pour son attentive relecture, ainsi qu'à Quentin Villotta dont la bonne humeur et le soutien ont sans nul doute contribué positivement à la réalisation de ce document.

****

"Lorsque j'ai demandé que la discussion de notre rapport d'évaluation de la loi LRU intervienne avant la prochaine réforme des universités, je me suis vu expliquer que les débats allaient se percuter. Pourtant, quel sens y aurait-il de discuter de la loi précédente après l'adoption de la suivante ? "

David Assouline

Président de la commission sénatoriale pour le contrôle de l'application des lois
Compte rendu du mardi 16 avril 2013
http://www.senat.fr/compte-rendu-commissions/20130415/applois.html

[1] Master in Economics - Public Policy and Development. Toulouse School of Economics (TSE). Directeur de Master : Jean-Paul Azam.

[2] Institut national de la recherche agronomique - Laboratoire LERNA (Laboratoire d'économie des ressources naturelles) - http://www.tse-fr.eu/lerna/

[3] http://www.tresor.economie.gouv.fr/Cahiers-de-levaluation

Table des matières

Résumé

L'évaluation des politiques publiques s'ancre dans les bonnes pratiques des pays développés. L'objectif est que toute nouvelle politique ne soit adoptée qu'une fois ces avantages et inconvénients bien pesés (analyses coût-bénéfice). Cependant, il est parfois difficile d'estimer les avantages à attendre de nouvelles mesures, comme par exemple les aides au retour à l'emploi ou les incitations à la scolarisation des enfants tant les réactions des individus sont tributaires de leurs caractéristiques personnelles. D'où l'intérêt de n'appliquer ce type de mesures qu'à des groupes restreints, dans un premier temps, afin de bien identifier *in vivo* les changements de comportement susceptibles d'en découler. De telles expérimentations de politiques publiques ont été tenté, aux États-Unis, dès les années soixante, avant de se diffuser à l'international.

L'évaluation expérimentale d'une mesure utilise le même principe que les essais cliniques: au sein de la population deux groupes sont sélectionnés par tirage aléatoire, l'un bénéficie du traitement (*i.e.* de la mesure) et l'autre non. L'impact du traitement (*i.e.* de la mesure) s'obtient en comparant ex-post le groupe d'agents bénéficiaires au groupe de non bénéficiaires. Le caractère aléatoire de l'allocation de la mesure garantit que les deux populations sont identiques *ex-ante*. Ainsi, seul le traitement (*i.e.* la mesure) fait la différence entre les deux populations. 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 adéquats.

Faute de pouvoir réaliser une expérimentation avec tirage au sort (*i.e.* expérimentation randomisée), on utilise parfois des données d'observation préexistantes pour reproduire les résultats expérimentaux (économétrie sur données individuelles). Ces méthodes « quasi-expérimentales » sont commodes car elles mobilisent moins de moyens et permettent d'éviter les problèmes éthiques et politiques que peut induire une allocation randomisée. Il reste cependant à déterminer les conditions pour que ces « quasi-expériences » constituent un bon substitut à une expérimentation randomisée.

Abstract

Public policy evaluation is considered to be part of the best practice in developed countries. The ambition is that any new policy should be adopted only once its pros and cons have been carefully compared (e.g. using cost-benefit analysis). However, it is generally difficult to estimate the expected net benefit of a new policy (for example of a Job Training Program or of a Schooling program), in particular because of the difficulty to anticipate the behavioural response of individuals. One possible solution is to target such a policy, first, to small groups in order to identify *in vivo* the behavioural changes that may arise. Such public policy experiments have been first employed in the United States in the sixties, before spreading internationally.

Experimental program evaluation is based on the same principle as clinical experiments: two groups are selected by randomly drawing among a population, with only one of them benefiting from the treatment (*i.e.* the program). The impact of the treatment (*i.e.* the program effect) is obtained by an ex-post comparison of the groups, the recipients and the non-recipients. Indeed, the randomness of the program allocation guarentees that the two groups have identical characteristics *ex-ante*. Thus, the remaining difference between the two groups is the effect of the program. In practice, there exist various designs of randomized experiments that can adapt to the specificities of each policy.

However, sometimes the use of randomization may not be possible. In this case one may want to use the existing observational data in order to reproduce experimental results (econometrics methods on individual data). These "quasi-experimental" methods are convenient because they require fewer resources and avoid ethical and political problems which may occur with a randomized program allocation. However, how these "quasi-experiments" can be good substitutes for a randomized experiment still needs to be documented.

Introduction

L'évaluation permet l'apprentissage et donc le progrès, c'est-à-dire l'amélioration progressive des politiques. Bien que l'utilité d'un système d'évaluation soit reconnue à la fois par les décideurs publics et les experts, l'évaluation reste encore peu développée en France. L'obligation d'études d'impact des projets de loi instaurée par la Loi organique n° 2009-403[4] va dans le bon sens mais les méthodes économiques restent encore peu mobilisées. À cet égard la France gagnerait sans doute à s'inspirer des pratiques de pays plus avancés comme, par exemple, les États-Unis.

Une défiance historique envers l'État fédéral a doté ce pays d'une véritable culture de l'évaluation[59]. À partir du New deal, l'envolée des dépenses publiques a contraint les gouvernements successifs à rechercher l'adhésion des électeurs en justifiant (*value for money*) leurs politiques. Dès les années soixante, les avancées académiques ont permis un recours intensif à l'évaluation des politiques publiques en utilisant à la fois des analyses coûts / bénéfices et des expérimentations. Les analyses coûts / bénéfices sont d'abord devenues obligatoires pour les dépenses publiques (*Planning and Programming Budgeting System*) puis, le coût social des réglementations étant mieux connu, elles le sont devenues également pour les projets de règlementation des agences de l'exécutif (*Executive order*)[5]. Les expérimentations ont, elles, été utilisées très tôt pour les programmes sociaux[55], la mise en place d'un revenu minimum a, par exemple, été testée dès la fin des années soixante (*cf.* encadré 2)[38].

Ce sont les expérimentations qui sont au cœur de ce travail. Le principe des expérimentations consiste à utiliser le cadre statistique des tests médicaux (tirage aléatoire de deux populations, l'une bénéficiant du traitement, l'autre non (groupe témoin)) pour évaluer les effets de politiques publiques sur des individus. Si ces méthodes sont aujourd'hui communément admises aux États-Unis, suite à une longue pratique[6], il n'en est pas de même en France où leur introduction en tant qu'aide à la décision publique date de 2007. Deux mesures phares des politiques sociales font alors l'objet d'une expérimentation cette année-là, le « Revenu de solidarité active » (RSA)[51] et « l'accompagnement renforcé des demandeurs d'emploi »[52] mais, depuis, l'expérimentation ne s'est pas véritablement ancrée dans les pratiques d'évaluation des politiques publiques malgré de nouvelles tentatives intéressantes comme le « Fonds d'expérimentations pour la jeunesse ». La France est encore dans une période d'apprentissage, la progression des expérimentations étant limitée par divers facteurs dont, notamment, les problèmes éthiques et politiques que soulèvent, dans notre culture, le tirage aléatoire des échantillons (*cf.* encadrés 1, 3 et 5)[51] [3].

L'objectif est ici de préciser le concept d'expérimentation. Après un rappel des difficultés rencontrées pour identifier une relation causale entre une politique et un résultat (**chapitre 1**), ce document présente les différents designs d'expérimentations permettant d'établir cette relation (**chapitre 2**). Faute de pouvoir réaliser une expérimentation, on utilise parfois des « quasi-expériences », c'est-à-dire des méthodes économétriques visant à reproduire les résultats expérimentaux à partir de données d'observation. Le document présente un panorama de ces méthodes et en esquisse un bilan empirique (**chapitre 3**).

[4] http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000020521873dateTexte= categorieLien=id

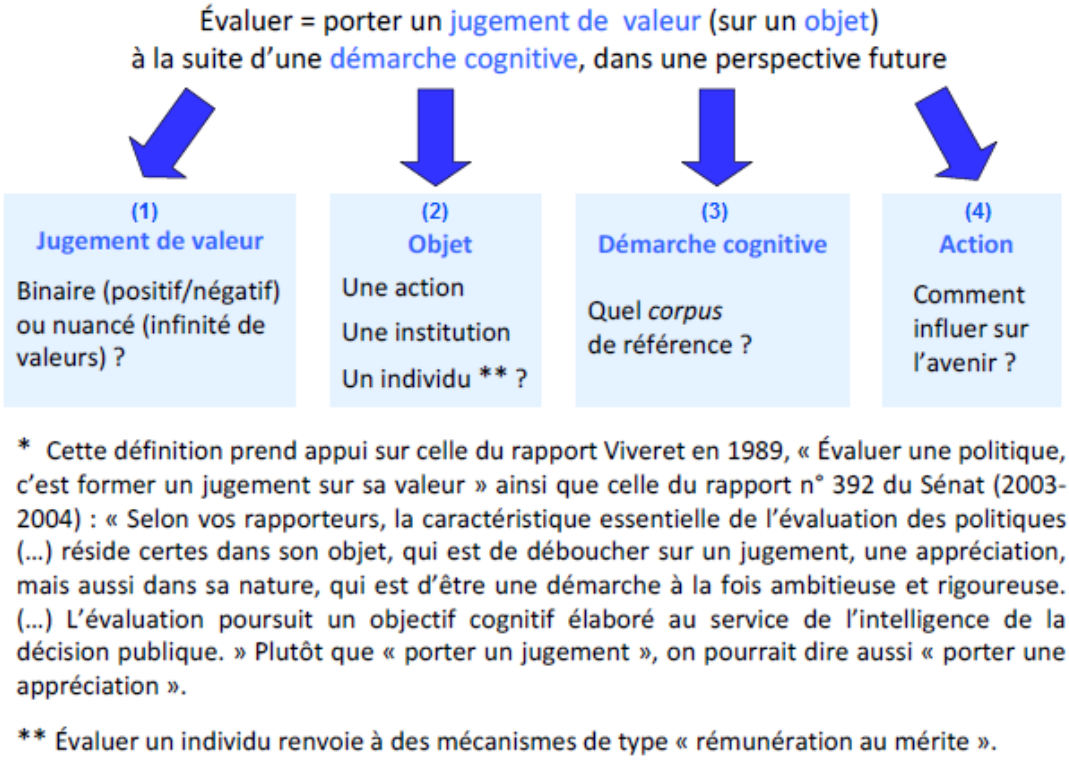
[5] Les études d'impact des réglementations sont imposées par des *Executive Orders* de la Présidence depuis 1975: 11821 (1975), (1978), 12291 (1981) et 12866 (1993). Pour plus de détails voir page 60 du rapport d'Eliana Valles, « L'évaluation aux États-Unis » (2012), TSE. Ce rapport de stage, réalisé sous la direction de Martine Perbet, est disponible sur le site de la Direction générale du Trésor, https://www.tresor.economie.gouv.fr/6 6 rapports-des-stagiaires.

[6] Le gouvernement américain soulignait, en 2003 sa volonté de donner priorité aux projets éducatifs ayant fait l'objet d'expérimentations (*Federal Register* 2003).

Chapitre 1 - Le problème fondamental d'inférence causale et les biais des comparaisons intuitives

Définissons tout d'abord la notion d'évaluation. Évaluer signifie porter un jugement de valeur sur un objet à la suite d'une démarche cognitive dans une perspective future[2]. La figure 1.1 décompose et détaille les différents éléments de cette définition.

Figure 1.1 – Définition de l'évaluation



* Cette définition prend appui sur celle du rapport Viveret en 1989, « Évaluer une politique, c'est former un jugement sur sa valeur » ainsi que celle du rapport n° 392 du Sénat (2003-2004) : « Selon vos rapporteurs, la caractéristique essentielle de l'évaluation des politiques (…) réside certes dans son objet, qui est de déboucher sur un jugement, une appréciation, mais aussi dans sa nature, qui est d'être une démarche à la fois ambitieuse et rigoureuse. (…) L'évaluation poursuit un objectif cognitif élaboré au service de l'intelligence de la décision publique. » Plutôt que « porter un jugement », on pourrait dire aussi « porter une appréciation ».

** Évaluer un individu renvoie à des mécanismes de type « rémunération au mérite ».

Source : *Cahiers de l'évaluation* n°4.

En référence à cette définition, ce document s'intéresse à l'évaluation des politiques publiques (objet) à partir de l'analyse économique et, plus précisément à partir de méthodes économétriques sur données individuelles (démarche cognitive). L'objectif est de savoir si la politique considérée a bien produit les effets attendus (jugement de valeur) afin de pouvoir conseiller le décideur public pour une éventuelle réorientation de cette politique (action).

Estimer les effets d'une politique implique de comparer une situation avec la politique par rapport à une situation sans cette politique (situation dite de référence). Dans une démarche d'évaluation *ex-ante* (figure 1.2), la situation de référence et la situation en présence de la politique sont inobservées. Le travailleur de l'évaluateur est de simuler ces deux situations.

Dans une démarche d'évaluation *ex-post* (figure 1.3), la situation en présence de la politique est observée. La situation de référence, ce qu'il se serait passé dans la situation hypothétique où la politique n'aurait pas été introduite, n'est pas observée. En conséquence, elle est qualifiée de « contrefactuelle [36] ».

La principale difficulté de l'évaluation *ex-post* est la reconstitution de la situation de référence. L'évaluateur fait face à un manque fondamental de données[63] : 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 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 par la situation qui préexistait avant la mise en place de la politique ou par la situation des individus ne bénéficiant pas de la politique. Malheureusement, ces deux comparaisons intuitives sont généralement biaisées. Les sources de biais de ces deux approches sont décrites dans les sections suivantes.

Figure 1.2 – Observabilité des situations et démarche cognitive dans une évaluation *ex-ante*

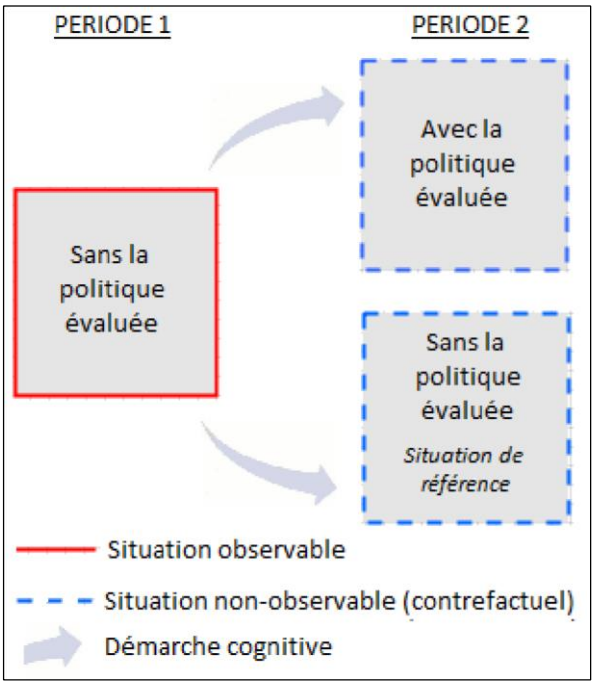
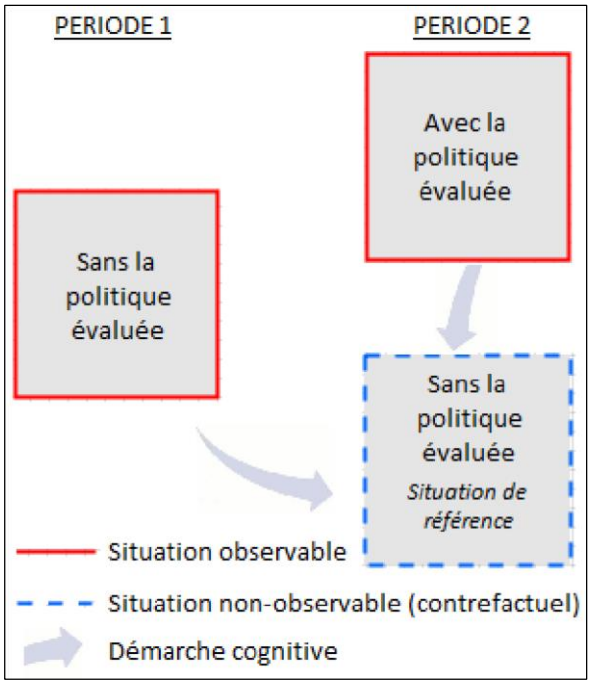


Figure 1.3 – Observabilité des situations et démarche cognitive dans une évaluation *ex-post*



Les concepts développés dans ce document s'appuient sur quatre exemples d'évaluation : la politique du revenu de solidarité active (RSA), importante réforme des minima sociaux succédant au revenu minimum d'insertion (RMI) en 2009[36] ; un projet de construction de barrage[7] ; un programme social de formation professionnelle et une campagne de vaccination.

1.1 La comparaison avant / après

Prenons l'exemple du RSA. Pour évaluer l'impact du RSA, une première approche (simplissime) 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 en place, par rapport au taux d'activité en vigueur aujourd'hui avec le RSA. On appelle cette méthode une comparaison avant / après.

Cette approche est insatisfaisante car d'autres facteurs que le RSA ont influencé 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.

