Guide

Septembre 2016

Évaluation

Comment évaluer l'impact des politiques publiques ?

Un guide à l'usage des décideurs et praticiens

Rozenn