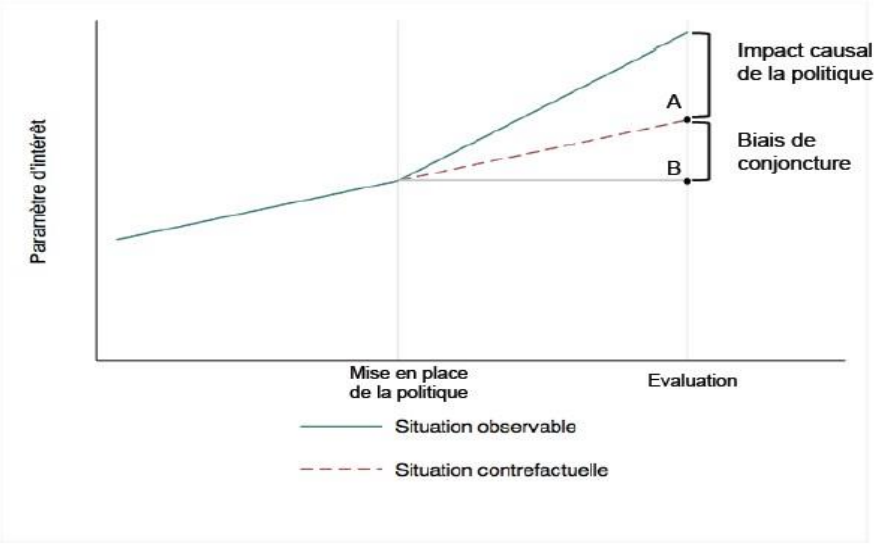
Autre exemple, la construction d'un barrage. On peut s'intéresser à l'impact du barrage sur la productivité agricole en comparant la situation initiale à celle 10 ans après sa construction.

[7] À titre de référence Esther Duflo et Rohini Pande ont publié Dams, un article étudiant les impacts des barrages sur la production agricole et la pauvreté en Inde[33].

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 (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 1.4 : Biais de conjoncture et effet réel



La figure 1.4 représente l'effet réel 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 le biais de conjoncture. Dans le cas présent, cela mènerait à une surestimation de l'impact 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 RMI 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 : conditions météorologiques, technologie de production et dépenses publiques jouent aussi un rôle. Il faut donc être très prudent avec ce type d'analyse qui peut aisément mener à des conclusions erronées.

1.2 La comparaison avec / sans

Dans la mesure où une politique n'est pas mise en œuvre sur l'ensemble du territoire mais seulement sur des zones restreintes, une méthode alternative pour évaluer cette politique consiste à comparer un groupe d'agents bénéficiaires (de cette politique) à un groupe de non bénéficiaires (de cette politique). 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.

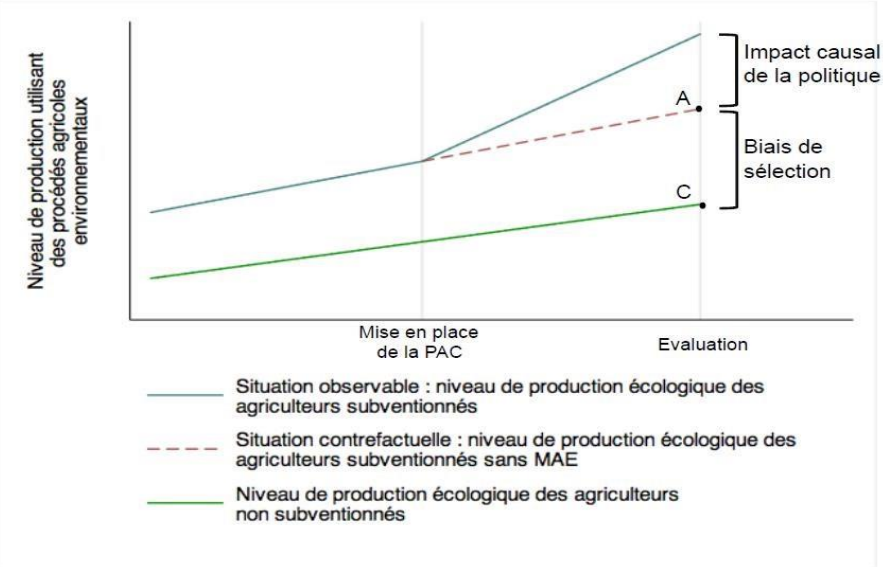
Cependant, pour être comparées, deux populations doivent être comparables. Ce n'est malheureusement généralement pas le cas. En effet, la différence entre un groupe bénéficiaire et un groupe de non bénéficiaires est due, d'une part à l'effet de la politique et d'autre part à la différence initiale entre les deux groupes. La différence initiale entre ces deux groupes est connue sous le nom de **biais de sélection** et doit absolument être prise en compte dans une évaluation afin d'éviter des résultats biaisés.

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 afin d'évaluer l'impact d'un nouveau barrage. Imaginons également que ce projet d'infrastructure vise les zones rurales les plus pauvres afin de favoriser leur développement. Si la zone de comparaison est une zone dynamique avec un niveau de pauvreté plus bas, une simple comparaison de cette zone avec la zone de traitement conduirait à une sous-estimation de l'impact causal de la politique sur la pauvreté.

Prenons un dernier exemple. L'un des objectifs des mesures agro-environnementales (MAE) de la politique agricole commune, la PAC, est d'encourager l'utilisation de méthodes de production agricole écologiques. Des subventions sont versées aux agriculteurs qui emploient ces technologies. Supposons que l'on veuille évaluer l'impact de ces subventions sur le niveau d'utilisation de méthodes agricoles écologiques[37]. On cherche donc la différence entre le niveau d'utilisation lorsque les agriculteurs sont subventionnés et le niveau d'utilisation lorsqu'ils ne le sont pas qui représente le contrefactuel. Cette situation contrefactuelle est représentée par la droite bleue figure 1.5 ci-après. Lors de l'évaluation 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 très probable qu'un nombre substantiel d'agriculteurs recevant les subventions utilisait 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 figure 1.5, surestime donc a priori l'effet causal de ces subventions sur le niveau d'utilisation des méthodes environnementales parce que le comportement des agriculteurs en absence des subventions n'est pas pris en compte.

Figure 1.5 – Biais de sélection et effet réel : exemple de la PAC



On voit donc bien que l'estimation d'effets causaux n'est pas triviale. Les comparaisons avant / après et avec / sans paraissent intuitives mais ne constituent généralement pas d'outils d'évaluation valides.

1.3 Un exemple numérique

Les biais des comparaisons avant / après et avec / sans se comprend mieux à partir d'un exemple numérique. Les programmes de formation professionnelle font l'objet de nombreuses études dans la littérature économique concernant l'évaluation des politiques publiques[7]. Nous prendrons l'un d'entre eux comme exemple pour illustrer numériquement la comparaison avant / après et la comparaison avec / sans et montrer pourquoi ces méthodes d'évaluation avant / après mènent à des résultats biaisés. Ce même exemple sera repris plus loin pour illustrer les expériences randomisées.

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. Les individus éligibles au programme sont les individus au chômage au moment où le programme est mis en place. Prenons comme mesure d'activité (retour à l'emploi), le pourcentage d'individus ayant un emploi après 6 mois.

Afin d'exposer de façon claire les biais potentiels induits par les méthodes présentées ci-dessus, 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. Il peut s'agir d'individus se trouvant dans une situation de chômage de longue durée, d'individus avec un niveau de revenu extrêmement faible ou encore 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 rouge figure 1.6. 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. Par exemple, cela peut inclure les individus plus jeunes, les individus pour qui les aides sociales constituent une stigmatisation ou ceux ayant un niveau de qualification élevé. Ces individus ont une probabilité de trouver un emploi supérieure à celle des individus de type 1[8]. Le taux d'activité au sein de ce groupe d'individus est de 90 % en absence du programme. Ils sont signalés par un point bleu figure 1.6. 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 1.6 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.

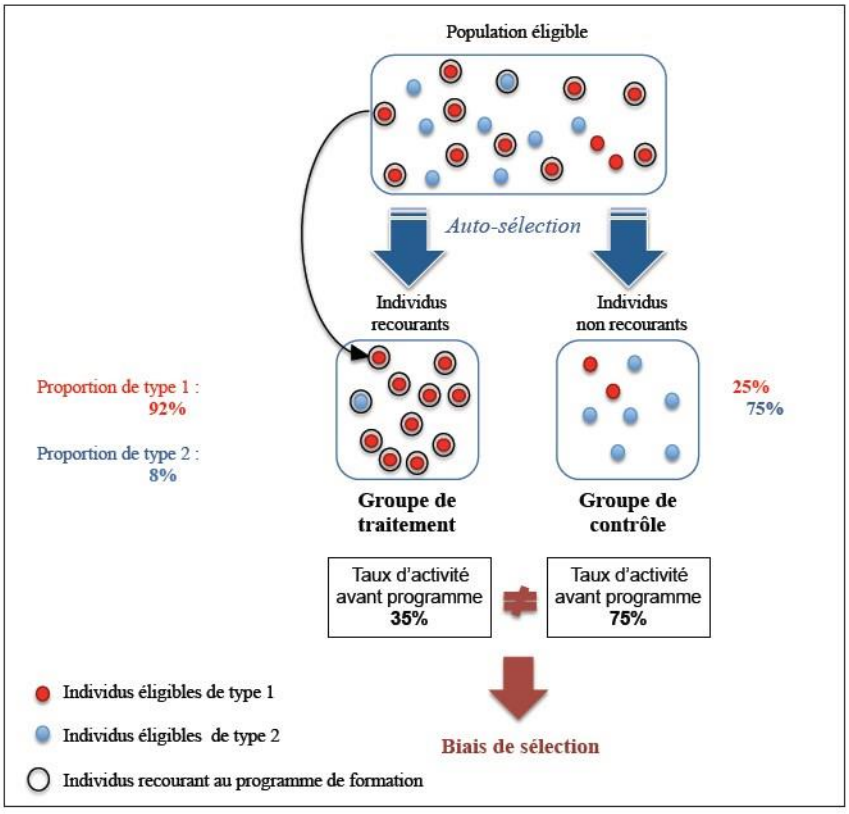
Procéder à l'évaluation de ce programme par une comparaison avec / sans revient à comparer le niveau d'activité *ex-post* d'un groupe de recourants au niveau d'activité d'un groupe d'individus non recourants. Le groupe de traitement et le groupe de contrôle étudiés sont formés par un système d'auto-sélection, autrement dit, 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.

[8] Un individu de type 1 et un individu de type 2 se distinguent par leurs caractéristiques sociales, néanmoins leur décision d'entrer dans le programme de formation dépend également de facteurs indépendant de leur type. Leur connaissance du programme ou bien encore les aides sociales qu'ils peuvent percevoir déterminent aussi s'ils auront recours au programme de formation.

Le tableau 1 montre que 80 % des individus de type 1 souscrivent au programme, alors que seul 10 % des individus de type 2 y ont recours. Le groupe de recourants est ainsi constitué de 92 % d'individus de type 1. Le groupe de non recourants, lui, est constitué d'une majorité d'individus de type 2 représentant ainsi 75 % du groupe. Le groupe de non recourants utilisé comme groupe de référence comprend davantage d'individus, qui intrinsèquement, ont une plus grande probabilité de retrouver un emploi. Ce groupe ne forme donc une bonne approximation de la situation contrefactuelle.

Ce problème de sélection biaise la comparaison avec / sans qui conclue que le programme de formation professionnelle impacte négativement le niveau d'activité des bénéficiaires. Plus précisément, avec les données établies tableau 1, on conclut que le programme diminue de 12 points de pourcentage le taux d'activité moyen alors que l'effet réel est une augmentation du taux d'activité de 2 points de pourcentage. L'effet du type domine l'effet du programme, autrement dit, la comparaison avec / sans 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**.

Figure 1.6 – Comparaison avec / sans : l'exemple d'un programme de formation professionnelle



Cet exemple témoigne bien des limites de la comparaison avec / sans comme méthode d'évaluation. La comparaison avec / sans serait valide si les deux types d'individus se répartissaient dans les mêmes proportions dans le groupe de traitement et dans le groupe de contrôle.

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 socioéconomique défavorable ou du programme lui-seul ? L'exemple numérique 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 l'effet du programme de formation.

Tableau 1 - Comparaison avec/sans et comparaison avant / après : L'exemple d'un programme de formation

	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

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

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

COMPARAISON AVEC/SANS				
	Groupe des recourants	**Groupe des non recourants**		
Proportion de Type 1 dans le groupe	0,92	0,25		
Proportion de Type 2 dans le groupe	0,08	0,75		
Taux d'activité avant randomisation (t = 0)	0,35	0,75		
Taux d'activité avec programme (t = 1)	0,53	0,80		
Taux d'activité sans programme (t = 1)	0,25	0,65		
Impact causal moyen réel sur les recourants (ICM)		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

COMP. AVANT/APRES	
Groupe de recourants	
0,92	
0,08	
0,35	
0,53	
0,25	
0,28	
0,18	
-0,10	

	= Non observable

Chapitre 2 - 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 du développement.

2.1 Définition et randomisation

2.1.1 Définition

Une expérience randomisée consiste à comparer deux groupes formés **aléatoirement** à partir d'un échantillon d'individus[9] : un groupe de traitement[10], au sein duquel les individus sont sujets à une intervention expérimentale et un groupe de contrôle[11] qui est utilisé comme groupe de comparaison.

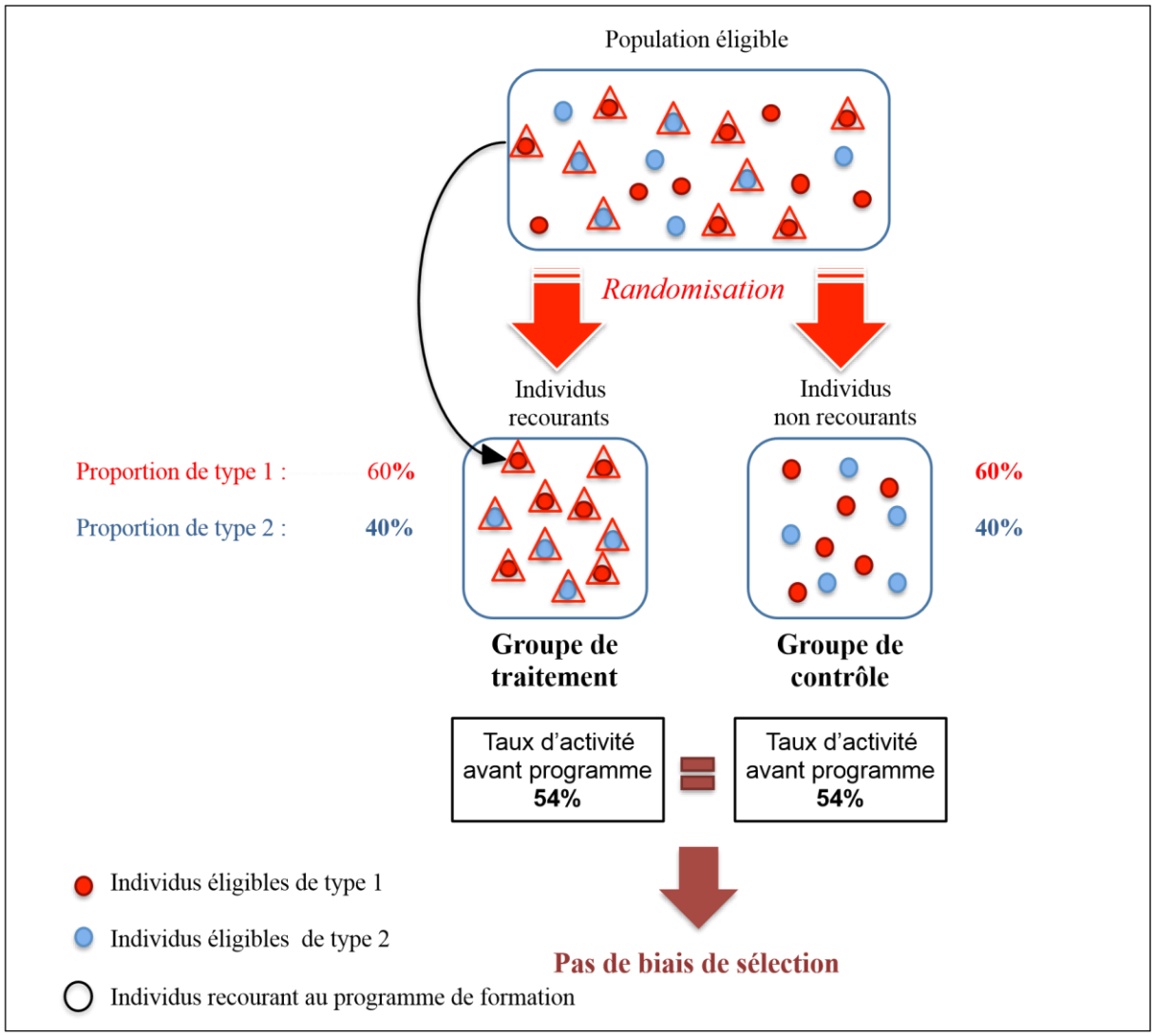
2.1.2 Le rôle de la randomisation

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 garantit que les individus ayant des caractéristiques différentes sont répartis 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 obtiendra la même proportion de chaque type dans chaque groupe. Cette méthode permet d'obtenir des groupes de traitement et de contrôle 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 section 1.3, 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. Les deux groupes deviennent comparables 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 2.1 schématise ce mécanisme[53][19].

[9] Une expérience randomisée n'est pas nécessairement exécutée au niveau individuel : des ménages, des régions ou encore des villages peuvent tout aussi bien être alloués à ces différents groupes.
[10] La notion de traitement réfère à une politique ou à un programme. Être traité signifie appartenir au groupe de traitement, population soumise à la politique évaluée. Être non traité signifie appartenir au groupe de contrôle, le groupe de comparaison non soumis à la politique.
[11] Une expérience randomisée peut très bien concevoir plusieurs groupes de traitement et plusieurs groupes de contrôle.

Figure 2.1 – Design 1 : Randomisation d'un traitement, exemple d'un programme de formation professionnelle



En pratique, il existe plusieurs manières de mettre en œuvre une expérience randomisée. Chacune de ces approches 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 décrit à la section 1.4. La dernière section de ce chapitre présente finalement les différentes limites que l'on rattache à l'expérimentation randomisée et les solutions existantes.

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 [*cf.* encadré 2], 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, 2009). […]. [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 6M€. 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*.

*En effet « Dans le cadre de l'expérimentation du RSA, ni la liste des départements expérimentateurs, ni le périmètre des zones tests dans chaque département, ni la liste des allocataires du RSA n'ont été choisis au hasard. Les départements, qui sont les véritables pilotes du RMI depuis la loi de décentralisation de décembre 2003, ont défini le périmètre des zones test sur la base d'une sélection raisonnée, en appliquant des critères et selon des contraintes qui leur sont propres. À l'intérieur de ces zones, tous les allocataires du RMI bénéficient du RSA. L'expérimentation du RSA s'inscrit donc dans le registre des quasi-expériences. Chaque département réalise une expérience qui est au niveau national répétée plus de trente fois, selon des modalités qui varient à la marge. On est bien dans la situation la plus courante de l'évaluation des politiques publiques où les bénéficiaires de la politique ne font pas l'objet d'un tirage au sort. La définition des zones témoins a été adaptée en conséquence » (Goujard et L'Horty, 2010)

Source : L'Horty Y., Petit P. (2010), « Évaluation aléatoire et expérimentations sociales », *document de travail* n° 135, Centre d'études de l'emploi, décembre[38].

2.2 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 illustré figure 2.1 et les résultats de cette évaluation sont présentés 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 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 $t = 0$ où 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 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 $t=1$.

2.2.1 Application numérique

D'après notre exemple numérique présenté 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 $\mathbf{0.6} * 0.3 + \mathbf{0.4} * 0.1 = 0.22$, correspondant à (1) − (3) dans le tableau 4.

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 (1) − (4) dans le tableau 2 soit $0.66 - 0.44 = 0.22$. On retrouve exactement l'impact causal moyen réel.

Tableau 2 – Design 1 : randomisation dans l'ensemble de la population

	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	0,6
Proportion de Type 2 dans le groupe	0,4	0,4
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	**0** (ICM - ICM estimé)	
Groupe utilisé pour simuler le contrefactuel	Groupe des non traités après randomisation (Groupe de contrôle en *t = 1*)	

= Non observable

2.2.2 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 étudiés comme *traités* alors qu'ils ne l'auraient pas été dans une version non expérimentale 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 sur 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 réagissaient de la même la 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. Les parties suivantes les présentent.

Encadré 2 - *Negative Income Tax Experiment*

«Le *Negative Income Tax Experiment* a été l'une des premières expérimentations sociales conduites aux États-Unis sur une grande échelle. Cette expérience a été réalisée à la fin des années 1960, dans un climat politique marqué par les revendications des mouvements de gauche réclamant une accentuation des programmes de lutte contre la pauvreté et l'instauration d'un programme de revenu minimum. Des économistes d'écoles de pensée très différentes étaient par ailleurs favorables à la mise en place d'un revenu minimum garanti par le biais d'un impôt négatif. Afin de lever les réticences du Congrès, qui craignait que les familles bénéficiaires de cette aide ne soient incitées à réduire leurs efforts de recherche d'emploi, un doctorant en économie du MIT. travaillant pour une agence de lutte contre la pauvreté de Washington, Heather Ross, proposa de mettre en place une expérimentation contrôlée sur échantillon aléatoire (Ross, 1966). Cette idée fut soutenue aussi bien par les Républicains que par les Démocrates.»

Source : Denis Fougeres, « Expérimenter pour évaluer les politiques d'aide à l'emploi : les exemples anglo-saxons et nord-européens », *Revue Française des Affaires Sociales* (1er trimestre 2000)[17].

Encadré 3 - L'expérimentation du RSA vue par Martin Hirsch en 2007

« **Depuis longtemps, vous défendez l'idée que l'expérimentation faciliterait la démarche de réforme. La voie que vous avez choisie pour le revenu de solidarité active (RSA), mis en place dans vingt-cinq départements avant d'être généralisé, reste rare. Pourquoi ?**

La France est en retard en matière d'expérimentation sociale, parce qu'elle est organisée pour que ce processus soit très peu possible. Juridiquement, d'abord, le principe d'égalité entre tous les citoyens ou tous les territoires est invoqué contre l'expérimentation. Culturellement, ensuite, l'idée d'avancer par tâtonnements est contradictoire avec la vision messianique d'un État qui sait où il faut aller. Enfin, la capacité d'expertise n'est pas à la hauteur. L'État est incapable de dire quels départements mènent les politiques sociales les plus efficaces, parce qu'il ne le mesure pas, et qu'il ne dispose pas des outils pour le faire.

Les réformes se font quand même...

Oui, mais jusqu'à récemment, elles ont consisté à rajouter des strates plutôt qu'à transformer le système. La prime pour l'emploi est l'exemple typique. La réforme a été faite vite, sans évaluation *ex-ante*. Et c'est après coup que l'on se rend compte qu'elle coûte cher et ne remplit pas ses objectifs. On fait de l'évaluation *ex-post*, qui s'apparente plus à du contrôle. Si l'on avait pris le temps de l'expérimentation, de l'évaluation et du réglage fin de la réforme, l'efficacité aurait été infiniment supérieure. Autre exemple : l'accompagnement des chômeurs. Là encore, les choses ont été faites à l'envers. L'Unedic a confié le travail à des prestataires privés pour voir s'ils pouvaient être plus performants que l'ANPE. Mais les conditions mêmes de l'expérimentation n'ont pas été arrêtées de manière consensuelle. Les résultats sont donc contestés, on cherche *a posteriori* à lever les biais pour avoir des éléments de comparaison fiables.

D'autres pays font-ils différemment ?

Les États-Unis ont une approche intéressante depuis plus de vingt ans. Ils testent systématiquement les réformes qu'ils veulent lancer en choisissant des groupes témoins et en comparant les résultats avec d'autres. Des États ont vérifié qu'il était efficace d'aider financièrement des jeunes en leur promettant *in fine* un emploi à condition que leurs résultats scolaires s'améliorent. Le maire de New York, Michael Bloomberg, a aussi testé des aides au retour à l'emploi ciblées sur les personnes ayant des enfants à charge, avec des résultats très positifs à la clef. Les pays les plus friands d'expérimentations sont ceux qui ont le moins la culture de la dépense publique : ils doivent justifier à l'avance le fait que le moindre dollar dépensé aura un retour sur investissement !

La France est-elle en train d'évoluer ?

Oui. Les départements bougent parce qu'ils ont un intérêt direct à ce que les politiques sociales produisent des résultats. Et la Constitution a été modifiée en 2003 de manière à leur permettre de déroger au principe d'égalité dans des conditions bien cadrées. Je pense que, dans les années qui viennent, l'expérimentation sociale va constituer l'approche la plus féconde pour renouveler les politiques publiques, à mesure que les outils réglementaires classiques montreront leurs limites et que la dimension du comportement des acteurs dans les processus de réforme prendra de l'importance. Les expérimentations permettent de mesurer cette variable essentielle. Autre avantage : en s'adressant à des volontaires, elles créent de l'émulation. On passe de la réforme subie à la réforme choisie, co-construite, comme en témoigne l'appétence croissante des départements à tester le RSA ».

Propos de Martin Hirsch, recueillis par Étienne Lefebvre et Dominique Seux, « L'expérimentation, une approche féconde pour renouveler les politiques publiques », *Les Échos*, 22/11/07. Martin Hirsch est alors Haut-commissaire aux solidarités actives contre la pauvreté.

2.3 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 figure 2.2 et illustré numériquement tableau 3.

Restons dans le cadre de l'évaluation d'un programme de formation professionnelle. Ce deuxième type d'expérience randomisée commence par laisser le choix aux individus de candidater au programme de formation, on parle d'**auto-sélection**. Deux groupes sont ainsi formés : les individus candidats au programme et les individus non recourants. L'impact du programme sur les individus candidats est bien ce qui nous intéresse 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, tableau 3, que le taux d'activité sans programme en $t = 1$ du groupe de traitement (25 %), qui représente le contrefactuel, est bien égal au taux d'activité du groupe de contrôle, contrefactuel utilisé. 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.

Figure 2.2 – Design 2 : Randomisation après auto-sélection, exemple d'un programme de formation professionnelle

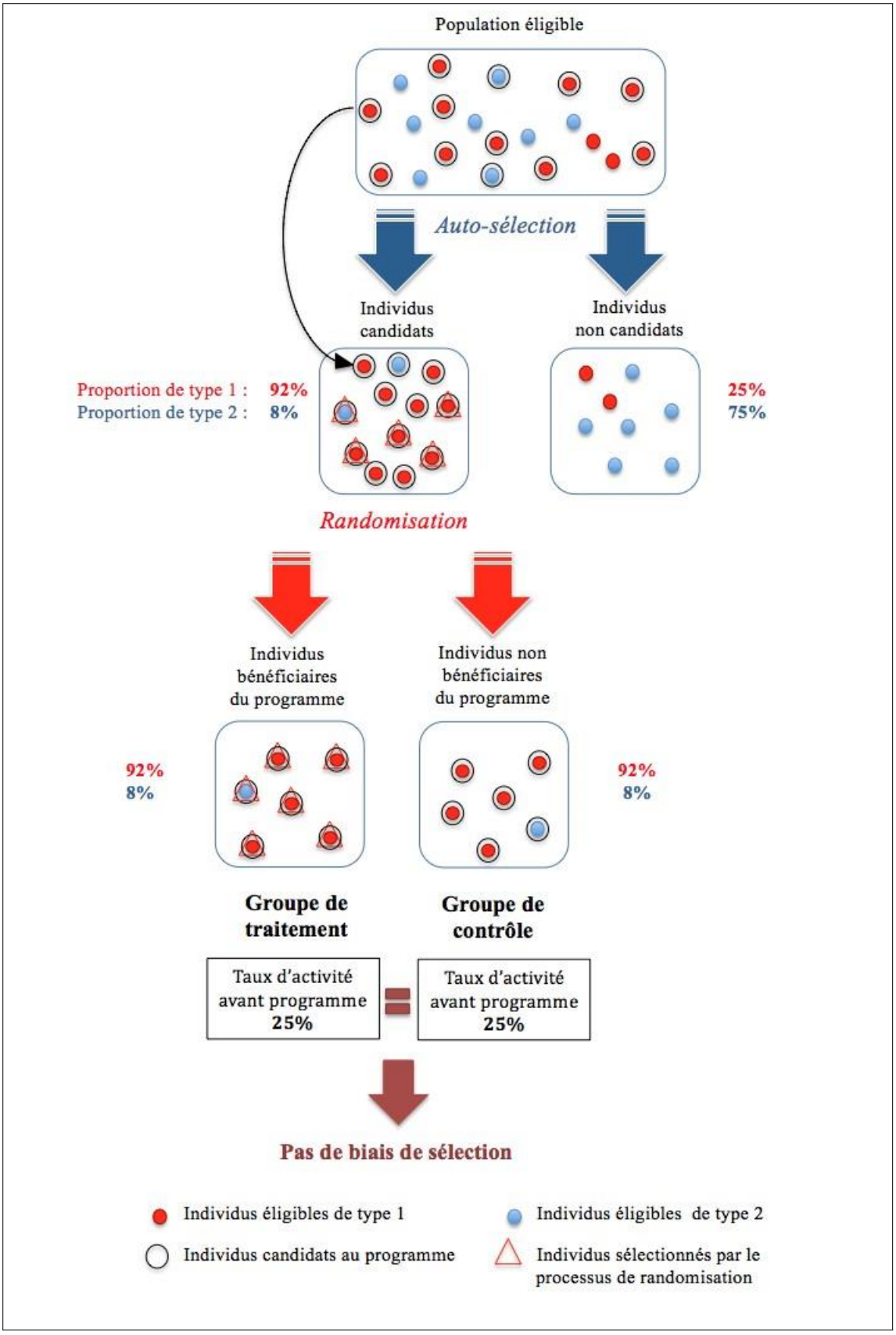


Tableau 3 - Design2 : Randomisation après auto-sélection : exemple numérique d'un programme de formation professionnelle

	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	-0,1

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

Design 2: RANDOMISATION APRES 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 (1)	0,53	0,80
Taux d'activité sans programme (t = 1)	0,25 (2)	0,25 (3)	0,65
Impact causal moyen réel sur les recourants (ICMR) (1) – (2)	0,28 (4)		
Impact causal moyen estimé sur les recourants (ICMR estimé) (1) – (3)	0,28 (5)		
Biais de sélection (5) – (4)	0		
Contrefactuel	Individus candidats non traités en t = 1		

= Non observable

2.3.1 Avantages et inconvénients du design 2

Ce type d'expérience randomisée replace la comparaison au niveau suivant l'auto sélection et permet ainsi de ne pas imposer de traitement à une population qui 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[4]. Le design 1 donnant l'effet moyen sur toute la population éligible ne conviendrait donc pas.

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 résout ce problème en proposant le traitement seulement à ceux qui pourront véritablement y avoir accès.

Application - Estimation de l'impact du programme de formation professionnelle *JTPA Title II-A* (Bloom *et al.* 1997)[14]

Le *Job Training Partnership Act (JTPA)* a été soumis à une évaluation randomisée du type du design 2 décrit ci-dessus. 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 à cette expérience peut être basée sur des facteurs corrélés avec l'effet du programme[12].

2.4 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 2.3 schématise ce troisième design et le tableau 4 reprend de nouveau l'exemple numérique du programme de formation professionnelle.

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 2.3.

Le groupe de contrôle et le groupe de traitement ont des caractéristiques identiques (même composition de types 60 % de type 1 et 40 % de types 2) et ne diffèrent que par l'accès au traitement.

[12] *Cf.* Section Menaces à la validité externe, paragraphe « Acceptation partielle et biais de randomisation ».

Ce design d'expérience randomisée permet d'obtenir deux impacts différents d'un programme :

— **1. 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, c'est-à-dire le groupe ayant accès au programme, et le groupe de contrôle. Dans le tableau ci-contre, 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 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.

— **2. 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[13]. 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 (52 %).

2.4.1 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 cependant, 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 du traitement à une partie de la population.

[13] On peut imaginer que des individus refusent de faire part d'une expérimentation mais qu'ils s'inscrivent, en réaction, à un autre programme de formation, l'impact causal du programme se trouverait sous-estimé.

Figure 2.3 – Design 3 : Randomisation de l'accès au traitement : exemple d'un programme de formation professionnelle

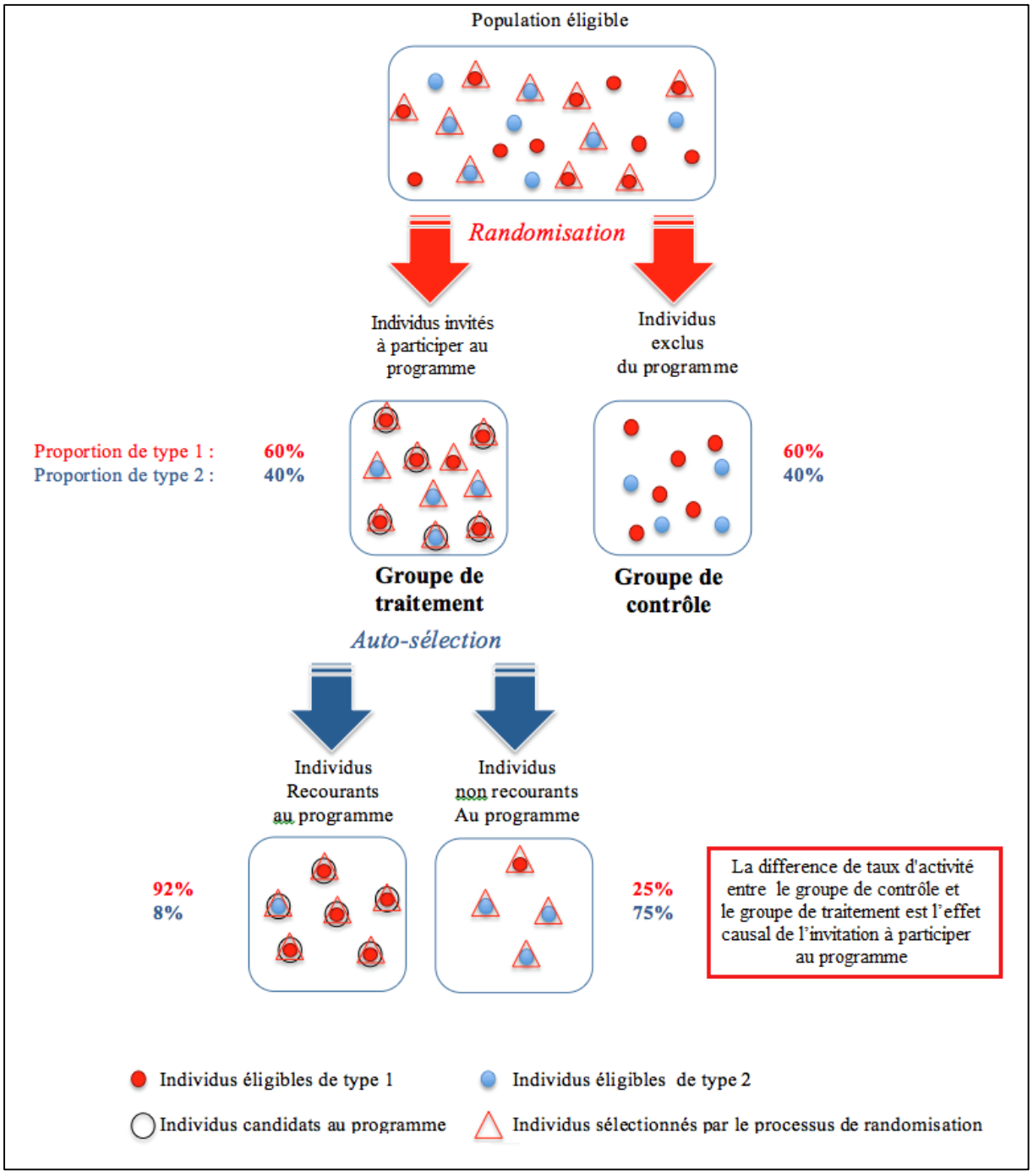


Tableau 4 – Design 3 : Randomisation de l'accès au traitement : exemple numérique d'un programme de formation

	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**

Design 3: RANDOMISATION DE L'ACCES AU PROGRAMME			
	Groupe de traitement *(accès libre)*		**Groupe de contrôle** *(accès bloqué)*
Proportion de Type 1 après randomisation (t = 0)	0,6		0,6
Proportion de Type 2 après randomisation (t = 0)	0,4		0,4
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 recourants	0,52 (3)		0
Taux d'activité moyen observé	0,59 (4)		0,44 (5)
Impact moyen de l'accès au traitement (ICMA) (4) –(5)	**0,15** (6)		
Impact causal moyen réel sur les recourants (ICMR réel)(1) – (2)		**0,28**	
Impact causal moyen estimé sur les recourants (ICMR estimé)(6) /(3)		**0,28**	
Contrefactuel	Individus candidats non traités en t = 1		

= Non observable

2.5 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 incitation à entrer dans le traitement évalué[26] : envoyer un courrier d'information sur une politique sociale, organiser des réunions d'information[58] ou encore donner une incitation financière ou matérielle[10].

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 2.4.

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 deux groupes comprennent les mêmes proportions d'individus de type 1 et d'individus de type 2 (60 % et 40 %) et ne différent 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.

Figure 2.4 – Design 4 : Randomisation d'un encouragement : exemple d'un programme de formation professionnelle

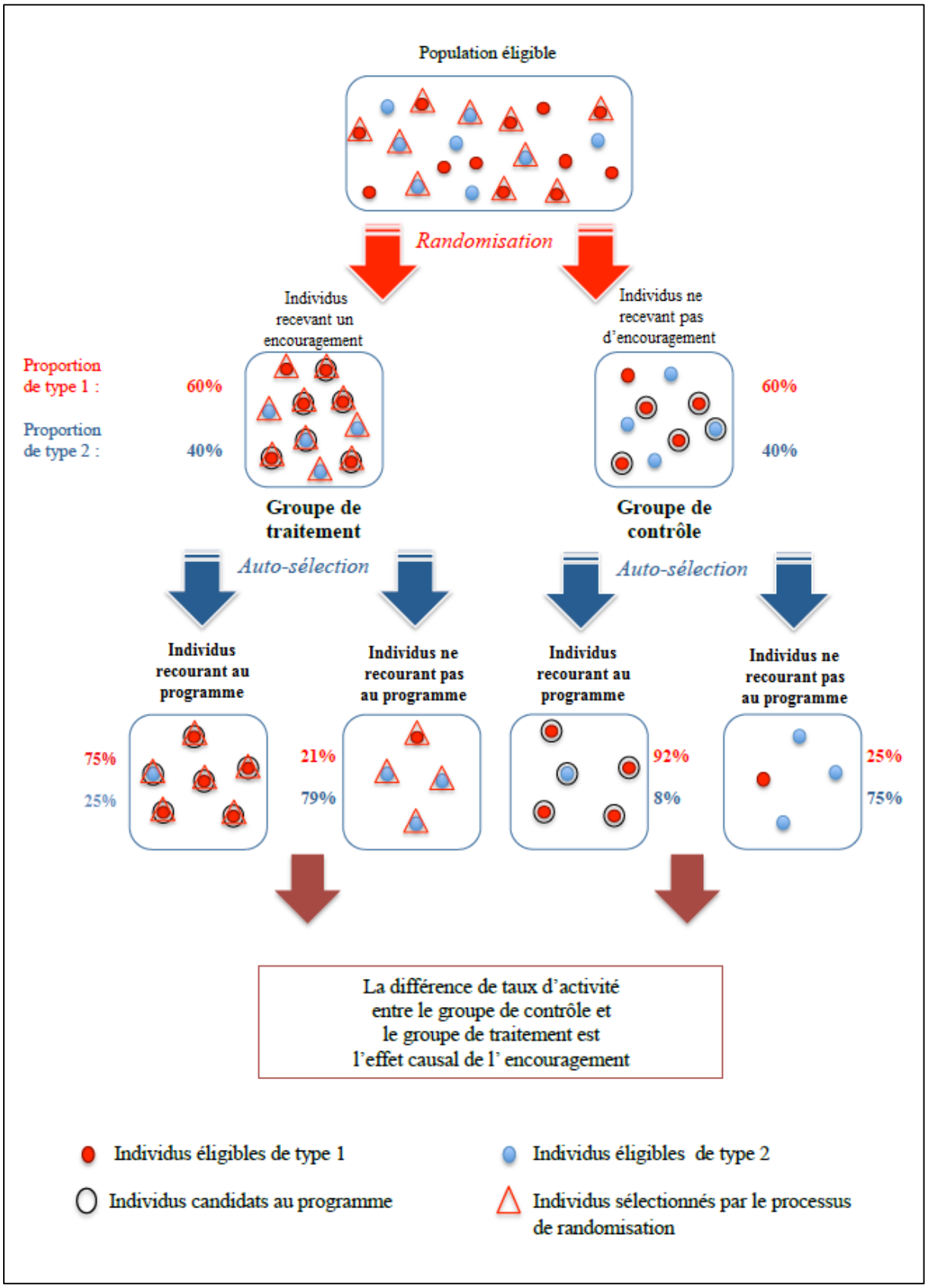


Tableau 5 - Design 4 : Randomisation d'un encouragement : exemple numérique d'un programme de formation professionnelle

	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 2 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)	0,6		0,6	
Proportion de Type 2 après randomisation (t = 0)	0,4		0,4	
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 (4)	0,81	0,53 (6)	0,80
Taux d'activité sans programme (t = 1)	0,35 (5)	0,67	0,25 (7)	0,65
Taux d'activité moyen	0,62 (1)		0,59 (2)	
Impact moyen de l'encouragement (1) – (2)	**0,03** (3)			
Impact moyen du programme sur les switchers (3)/proportion de switchers	**0,15**			
Impact causal moyen réel sur les recourants (ICMR réel)	**0,25** (4) - (5)		**0,28** (6) – (7)	

= Non observable

Ce design d'expérience randomisée permet d'obtenir deux impacts différents d'un programme :

— **1. 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 0,03 point de pourcentage.

— **2. L'impact causal du traitement sur les *switchers* (ICMS)**. Les *switchers* 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. Le groupe de contrôle et le groupe de traitement ne différant que par la réception d'un encouragement, l'écart de niveau du paramètre d'intérêt, *taux d'activité*, entre ces deux groupes est dû à la différence de taux de participation, c'est-à-dire à la proportion de *switchers*. On peut obtenir l'impact causal du traitement sur les *switchers* en divisant l'impact causal moyen de l'encouragement (ICME) par la proportion de *switchers*. L'estimation de ce paramètre suppose donc que l'on ait connaissance de la proportion de *switchers* dans la population.

  Dans notre exemple, participer au programme de formation professionnelle augmente de 15 points le taux d'activité moyen des *switchers*, c'est-à-dire l'ICME (0,03) divisé par la proportion de *switchers*[14]. Il a été choisi arbitrairement que les individus de type 2 comportaient davantage de *switchers* (35 %) que les individus de type 1 (10 %). En ajustant par les proportions de chaque type on obtient 20% de *switchers* dans la population. Dans la réalité ces chiffres ne sont pas directement observables mais la proportion de *switchers* se retrouve en calculant la différence de proportion de recourants entre le groupe de contrôle et le groupe de traitement, soit $0{,}72 - 0{,}52 = 0{,}20$.

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 v.s. 0,25). 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 recourants recevant l'encouragement comprend donc d'avantage de types 2 ce qui tire vers le bas l'impact du programme sur le taux d'activité.

2.5.1 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 *switchers*. La limite majeure de ce design est qu'il permet d'évaluer l'effet d'un programme seulement sur une population limitée représentée par les *switchers*. Cependant, ce design est parfaitement adapté aux politiques qui visent à augmenter le taux de participation à un programme donné et qui ciblent précisément les éventuels *switchers*.

Dans le cas du RSA, cette méthode est 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[4].

L'application 2, ci-après, présente une expérience réalisée en Gironde dont le but était d'évaluer l'effet d'une campagne d'information sur le taux de participation au RSA.

[14] La proportion de *switchers* est obtenue par le calcul : $0{,}1 * 0{,}6 + 0{,}35 * 0{,}4$.

Application 1 - Randomisation d'un suivi renforcé par l'ANPE (Crépon *et al*. 2012-2013)[11][12]

Behagel, Crépon et Gurgand ont écrit un article présentant les résultats d'une large expérience randomisée réalisée en France, 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 classique. En pratique, cette expérimentation s'apparente donc au design 3 : la population de switchers 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.

Application 2 - Randomisation d'un encouragement en Gironde[57]

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.

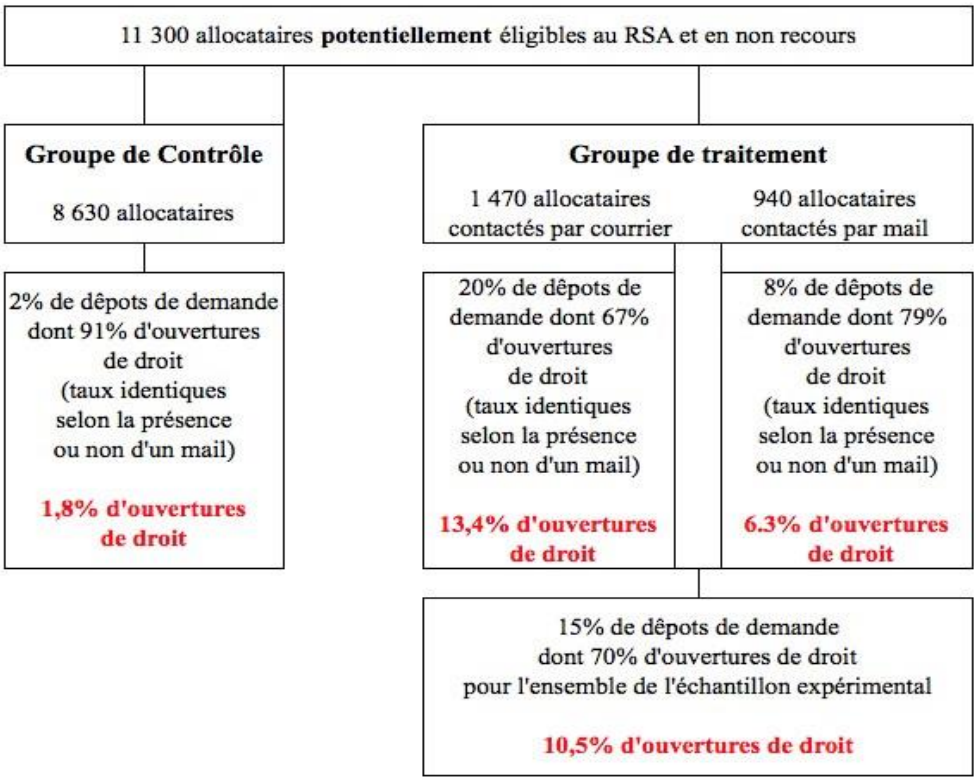
In fine trois groupes ont été formés : un groupe de traitement constitué d'individus recevant un encouragement, mail ou courrier, et un groupe de contrôle. Ces groupes ont été construits par tirage aléatoire. Plus précisément, les expérimentateurs ont procédé à un tirage aléatoire par rapport à plusieurs variables de stratification : situation familiale, estimation des montants des droits au RSA et points d'accueil physiques de rattachement[15].

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.

[15] Un échantillonnage stratifié sélectionne des échantillons de manière indépendante dans les strates spécifiées. Cette méthode d'échantillonnage permet d'améliorer la précision des estimations globales.



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

2.6 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 ci-après décrit plus en détail ce programme.

Une distorsion induite par ce design est que les individus sachant qu'ils auront droit au traitement dans une période future, 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[32].

Application – 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. Il touchait en 2012 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 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[16]. 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[34], 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 (pp) 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. PROGRESA semble également avoir des impacts bénéfiques sur la santé : la probabilité de tomber malade a été réduite de 12 pp pour les enfants et de 19 pp pour les adultes. Les ménages traités semblent aussi mieux s'alimenter.

16 . *Cf.* section Biais de diffusion paragraphe « Menace à la validité interne ».

Le tableau 6 résume les principales caractéristiques des différents designs évoqués.

Tableau 6 – Résumé des caractéristiques des différents designs d'expériences randomisées

Design de l'expérimentation	Paramètre(s) estimé(s)	Avantages	Inconvénients
1. Randomisation simple	Impact causal moyen du traitement sur toute la population	Pas de biais de sélection	Le paramètre estimé (ICM) ne correspond pas à l'impact causal d'intérêt (ICMR). Traitement imposé aux individus.
2. Randomisation après auto-sélection	Impact causal moyen du traitement sur les recourants (ICMR)	Pas de biais de sélection. Estimation de l'impact causal d'intérêt. Traitement non imposé aux individus	Traitement non accessible à toute la population évaluée.
3. Randomisation de l'accès au traitement	Impact causal de l'accès au traitement (ICMA) et ICMR	Pas de biais de sélection. Traitement non imposé aux individus. Estimation de l'impact causal d'intérêt.	Traitement non accessible à toute la population évaluée.
4. Randomisation d'un encouragement	Impact causal moyen de l'encouragement (ICME) et Impact causal moyen du traitement sur les *switchers* (ICMS)	Pas de biais de sélection. Accès au traitement non restreint.	Évaluation de l'impact du traitement seulement sur une sous-population : les *switchers*.
5. Randomisation de l'ordre d'allocation du traitement	ICMR	Pas de biais de sélection. Accès au traitement non restreint. Estimation de l'impact causal d'intérêt.	Biais d'anticipation. Évaluation des effets de long terme problématique.

2.7 Problèmes et biais potentiels des expériences randomisées

Cette partie résume tout d'abord les problèmes éthiques soulevés par les différents designs 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 aspects 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.

Exemple : *Si l'on estime qu'un programme de formation fait augmenter le taux d'activité de 30 points de pourcentage, dans quelle mesure cette valeur correspond-elle bien à l'impact causal réel du programme sur l'échantillon de population étudiée?*

— **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, autrement dit, si l'estimation est représentative de l'effet de la politique que l'on souhaite réellement mettre en œuvre.

Exemple *: Si l'on met en place un programme de formation en l'absence de randomisation ou dans un autre contexte, son effet sur le taux d'activité sera-t-il véritablement de 30 points de pourcentage supplémentaires ?*

La validité interne et la validité externe peuvent être affectées par certains phénomènes présentés dans les parties 2 et 3 de cette section. Par ailleurs, un autre inconvénient souvent accordé à l'expérimentation est son coût [*Cf.* encadré 4].

2.7.1 Problèmes éthiques et politiques

Comme nous l'avons suggéré précédemment, l'expérimentation randomisée est sujette à différents problèmes éthiques. Il est en effet difficile de donner l'accès à une politique potentiellement bénéfique à une partie restreinte de la population ou encore d'obliger des individus à participer à un programme social.

Le design 1 (randomisation du traitement) force des individus à entrer dans un programme. Le design 2 (randomisation après auto-sélection) informe des individus qu'ils ont accès à un programme puis, une fois qu'ils ont accepté, leur bloque finalement l'accès. Le design 3 (randomisation de l'accès au traitement), quant à lui, n'offre le traitement qu'à une partie de la population.

Les problèmes éthiques sous-jacents à l'allocation aléatoire de l'expérimentation randomisée sont récurrents dans les débats politiques et méthodologiques. On comprend bien que ne pas allouer un traitement à l'ensemble d'une population pose un problème d'équité. Cependant, une nouvelle politique fait souvent place à une ancienne politique et son évaluation se fait par rapport à une situation de référence qui est rarement une situation sans politique. On ne prive généralement pas les individus de toute politique, mais simplement de la version qui fait l'objet de l'évaluation. De plus, sans évaluation, il n'est pas possible de dire si une politique est plus efficace qu'une autre et il est donc impossible de dire si un groupe est désavantagé par rapport à un autre.

Si l'on résume les avantages et les inconvénients de chaque design, on s'aperçoit que l'on est face à un arbitrage *équité / apprentissage*. Les designs les plus contraignants éthiquement sont les designs 2 et 3 (randomisation après auto-sélection et randomisation de l'accès au traitement). Ils sont pourtant très informatifs sur l'efficacité du programme parce qu'ils permettent l'estimation de l'impact causal sur les recourants. *A contrario* le design 4 de randomisation d'un encouragement qui permet de résoudre les problèmes éthiques ne mesure l'impact causal de la politique que sur les individus affectés par l'encouragement (les switchers). La randomisation progressive, lorsqu'elle est appliquée avec certaines précautions, représente, quant à elle, un bon compromis entre l'apprentissage nécessaire à propos d'une politique et les questions d'équité lorsque que l'évaluation peut se faire sur l'ensemble de la population.

Pour finir, lorsque des contraintes budgétaires ne permettent pas de faire bénéficier de la politique tous ceux qui le souhaiteraient, une allocation aléatoire des individus à un groupe peut représenter la manière la plus juste d'assigner un programme, car ce type d'allocation ne se base sur aucun critère et n'est donc soumis à aucune discrimination. Par ailleurs les expériences randomisées sont fréquemment pratiquées, et depuis longtemps, dans le domaine biomédical qui, lui, soulève ces questions éthiques d'autant plus qu'il impacte la santé des populations étudiées[15].

Encadré 4 - Identifier les mesures les plus rentables pour la collectivité

« Dans un bulletin du J-PAL (*Laboratoire d'Action contre la Pauvreté Abdul Latif Jameel*), on compare le coût par enfant, d'une année d'éducation supplémentaire induite par les différentes stratégies mises en œuvre. Les coûts varient de +$3,50 pour une année supplémentaire pour le traitement vermifuge, à 6000$ par année supplémentaire par enfant pour le volet Éducation Primaire du programme PROGRESA, programme mexicain de transferts sociaux conditionnels. Même si l'on exclut PROGRESA, dont l'éducation n'est pas le seul objectif, le coût par année d'éducation supplémentaire va de $3,25 à plus de $200 d'un programme à l'autre. Le taux de rentabilité des investissements publics est loin d'être égal ».

Banerjee Abhijit V. et Duflo E. (2009), « L'approche expérimentale en économie du développement », *Revue d'économie politique*, Vol. 119, p. 691-726.

2.7.2 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. Ceci crée un manque de données et en particulier sur les résultats finaux qui sont essentiels à la conclusion de l'expérience. Cela constitue un problème majeur car l'attrition n'est pas nécessairement aléatoire c'est-à-dire que les individus quittant l'expérience n'ont pas des caractéristiques homogènes ce qui fausse la randomisation initiale et donc ne permet pas d'obtenir une estimation correcte de l'effet du traitement.

Solution : l'attrition représente la menace la plus importante de l'expérimentation randomisée. Les solutions disponibles concernent 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 en estimant les bornes minimale et maximale de l'impact d'un programme[46][49].

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 améliore donc la santé des individus du groupe de contrôle de façon indirecte qui, autrement aurait potentiellement été contaminés par les individus du groupe de traitement[54].*
>
> ***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. On parle d'effet d'équilibre général[27].*

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 relative au paramètre d'intérêt, deux marchés distincts par exemple.

> ***Exemple 1 :*** *Lors de l'évaluation du RSA une randomisation au niveau des villes peut éviter les biais de diffusion. Il est en effet peu probable qu'un individu touchant le RSA dans la ville de Lyon impacte le niveau d'activité d'un habitant de Toulouse.*
>
> ***Exemple 2 :*** *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[54].*
>
> ***Exemple 3 :*** *Une étude publiée en 2012[27] 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.*
>
> ***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 ont moins d'incitation à rechercher un substitut.

Effet de l'expérimentateur

Il s'agit de l'effet produit par les expérimentateurs sur les individus étudiés. Une expérience randomisée passe par la rencontre avec les sujets, par la réalisation de différents questionnaires. Tous les contacts des expérimentateurs avec les sujets semblent affecter leurs comportements. Rothenthal[60] explique 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.

Solution : Nécessité pour les expérimentateurs d'être le plus neutre possible.

Durée limitée d'une expérience

Afin de pouvoir estimer correctement un impact causal avec une expérience randomisée, il est nécessaire que l'expérimentation dure suffisamment longtemps. Tout d'abord, une période de temps longue est nécessaire afin que les effets aient le temps de se matérialiser. Par exemple, une politique de subvention sur l'achat d'immobilier ne peut pas être évaluée sur six mois. 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 donc de faire durer l'expérimentation aussi longtemps que les contraintes financières et politiques le permettent.

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[30]. Il s'agit d'un effet psychologique qui peut favoriser l'estime de soi. Le plus souvent cela se traduit par une plus grande motivation qui peut aussi provenir du simple fait d'être observé ou d'une certaine reconnaissance[32].

L'effet estimé sera donc l'impact causal réel de la politique plus l'effet expérimental, ce qui altère donc l'évaluation de la politique.

On parle d'effet de John Henry lorsque que les individus du groupe de contrôle sont touchés par une distorsion de comportement. Les individus du groupe de contrôle peuvent se trouver offensés de se voir refuser la politique ce qui peut affecter leurs décisions regardant les paramètres d'intérêt. 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^{ième} 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 se dissipent dans le temps. Une autre solution peut également se trouver dans le design même de l'expérience. Par exemple Ashraf, Karlan et Yin[9] 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.

2.7.3 Menaces à la validité externe

Indépendance environnementale[32]

Il s'agit de la plus grande menace à la validité externe de l'expérimentation randomisée. 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 (À quel point l'expérimentation d'une politique spécifique est-elle informative au regard d'une politique légèrement différente ?). Le design d'une expérience est souvent très étudié et exécuté avec beaucoup d'application alors que dans un mode routinier, la mise en place d'une politique publique peut être moins élaborée. 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.

Solutions[32] **:** Il est nécessaire d'expérimenter des programmes réalistes qui peuvent être applicables. 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 consiste notamment à modéliser mathématiquement les comportements individuels, est un moyen d'analyser les mécanismes sous-jacents d'une intervention économique.

Acceptation partielle et biais de randomisation

Le problème d'acceptation partielle est le refus de la part des individus assignés au groupe de traitement de recevoir le traitement. Ce refus n'est pas aléatoire, les individus acceptant de participer à une expérience randomisée ont probablement des caractéristiques différentes de ceux qui refusent. Il s'agit d'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[43]. 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, certains économistes ont fourni des preuves montrant que les individus refusent davantage de participer à une expérience randomisée qu'à une expérience non randomisée. À titre d'exemple Kramer et Shapiro[56] 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.

Solution : L'acceptation partielle émerge lorsque que 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 une expérience randomisée. Procéder à une expérience progressive (design 5) peut permettre d'atténuer ce biais.

Encadré 5 - Utiliser l'expérimentation en France : Interview d'Esther Duflo

Des décennies d'économie du développement n'ont pas réussi à éradiquer la pauvreté, d'où un certain scepticisme. Quelle est la spécificité de votre démarche?

Le but d'éradiquer la pauvreté est sans doute un peu trop grandiose. Il y a une forte demande politique pour des baguettes magiques qui permettraient de résoudre le problème d'un seul coup, mais il est un peu vain de les chercher. Mon approche consiste à poser au contraire des questions très concrètes, auxquelles il est possible de donner des réponses claires. Non pas : « Comment éradiquer la pauvreté ? » mais : « Donner aux enseignants de meilleures incitations les convaincrait-il de venir travailler plus souvent ? Est-ce que cela aurait des conséquences positives sur les résultats scolaires des enfants ? » Une fois que l'on se pose ce genre de questions beaucoup plus concrètes, on peut y répondre de manière plus rigoureuse. D'où ma deuxième innovation, l'usage d'expérimentations randomisées (similaires aux essais cliniques) pour l'évaluation des politiques publiques.

Quels sont les résultats les plus frappants des expérimentations que vous avez conduites ou inspirées?

Il est difficile d'isoler un résultat ou l'autre. J-PAL, un réseau de chercheurs que j'ai fondé avec Abhijit Banerjee et qui est consacré à l'utilisation de la méthode expérimentale pour l'évaluation des programmes de lutte contre la pauvreté, a réalisé presque trois cents évaluations, en comptant celles encore en cours. Chacune d'entre elles, prise isolément, est plus ou moins intéressante. Or c'est quand on combine tous les résultats de diverses expérimentations réalisées dans un même domaine, mais dans des contextes différents et avec des programmes légèrement changeants, que l'on commence vraiment à en voir la force et que l'on peut en tirer des leçons plus générales. C'est ce que nous essayons de faire dans le livre Poor Economies - Repenser la pauvreté (Banerjee et Duflo, 2011-2012). Une leçon fondamentale, c'est que la source des échecs de certains programmes tient souvent au défaut de réflexion et de soin apporté à l'élaboration du programme : il faut faire extrêmement attention aux détails.

Comment passe-t-on du résultat d'une expérimentation aléatoire à des préconisations concrètes de politique publique?

Cela dépend de l'expérimentation. Si celle-ci a été réalisée dans des conditions «réalistes» pour le compte d'un gouvernement qui a la possibilité de généraliser un programme identique, alors le résultat peut se traduire presque immédiatement par une recommandation de politique publique. Par exemple, nous avons réalisé avec la police du Rajasthan une expérience à très grande échelle portant sur différentes interventions destinées à améliorer la performance et l'image de la police. L'une de celles qui se sont révélées efficaces est une session de formation aux techniques d'enquête et à la médiation. C'est quelque chose que le Rajasthan est maintenant à même de généraliser, et peut-être d'autres États indiens le pourront-ils également. Dans d'autres cas, l'expérimentation sert davantage à comprendre un mécanisme. Par exemple, donner de petites incitations aux gens permet d'augmenter énormément les taux de vaccination. Mais, pour un gouvernement, il y aurait tout un nombre d'étapes à franchir pour être capable de mettre un programme général en place tout en évitant la corruption. En l'occurrence, l'expérimentation a davantage débouché sur un principe que sur une intervention clés en main.

L'expérimentation aléatoire peut-elle être appliquée avec autant de succès dans les pays développés que dans les pays en développement?

Oui. L'expérimentation aléatoire vient des pays développés : les premières grandes expérimentations sociales ont été réalisées aux États-Unis. On observe d'ailleurs un fort mouvement dans ce sens en France depuis quelques années, sous l'impulsion de chercheurs comme Marc Gurgand et Bruno Crépon, qui sont réunis dans le laboratoire J-PAL Europe. Une bonne vingtaine d'expérimentations sociales sont en cours à l'heure actuelle en France, sur des sujets aussi variés que l'éducation, la création d'entreprise, le rôle des permis de conduire, l'autonomie des personnes âgées, la santé des jeunes, etc.

Quel type de structures faudrait-il mettre en place en France pour systématiser l'évaluation des politiques publiques et améliorer l'action de l'État ?

Le fonds d'expérimentation sur la jeunesse qu'a lancé Martin Hirsch lorsqu'il était commissaire à la jeunesse et aux Solidarités actives est un modèle du genre. Il s'agissait d'un fonds, bien ouvert, qui lance des appels d'offres régulièrement, auxquels peuvent répondre des membres de la société civile, des administrations locales, des lycées. Ils peuvent soumettre leur idée pour une nouvelle politique et l'accompagner d'une évaluation la plus rigoureuse possible, souvent réalisée en collaboration avec un groupe de chercheurs. Un comité scientifique évalue les propositions à l'aune des critères suivants : est-ce que les leçons que nous pourrons en tirer seront utiles, Généralisables ? Le fonds a survécu au départ de Martin Hirsch, mais la grande question est de savoir s'il sera renouvelé ou étendu : ce serait une excellente chose, car ce fonds constitue un exemple pour le monde. La France est, pour une fois, très nettement en avance par rapport à ses homologues européens.

Propos recueillis par Claire Montialoux (*Regards croisés sur l'économie* n°10, Février 2011) ; Le livre évoqué est celui de Banerjee A. et Duflo E. (2011), *Poor Economies: a Radical Rethinking other Way to Fight Global Poverty*, Public Affairs, New York; trad. fr., *Repenser la pauvreté*, Seuil, Paris, 2012.

Chapitre 3 - 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. Ce chapitre présente des démarches cognitives *ex-post* alternatives à l'expérimentation randomisée, qui proposent des solutions au problème fondamental d'inférence causale, problème d'inobservabilité, et qui permettent la comparaison de ces différentes situations.

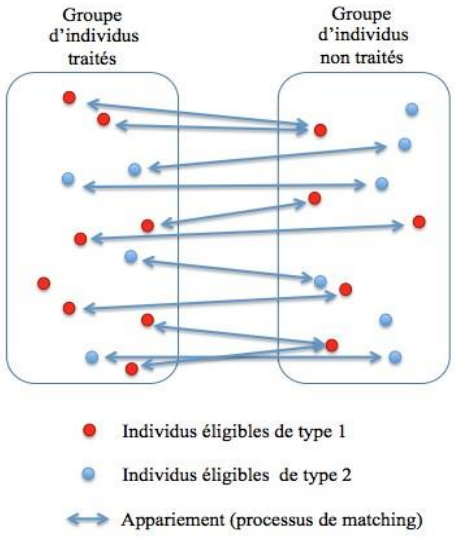
Ces méthodes, appelées **quasi-expérimentales**, permettent d'évaluer une politique ou un programme qui est alloué de façon non aléatoire au sein d'une population. On parle également d'*expériences naturelles*. Ces méthodes exploitent en effet des situations réalisées sans intervention expérimentale - naturellement - mais qui possèdent certaines caractéristiques permettant de les analyser comme si l'allocation du traitement était aléatoire, c'est-à-dire qui permettent la prise en compte des différents biais induits par une allocation non aléatoire.

3.1 Méthode de *Matching*

Définition[63] **:** 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[20]. 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 3.1 illustre de façon simplifiée ce mécanisme[17].

Figure 3.1 – Principe du *matching*



[17] Selon le type de procédure de *matching* utilisée, différents individus traités peuvent être appareillés à un même individu non traité ayant des caractéristiques similaires. Pour cette raison on observe figure 4.1 plusieurs flèches allant vers un même individu non-traité.

Cette méthode repose sur plusieurs hypothèses[20] :

— **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.

Autrement dit, si l'on prend l'exemple d'un 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.

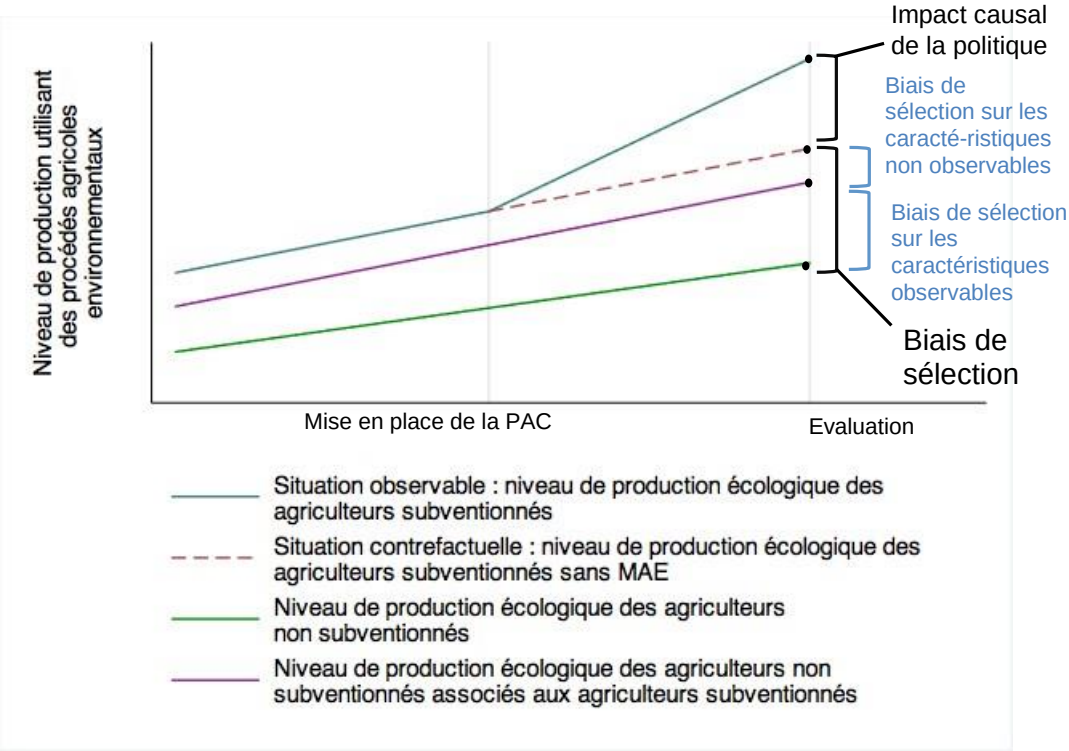
— **Hypothèse 2 :** Pour chaque caractéristique, il existe à la fois des individus traités et des individus non traités.

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. Mais 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. On voit bien que ne pas prendre en compte les caractéristiques non observables peut produire des résultats erronés.

La figure 3.2 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 section 1.3. Pour rappel, ces mesures consistent à verser des subventions aux agriculteurs qui utilisent des technologies de production écologiques.

Figure 3.2 – Décomposition du biais de sélection



— **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. L'hypothèse 2 ajoute indirectement que cette source de variation peut être traitée 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 dans ce document.

Application : Estimation de l'impact des mesures agro-environnementales de la PAC (Chabé-Ferret et Subervie)[23]

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 mesures agricoles environnementales (diversités des cultures, agriculture biologique etc.). La figure 3.2 illustre le problème de biais de sélection qui peut intervenir dans l'évaluation des MAE. En effet, comme nous l'avons mentionné, il est très probable qu'un nombre substantiel d'agriculteurs recevant les subventions utilisait déjà ces méthodes 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 méthodes 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 violette figure 3.2 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 présentons dans la partie qui suit. Ils trouvent finalement un impact causal de 10,66 ha.

3.2 Double différence

Nous avons vu que la méthode de *matching* ne prend pas en compte l'hétérogénéité des caractéristiques non observables des individus. La double différence est une méthode d'évaluation qui permet de résoudre, sous certaines hypothèses, le problème de sélection relatif aux caractéristiques observables ainsi qu'aux caractéristiques non observables.

L'idée de base de la double différence est d'estimer l'impact causal d'une politique en calculant la différence avant / après du paramètre d'intérêt pour un groupe traité et par ailleurs pour un groupe non traité et de comparer ensuite ces deux différences[21]. Autrement dit, cette méthode soustrait à la comparaison avant / après, la variation du paramètre d'intérêt du groupe non traité, qui est supposé être affecté de la même façon par la conjoncture socioéconomique que le groupe traité. Cette méthode peut aussi être comprise comme mesurant la variation de la *comparaison avec / sans avant et après la mise en place de la politique*. Sous l'hypothèse testable de **biais de sélection constant dans le temps** présentée ci-après, la double différence mène à une estimation non biaisée de l'impact causal d'un traitement.

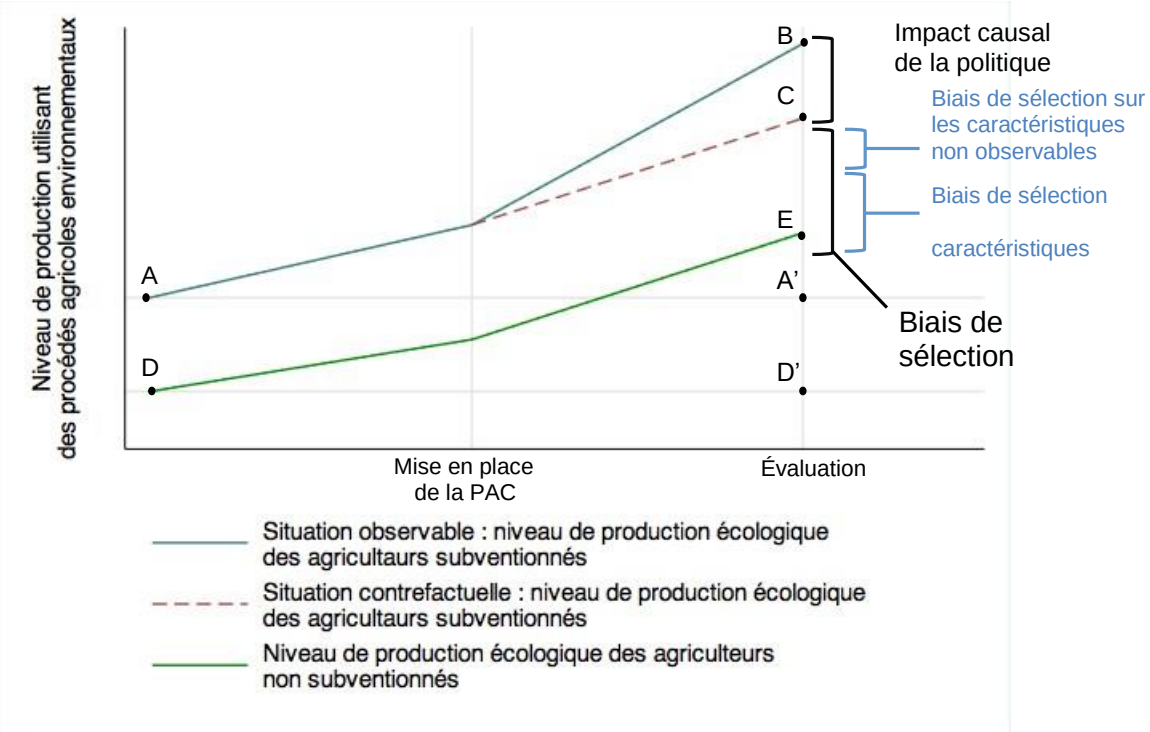
Nous présentons maintenant cette méthode de façon plus précise à travers l'exemple de la politique de la PAC. Nous expliquons ici la double différence graphiquement à l'aide de la figure 3.3.

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 3.3. 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. Figure 3.3, cela se traduit par un écart constant entre les droites (*AC*) et (*DE*), qui sont donc parallèles.

Figure 3.3 – Décomposition du biais de sélection



L'impact causal sur les recourants est estimé par la différence entre la comparaison avant / après du groupe traité (segment [$A'B$] ou ($B - A$) fig. 3.3) 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 (segment [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 bénéficiaire 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'].

Application 1 : L'impact d'une hausse du salaire minimum sur le taux d'emploi et les prix (Card et Krueger 1993)[16]

Comment les entreprises réagissent-elles à une hausse du salaire minimum? Il s'agit d'une question socioéconomique 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. 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.

Application 2 : L'impact de la construction de nouvelles écoles sur le niveau d'éducation et le revenu (Duflo 2001)[31]

Esther 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.

Esther 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 : Esther 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, Esther 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'à une augmentation de revenu de 1,5 % à 2,7 %.

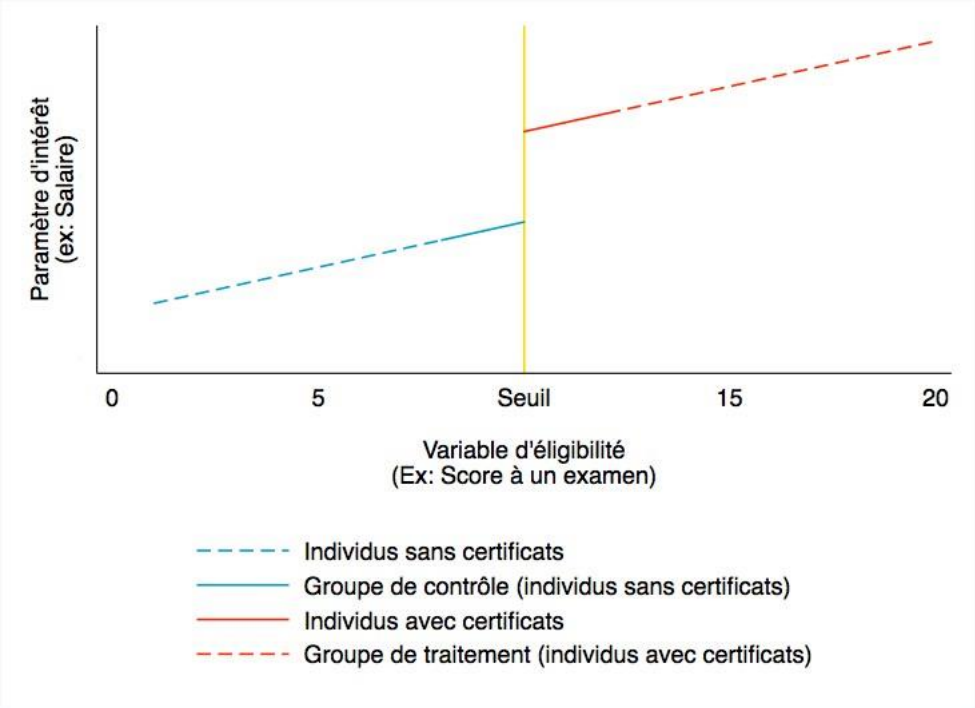
3.3 Régression par discontinuité

On a vu que le *matching* résout les problèmes de sélection pour les caractéristiques observables et que la double différence les résout pour les caractéristiques inobservables sous certaines hypothèses.

La méthode présentée ci-après est une méthode non expérimentale qui ne fait pas d'hypothèse sur le biais de sélection et qui parvient à prendre en compte l'hétérogénéité des caractéristiques non observables des individus étudiés, résolvant ainsi le problème de biais de sélection.

Cette méthode consiste à identifier une situation où l'allocation à un traitement 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.

Figure 3.4 – La régression par discontinuité



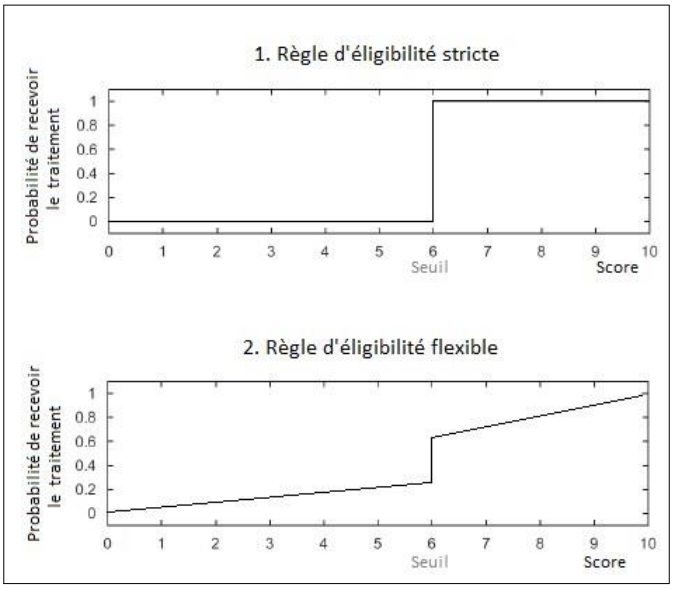
Définition : Cette méthode utilise le fait que lorsqu'un programme est distribué par rapport à un seuil d'éligibilité, une discontinuité est créée dans l'allocation du programme (voir premier encadré ci-après). Les individus juste en-dessous du seuil et ceux juste au-dessus ont des caractéristiques similaires 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 près de ce seuil [*cf.* figure 3.4].

Un étudiant arrivant 121ème à un concours où seulement 120 personnes sont admises a le même niveau de connaissance que 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[62] ont par exemple 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 soit 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[50].

Remarque : 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[22]. La figure 3.5 illustre cette différence. Un concours d'entrée pour une formation 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.

Figure 3.5 – Règle d'éligibilité flexible et règle d'éligibilité stricte

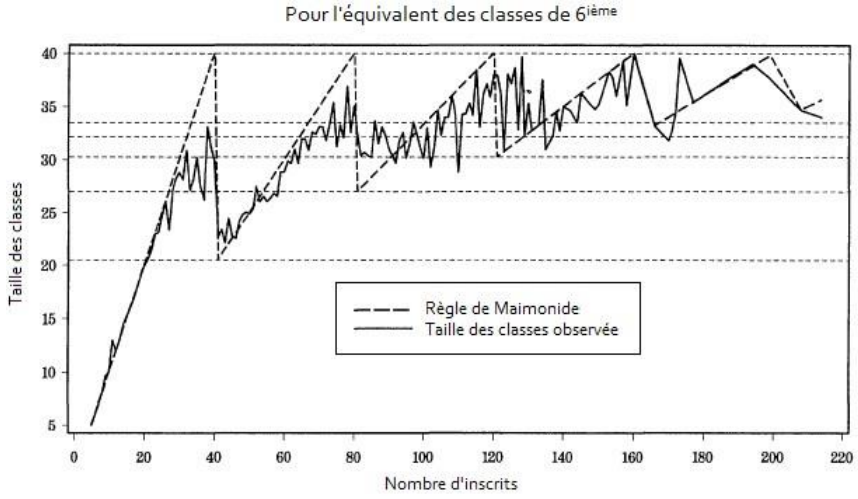


Application 1 : L'impact de la taille des classes sur les résultats scolaires (Angrist and Lavy 1999)[7]

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 Israel pour déterminer la taillees 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 3.6).

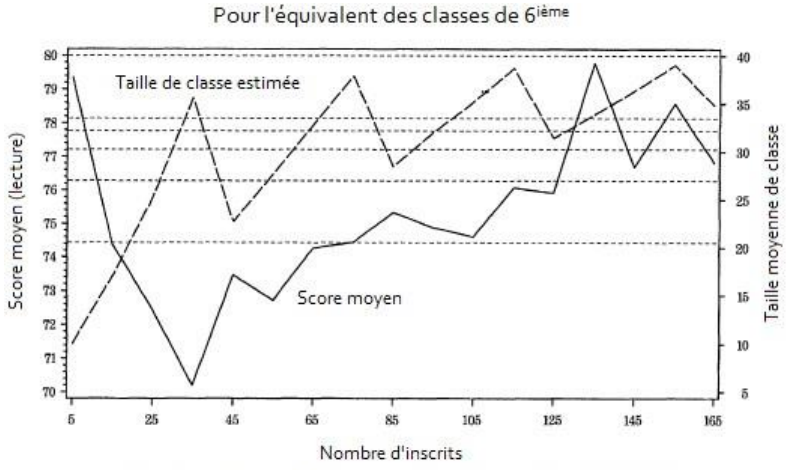
Figure 3.6 – Relation entre le nombre d'inscrits et la taille des classes



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

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.

Figure 3.7 – Relation entre résultats scolaires et taille des classes

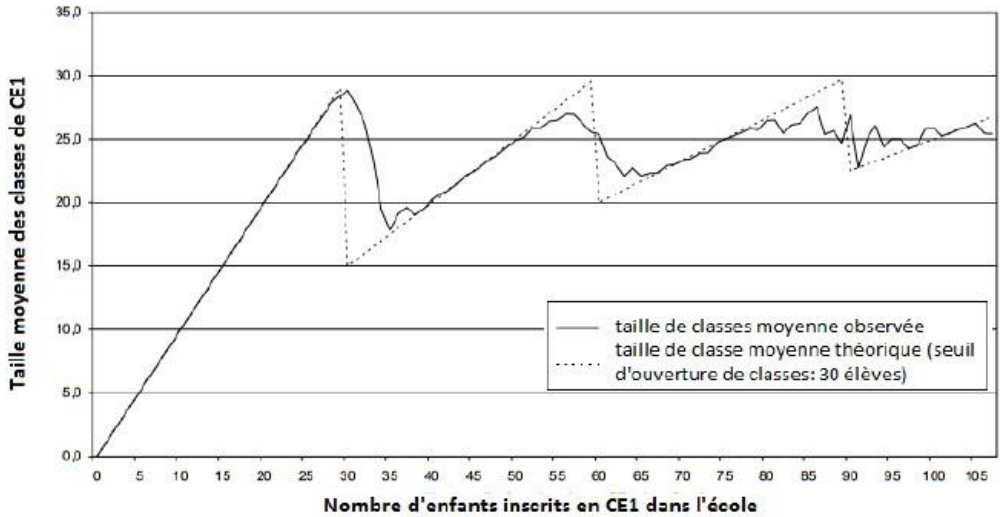


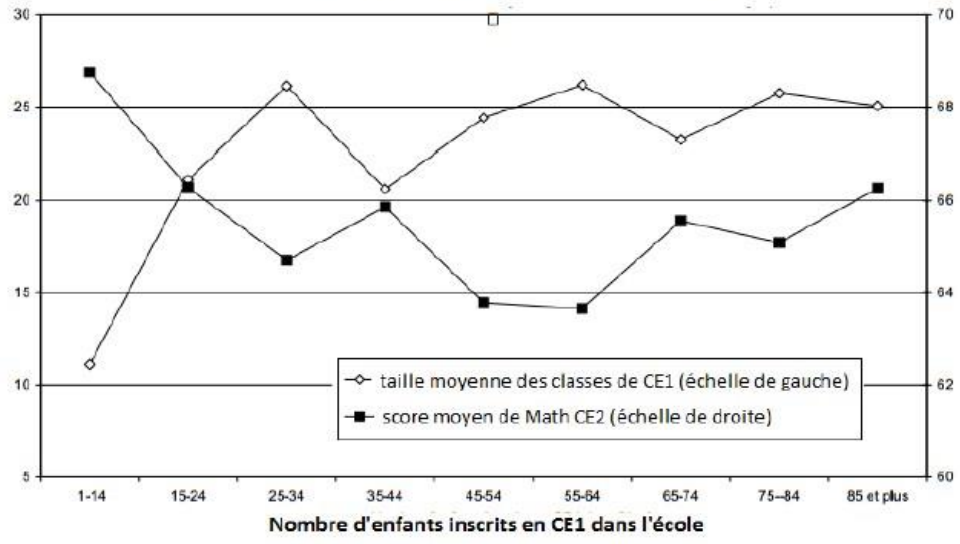
Source: The Quartely Journal of Economics, Vol. 114, No. 2 (May, 1999), 533-575.

En exploitant la discontinuité entre taille des classes et nombre d'élèves inscrits, Angrist et Lavy trouvent que réduire la taille des classes améliore la réussite des élèves.

Application 2 : L'impact de la taille des classes sur les résultats scolaires en France (Piketty et Valdenaire 2006)[39]

Une étude sur l'impact de la taille des classes sur la réussite scolaire a également été réalisée en France par Piketty et Valdenaire en 2006. 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.





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).

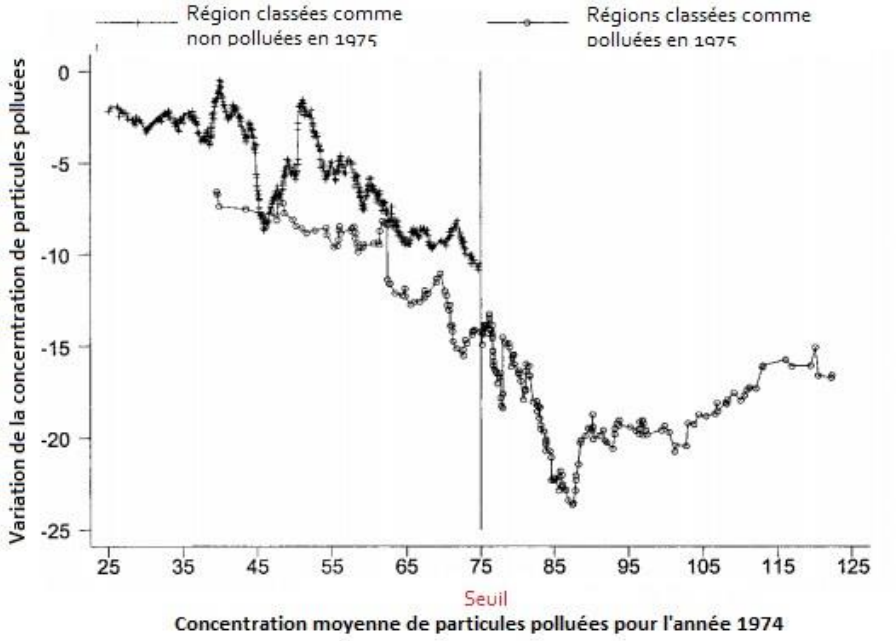
Piketty et Valdenaire discutent également de la politique des ZEP (zones d'éducation prioritaire), dont les classes ont en moyenne 20,9 élèves contre 22,8 dans les écoles non ZEP. Leurs résultats suggèrent qu'une diminution plus grande de l'effectif de ces classes diminuerait substantiellement les inégalités de réussite scolaire. Concernant l'éducation secondaire, l'impact de la taille des classes est moins important mais va dans la même direction.

Application 3 : Estimation de la valeur économique que les individus attribuent à l'air non pollué et impact des politiques de régulation (Chay et Greenstone 2005)[24]

Les politiques de régulation concernant la pollution sont aujourd'hui controversées et le fait que l'on puisse difficilement estimer les coûts et les bénéfices de la pollution contribue largement à ces débats.

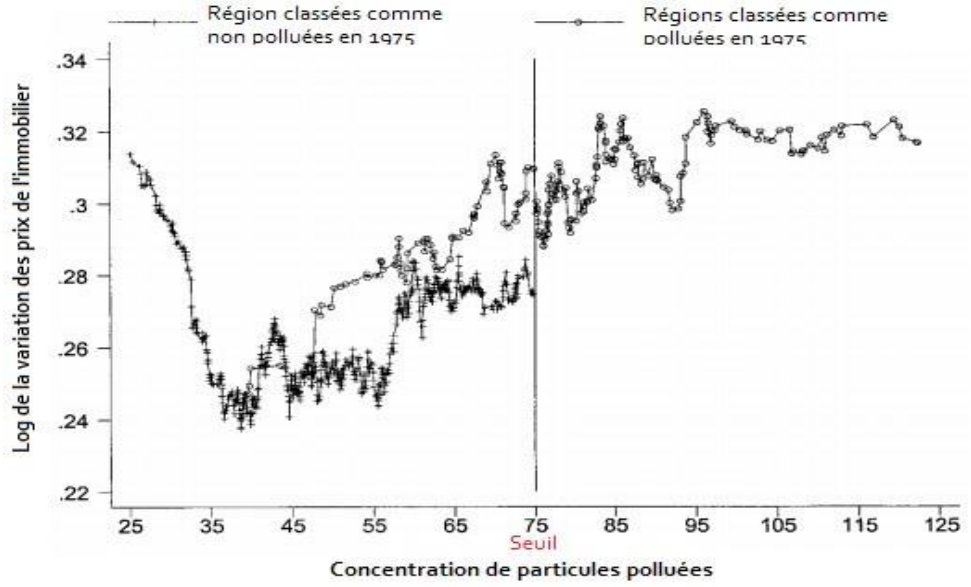
Chay et Greenstone exploitent la structure des *Clean Air Act Amendements* (CAAAs) de 1970 afin d'estimer la valorisation de l'air non pollué par la méthode de régression par discontinuité. Les CAAAs, mis en œuvre au niveau régional, visent à diminuer le niveau de pollution. Si le niveau de pollution d'une région dépasse un certain 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.

Figure 3.8 – Variation de la pollution entre 1970 et 1980 selon le statut et le niveau de pollution



Sur ce graphique, les régions ne dépassant pas le seuil annuel de pollution de 1μ/m^3 sont considérées comme polluées si elles dépassent un certain seuil journalier de pollution. On observe pour cette raison des régions désignée comme polluées à gauche de la droite verticale matérialisant le seuil annuel.

Figure 3.9 – Variation des prix de l'immobilier entre 1970 et 1980 selon le statut et le niveau de pollution



La figure 3.8 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 3.9 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 $1\mu/m^3$ de la concentration de particules polluées induit une augmentation de la valeur immobilière allant de 0,2 % à 0,35 %.

La politique CAAAs a donc permis une baisse du niveau de pollution qui a été reflétée par une hausse des prix de l'immobilier.

3.4 Variable instrumentale

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, *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 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 *switchers*, 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 un instrument valide si elle n'est pas directement corrélée à la variable d'intérêt mais qu'elle l'affecte uniquement par l'intermédiaire du taux de participation :

Envoyer un courrier informatif à propos d'un programme de formation professionnelle augmente le nombre de recourants, et, par un taux de participation plus élevé le 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. Estimer l'impact du programme de formation sur les *switchers* surestimerait l'effet de la formation.

Application 1 : Les anciens combattants ont-ils des revenus plus faibles que le reste de la population? (Angrist 1990)[5]

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 :

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.

Les auteurs relèvent 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.

Application 2 : Évaluer l'impact de l'instruction obligatoire sur le revenu (Angrist et Krueger 1991)[6]

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. Etant donné que les enfants naissent à différents mois de l'année, tous ne commencent pas l'école exactement au même âge. En effet, si doivent commencer l'école en septembre 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.

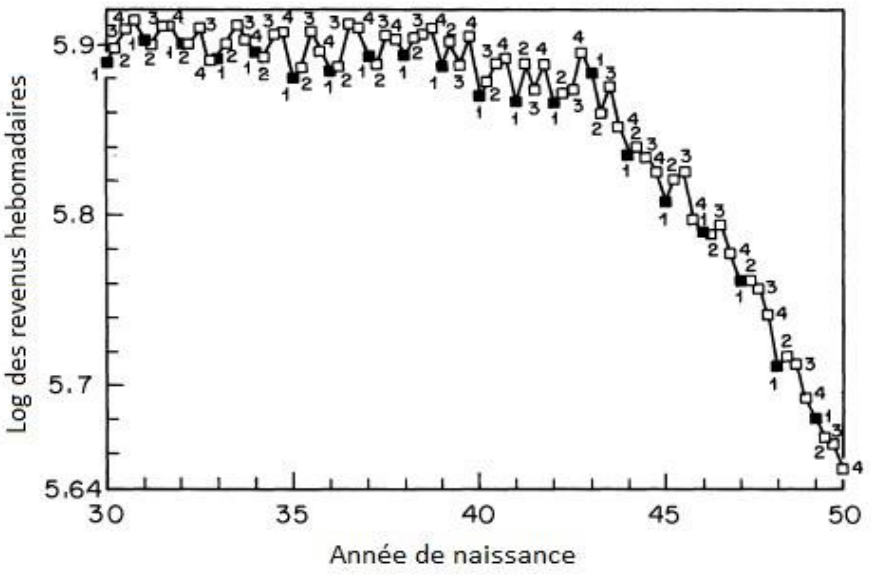
La variable instrumentale utilisée dans cette étude est **le trimestre de l'année durant lequel un individu est né** : janvier-mars, avril-juin, juillet-septembre, et octobre-décembre. Être né à une période de l'année ou une autre est *a priori* indépendant du revenu futur : cela n'impacte pas le milieu socioéconomique 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 :

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 ci-dessous. 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.

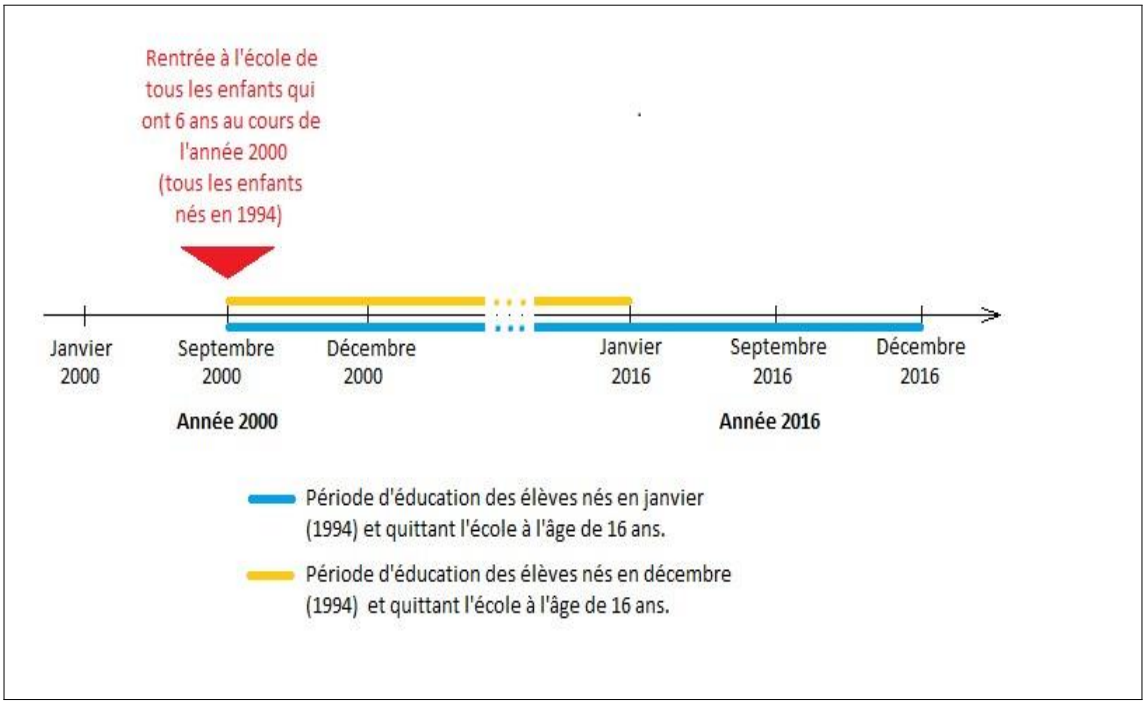
Les auteurs précisent que cette étude est pertinente pour la population d'élèves quittant l'école tôt. En effet, il est moins évident que l'on observe ce même phénomène pour les élèves poursuivant des études supérieures. Julien Grenet a d'ailleurs publié un article[42] sur l'impact du trimestre de naissance 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 3.10 – 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), 979-1014

Figure 3.11 – Durée d'éducation selon la période de naissance



3.5 Quasi-expériences 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. Afin de répondre à cette question le tableau ci-après recense les résultats d'études qui comparent l'estimation par quasi-expérience et par expérimentation randomisée d'une même politique.

D'après ce tableau, on observe qu'à l'exception de la régression par discontinuité, les méthodes non expérimentales, en particulier le *matching* et la double différence, ne parviennent généralement pas à reproduire les résultats expérimentaux. Des recherches sont en cours pour comprendre les sources de ces échecs et tenter de proposer des méthodes plus performantes[18].

Selon Esther Duflo les données non expérimentales, à quelques exceptions près, ne fournissent pas de résultats suffisamment rigoureux pour élaborer des politiques efficaces[18].

Etude	Methode quasi-expérimentale utilisée	Politique évaluée	Variable de résultat	Design utilisé	Source des données non expérimentales	Taille d'échantillon (Traités - Non traités)	Attrition	Résultat
LaLonde (1990)[48]	Regression lineaire (MCO) conditionant sur l'âge, l'éducation et le revenu et Double Différence	National Supported Work Demonstration (NSW) : emplois aidés	Revenus	Design 2	Non recourants dans enquêtes nationales	297-425 (jeunes) ; 600 -585 (mères)	oui (20 à 30%)	MCO et double diff. très biaisés pour l'effet sur les jeunes. MCO proche du vrai impact pour les mères.
Fraker (1987)[40]	Matching	National Supported Work Demonstration (NSW) : emplois aidés	Revenus	Design 2	Non recourants dans enquêtes nationales	566-678 (jeunes) ; 800 - 802 (mères)	Non	Matching sous-estime l'effet pour les jeunes (avec erreur de signe). Le matching estime sans biais l'effet sur les mères.
Heckman (1989)[44]	MCO conditionant sur âge, éducation et revenu, Double Différence et Triple Différence	National Supported Work Demonstration (NSW) : emplois aidés	Revenus	Design 2	Non recourants dans enquêtes nationales	566-678 (jeunes) ; 800 - 802 (mères)	Non	MCO sous-estime l'effet pour les jeunes adultes (avec erreur de signe). MCO estime sans biais l'effet sur les jeunes mères. Double différence biaisée pour les jeunes mais non biaisé pour les mères. Triple différence non biaisée pour les jeunes. Validité de la double différence détectable avec données pre-traitement.
Friedlander (1997)[41]	Avec/Sans ; Avant/Après ; Matching	WORK ; OPTIONS ; SWIM ; ESP	Emploi	Design 2	Groupe de contrôle des autres expérimentations	1000 à 3000 (traités+non traités)	?	Biais du matching très grand, estimation deux fois suppérieure. Biais Avant/Après plus faible (mais égal à effet du traitement)
Heckman (1998)[45]	Matching ; Double Différence	Job Training Partnership Act (JTPA)	Revenus	Design 2	Eligibles non recourants des mêmes bassins d'emploi	508-388	30-50%	Biais du matching égal à l'effet du traitement. Matching sous-estime l'effet. Double diff. moins biaisée lorsqu'appliquée symétriquement par rapport à la date de traitement et combinée avec matching.
Dehejia (2002) (1999)[28, 29]	Matching	National Supported Work Demonstration (NSW) : emplois aidés	Revenus	Design 2	Non recourants dans enquêtes nationales	297-425 (jeunes)	oui (20 à 30%)	Matching reproduit parfaitement les résultats de l'expérimentation en ajountant deux années de revenu pré-traitement aux variables de contrôle.
Smith (2005)[61]	Matching et Double Différence	National Supported Work Demonstration (NSW) : emplois aidés	Revenus	Design 2	Non recourants dans enquêtes nationales	297-425 (jeunes)	oui (20 à 30%)	Biais du matching égal à l'effet du traitement. Matching sous-estime l'effet. Double diff. moins biaisée. Les résultats de D&W sont dus à un choix de l'échantillon dans lequel le biais de sélection est absent. Il est dû au fait de vouloir utiliser un grand nombre d'observations avant le traitement.
Arceneaux (2006)[8]	Matching	Voter Mobilization Experiment	Participation électorale	Design 3	Eligibles non recourants	84503 -2,390,424	non	Biais du matching très large. Matching surestime l'effet du traitement (x3)
Cook (2008)[25]	Regression par discontinuité	Remedial writing class ; Progresa ; Kentucky Job Training Program	Notes ; Assiduité à l'école ; Revenus	Design2	Non éligibles (sous le seuil)	39-69 ; 6000-4000 ; 1222-742	?	Biais de RDD très faible. Signe et significativité la plupart du temps en accord avec les résultats expérimentaux.
Blehaut Aparaitre[13]	Matching	Expérimentation Jeunes Diplômés	Emploi	Design 3	Eligibles non recourants	13000-5812	oui (20%)	Biais du matching égal à l'effet du traitement. Matching sous-estime l'effet.

4. Conclusion

Les politiques publiques sont évaluées en amont de la décision (évaluation *ex ante*) afin qu'en soient bien pesés les avantages et inconvénients (analyses coût-avantage dites aussi analyses coût bénéfice[18]). En pratique les coûts sont souvent plus faciles à estimer que les avantages (amélioration de la santé, de l'éducation, de l'environnement…) et ceci est particulièrement vrai pour les politiques sociales dont les effets dépendent crucialement du facteur humain. Les réactions des individus qui bénéficient de ces politiques (aide au retour à l'emploi, incitation à la scolarisation etc.) sont difficiles à anticiper car elles dépendent des caractéristiques propres à chacun (situation personnelle, croyances etc.). 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 à l'ensemble de la population.

Lutter contre l'échec scolaire

« Plusieurs expériences menées aux États-Unis ont démontré qu'une intervention publique suffisamment précoce était capable d'infléchir profondément les trajectoires scolaires d'enfants initialement parmi les plus en difficulté. Par exemple, les effets très durables des interventions *Abecedarian* (qui propose aux enfants une prise en charge en crèche avec un programme intensif en stimulation du langage) ou *Perry Preschool Project* (qui combine des prises en charge en milieu préscolaire et des visites à domicile) sont désormais bien établis. Les programmes reconnus aujourd'hui parmi les plus efficaces sont toutefois très coûteux et quasi impossibles à généraliser à grande échelle. Ils agissent simultanément sur plusieurs dimensions du problème et la question reste entière d'identifier celles de leurs composantes qui peuvent être à la fois efficaces et généralisables »

Dominique Goux, Marc Gurgand, Eric Maurin, avec Adrien Bouguen, « Évaluation d'impact du dispositif Coup de Pouce Clé - Rapport final pour le Fonds d'Expérimentation pour la jeunesse », mai 2013

Il n'est pas simple d'appréhender les résultats d'une expérimentation. Comparer le groupe d'agents bénéficiaires à un groupe de non bénéficiaires, suppose que les deux populations soient comparables. Malheureusement, les processus de sélection des bénéficiaires des politiques publiques (critères d'éligibilité, décision des individus ou des administrateurs du programme) produisent généralement des groupes de bénéficiaires et de non bénéficiaires non similaires. Une comparaison directe des résultats de ces deux groupes ne permet généralement pas de mesurer l'effet causal de la politique. Cette difficulté est combattue au plan technique par deux approches :

✓ L'expérimentation randomisée. On réalise l'expérimentation *in vivo* en sélectionnant les deux groupes par un tirage aléatoire au sein de la population. Cependant, la réalité du terrain étant souvent imparfaite au regard des conditions théoriques requises, des corrections doivent être introduites. Elles le sont grâce à des design spécifiques d'expérimentations dont le document montre qu'ils permettent une estimation non-biaisée de l'impact d'une mesure.

✓ Les quasi-expériences. On se passe de l'expérimentation *in vivo* et on utilise des données d'observation pré-existantes pour reproduire les résultats expérimentaux (économétrie sur données individuelles). Pratiques par le fait qu'elles mobilisent moins de moyens, puisqu'elles ne nécessitent pas de tirage aléatoire, elles ne semblent pas toujours un bon substitut aux expérimentations (d'après la recension des études consacrées à ce sujet).

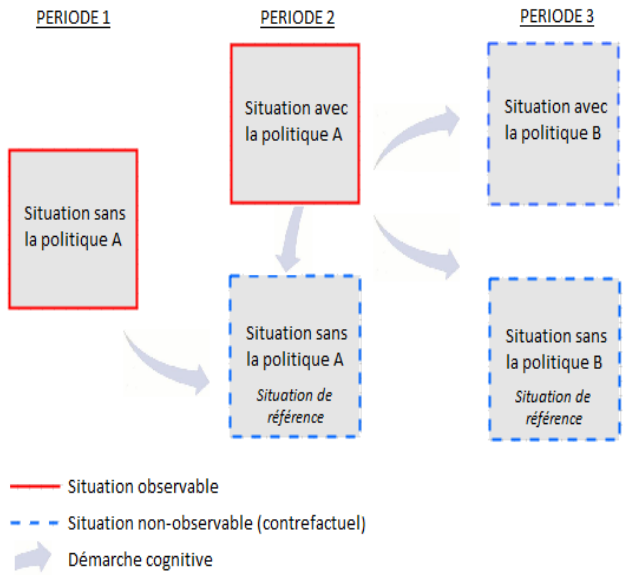
L'expérimentation semble ainsi une méthode efficace (*cf.* encadré sur la lutte contre l'échec scolaire) dont les potentialités font aujourd'hui consensus : « des progrès importants peuvent être accomplis en testant les dispositifs nouveaux sur une base territoriale, en sélectionnant

[18] Pour le décideur public, il s'agit de savoir si une politique crée plus de valeur pour la collectivité qu'elle n'en détruit. Elle crée de la valeur, du bien-être, en améliorant la santé, l'environnement, l'éducation etc. mais elle en détruit car elle consomme des ressources (coût de mise en œuvre mais aussi, parfois, effets négatifs collatéraux). Le principe de l'analyse coût-avantage (*i.e* analyse coût-bénéfice) est d'évaluer en termes monétaires ces deux postes afin de dégager la « valeur nette » de la politique : si celle-ci est positive, la politique est bénéfique pour la collectivité et doit donc être mise en œuvre (dans cette configuration étudiée ou dans une autre éventuellement plus avantageuse) sinon elle doit être rejetée. Quand les avantages ne peuvent être monétarisés on utilise l'analyse coût-efficacité : pour un avantage donné (année supplémentaire de scolarisation, par exemple, *cf.* encadré, De l'expérimentation à l'analyse coût-efficacité) on retient la mesure qui produit cet avantage à moindre coût.

les bonnes politiques pour les généraliser, et en arrêtant celles qui ne donnent pas de résultats probants»[19]. Cependant l'expérimentation mobilise des ressources importantes (150 millions d'euros, par exemple, pour le fonds d'expérimentations pour la jeunesse) et s'inscrit dans la durée (l'expérimentation du programme *Job training Partnership Act* (JTPA) aux États-Unis a été décidée en 1985 et ne s'est achevée qu'en 1993).

Réussir une expérimentation implique donc un engagement fort du décideur politique, qui doit accepter d'attendre les résultats de l'expérimentation pour prendre sa décision. Cela implique aussi, de la part du responsable de l'expérimentation, le respect de « bonnes pratiques » aujourd'hui bien documentées : une organisation rigoureuse, un protocole d'expérimentation bien défini, des variables d'intérêt soigneusement choisies pour éclairer les politiques, une bonne coopération des acteurs de terrain et une collaboration à long terme entre les chercheurs et le décideur politique.

Figure 4.1 – Un système d'évaluation continu



De l'expérimentation à l'analyse coût-efficacité

Dans un article* Esther Duflo compare différentes mesures visant à augmenter la durée de scolarité des enfants. Les résultats des expérimentations montrent que le coût, par enfant, d'une année d'éducation supplémentaire varie beaucoup suivant les stratégies adoptées. Si ce coût est modique, 3,50 $, pour la stratégie de déparasitation des enfants infectés par des parasites intestinaux - ces infections pénalisent l'assiduité à l'école -, il peut dépasser 200 $ pour d'autres stratégies. Le coût atteint même 6.000 $ pour le volet Education Primaire de PROGRESA, programme mexicain de transferts sociaux conditionnels mais celui-ci poursuit d'autres objectifs que l'éducation. Cette diversité de coût pour un même résultat souligne à quel point le taux de rentabilité des investissements publics peut diverger si l'on ne sélectionne pas soigneusement les stratégies à adopter.

* Banerjee Abhijit V. et Duflo Esther, « L'approche expérimentale en économie du développement », *Revue d'économie politique*, 2009/5 Vol. 119, p. 691726.

L'expérimentation est un « *processus d'apprentissage en continu* » apportant une « *base de connaissance exploitable* »[35] pour la décision » (Esther Duflo). Elle joue un rôle tout au long du cycle de vie de la politique étudiée. Tout d'abord, les informations issues de la phase expérimentale alimentent les analyses coût-avantage (ou les analyses coût-efficacité, *cf.* encadré ci-dessus) qui révèlent quelles sont les mesures les plus efficaces. Ensuite, une fois celles-ci mises en œuvre, les systèmes d'observation mis en place pour l'expérimentation renseignent sur d'éventuels effets pervers liés à l'extension de la politique. Enfin l'analyse ex post permet de vérifier que les objectifs sont atteints au moindre coût (analyse coût-efficacité) ou, mieux encore, que la politique maximise le bien-être collectif (analyse coût-bénéfice) et, si ce n'est pas le cas, de mettre en place une politique plus adaptée. Ainsi l'enjeu est-il d'articuler l'ensemble des outils économiques pour qu'ils servent au mieux la décision publique dans le cadre d'un système d'évaluation en continu (*cf.* figure 5.1).

[19] « Principaux enseignements des débats « France dans 10 ans » », CGSP, 20 novembre 2013.

Bibliographie

[1] Principaux enseignements des débats « France dans 10 ans », *CGSP*, 11/2013.

[2] *Cahiers de l'évaluation*, 2012.

[3] Évaluation des politiques publiques. *Note n'1 du Conseil d'Analyse Économique*, 8(24), 2013.

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

[5] Angrist J. D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records", *American Economic Review*, 80(3): 313– 36, June.

[6] Angrist J. D. and Krueger A. B. (1991), "Does compulsory school attendance affect schooling and earnings?", *The Quarterly Journal of Economics*, 106(4): 979–1014, November.

[7] Angrist J. D. and Lavy V. (1999), "Using Maimonides' rule to estimate the effect of class size on scholastic achievement", *The Quarterly Journal of Economics*, 114(2): 533–575.

[8] Arceneaux K., Gerber A.S., and Green D.P. (2006), "Comparing Experimental and Matching Methods using a large-scale voter Mobilization Experiment", *Political Analysis*, 14(1): 37 – 62.

[9] Ashraf N., Karlan D. and Yin W. (2006), "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines", *The Quarterly Journal of Economics*, 121(2): 635–672.

[10] Banerjee A. V. et Duflo E. (2009), « L'approche expérimentale en économie du développement », *Revue d'économie politique*, 119 :691–726, Mai.

[11] Banerjee A. V., Duflo E., Glennerster R., and Kothari D. (2010), "Improving Immunization Coverage in Rural India: Clustered Randomised. Controlled Evaluation of Immunization Campaigns with and without Incentives", *BMJ*, 340, 5.

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

[13] Behaghel L., Crépon C. and Gurgand M. (2012), "Private and Public Provision of Counseling to Job-seekers: Evidence from a Large Controlled Experiment", *Discussion Paper series*, Forschungsinstitut zur Zukunft der Arbeit, Institue for the Study of Labor, http://ftp.iza.org/dp6518.pdf.

[14] Bell S. H., Cave G., Doolittle F., Lin W., Bloom H. S., Orr L. L. and Bos J. M. (2007), "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study. *The Journal of Human Resources*, XXXII - 3: 549–576.

[15] Bléhaut M. and Rathelot R., « Expérimentation contrôlée contre appariement : le cas d'un dispositif d'accompagnement de jeunes diplômés demandeurs d'emploi », *Économie et Prévision*, À paraître.

[16] Bloom H.S., Orr Larry L., Bell S. H., Cave G., Doolittle F., Lin W. and 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*, 32(3) : 549–576

[17] Burtless G. (1985), "The Case for Randomized Field Trials in Economic and Policy Research", *The Journal of Economic Perspectives*, 9: 63–84.

[18] Card D. and Krueger A. B. (1993), “Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania”, *Working Paper* 4509, National Bureau of Economic Research, October.

[19] Chabé-Ferret S. (2015), "Analysis of the Bias of Matching and Difference-In-Difference under Alternative Earnings and Selection Processes," *Journal of Econometrics*, Volume 185, Issue 1, Pages 110–123.

[20] Chabé-Ferret S. (2013), “Lecture 3: Randomized Control Trials, Basics - Program Evaluation Methods”, *Toulouse School of Economics*.

[21] Chabé-Ferret S. (2013), “Lecture 5: Matching - Program Evaluation Methods”, *Toulouse School of Economics*.

[22] Chabé-Ferret S. (2013), “Lecture 6: Differences in Differences and fixed effects - Program Evaluation Methods”, *Toulouse School of Economics*.

[23] Chabé-Ferret S. (2013), “Lecture 7: Regression Discontinuity Designs - Program Evaluation Methods”, *Toulouse School of Economics*.

[24] Chabé-Ferret S. and Subervie J. (2012), “How Much Green for the Buck? Estimating Additional and Windfall Effects of French Agro-Environmental Schemes by DID-Matching”, *TSE Working Papers* 12-357, Toulouse School of Economics (TSE), July.

[25] Chay K.Y. and Greenstone M. (2005), “Does air Quality Matter? Evidence from the Housing Market”, Working Paper 6826, *Journal of Political Economy*, December.

[26] Cook T. D. and Wong V. C. (2008), “Empirical Tests of the Validity of the Regression Discontinuity Design”, *Annals of Economics and Statistics / Annales d’Economie et de Statistique*, 91/92:127–150, July.

[27] Crépon B., « Évaluation des politiques publiques par expériences contrôlées », *Présentaion - CREST, INSEE*, page http://www.crest.fr/ckfinder/userfiles/files/pageperso/givord/eval/slide_experiment_new.pdf.

[28] Crépon B., Duflo E., Gurgand M., Rathelot R. and Zamora P. (2012), “Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment”. *Working Paper* 18597, National Bureau of Economic Research, December.

[29] Dehejia R. H. and Wahba S. (1999), “Causal Effects in Non-Experimental Studies: Reevaluating the Evaluation of Training Programs”, *Journal of the American Statistical Association*, 94(448):1053–1062, Dec.

[30] Dehejia R.H. and Wahba S. (2002), “Propensity Score-Matching Methods For Nonexperimental Causal Studies”, *The Review of Economics and Statistics*, 84(1):151–161, February.

[31] Duflo E. (2001), “Schooling and Labor Market consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment”, *Working Paper* 7860, *The American Economic Review*, August.

[32] Duflo E., Glennerster R. and Kremer M. (2008), “Using Randomization in Development Economics Research: A Toolkit”, volume 4 of *Handbook of Development Economics*, chapter 61, pages 3895–3962. Elsevier.

[33] Duflo E. and Pande R. (2007), "Dams". *The Quarterly Journal of Economics*, 122(2): 601–646.

[34] Effet d’Hawthorne. *http://www.psychologuedutravail.com/psychologie-dutravail/effet-hawthorne/*.

[35] Ferracci M. et Wasmer E. (2011), « État moderne, État efficace ».

[36] Fougère D., « Expérimenter pour évaluer les politiques d'aide à l'emploi : les exemples anglo-saxons et nord – européens », *Revue Française des Affaires Sociales, 54(1), 111-144.*, 8(24).

[37] Fraker T. and Maynard R. (1987), "The Adequacy of Comparison Group Designs for Evaluations of Employment-related Programs", *The Journal of Human Resources*, 22(2):194–227, April.

[38] Friedlander D., Greenberg D. H., and Robins P. K. (1997), "Evaluating Government Training Programs for the Economically Disadvantaged", *Journal of Economic Literature*, 35(4) :1809–1855, December.

[39] Grenet J. (2006), "Academic Performance, Educational Trajectories and the Persistence of date of Birth effects", *Evidence* from France.

[40] Guitard J., Gurgand M., Behaghel L. et Crépon B. (2009), « Évaluation d'impact de l'accompagnement des demandeurs d'emploi par les opérateurs privés de placement et le programme cap vers l'entreprise », *Rapport final*.

[41] Heckman J.J. (1991), "Randomization and Social Policy Evaluation", *NBER Technical Working Papers* 0107, National Bureau of Economic Research, Inc, July.

[42] Heckman J. J. and Hotz V. J. (1989), "Choosing Among Alternative Non-Experimental Methods for Estimating the Impact of Social Programs: the Case of Manpower Training", *Journal of the American Statistical Association*, 84(408): 862–874.

[43] Heckman J. J., Ichimura H., Smith J. A. and Todd P. E. (1998), "Characterizing Selection Bias Using Experimental Data", *Econometrica*, 66:1017–1099.

[44] Horowitz C. F., Manski J. L. (2000), "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome data" *Journal of the American Statistical Association*, 95.

[45] L'Horty Y. (2009), « Paradoxes de l'évaluation du RSA », *revue Projet, n°*308*,* 8(24)*.*

[46] L'Horty Y. et Petit P. (2010), « Évaluation aléatoire et expérimentations sociales », *document de travail n'135, Centre d'études de l'emploi*, 8(24).

[53] Kramer M.S. and Shapiro S. H. (1984), "Scientific Challenges in the Application of Randomized Trials", *JAMA*, 252(19):2739–2745.

[47] LaLonde R.J. (1986), "Evaluating the Econometric Evaluation of Training Programs with Experimental Data", *American Economic Review*, 76: 604–620.

[48] Lee D. S. (2000), "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects", *Review of Economic Studies*, 76, 1071-1102.

[49] Lee D. S. and Lemieux T. (2009), "Regression Discontinuity Designs in Economics", *Working Paper* 14723, National Bureau of Economic Research, February.

[53] Magnac T. (2000), « L'apport de la microéconomie à l'évaluation des politiques publiques », *Cahiers d'économie et sociologie rurales*, n°54.

[54] Miguel E. and Kremer M. (2001), "Worms: Education and Health Externalities in Kenya", *NBER Working Papers* 8481, National Bureau of Economic Research, Inc, September.

[55] Moffitt R. (2002), "The Role of Randomized field Trials in Social Science Research: a Perspective from Evaluations of Reforms of Social Welfare Programs", *CeMMAP working papers* CWP23/02, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, December.

[56] Okbani N., « Le non recours au RSA activité : étude auprès des allocataires de la caf de la gironde », *Technical report*, Caf de la Gironde.

[57] Olken B. A. (2005), "Monitoring Corruption: Evidence from a Field Experiment in Indonesia", *Working Paper* 11753, National Bureau of Economic Research, November.

[58] Perret B. (2008), « L'évaluation des politiques publiques. Entre culture du résultat et apprentissage collectif. » *Esprit* (350) : 142-59.

[50] Piketty et Valdenaire (2006), « L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français », *Les dossiers - Enseignement scolaires*, 173.

[53] Rosenthal R. (1966), "Experiment Effects in Behavioral research", *Appeton-Century-Crofts*, page 464.

[54] Smith J. A. and Todd P. E. "Does Matching Overcome LaLonde's Critique of Non-Experimental Estimators?" *Journal of Econometrics*, 125(1-2):305–353, 2005.

[51] Skoufias E. and Bonnie McClafferty B. (2001), "Is Progresa Working? Summary of the Results of an Evaluation by ifpri", *IFPRI*, 118.

[53] Thistlethwaite D. L. and Donald T. (1960), "Campbell. Regression-Discontinuity Analysis: An Alternative to the ex-post facto Experiment", *Journal of Educational Psychology*, 51(6):309–17.

[54] Todd P. E. (2008), "Evaluating Social Programs With Endogenous Program Placement and Selection of the Treaded", *Handbook of Development Economics*, 4.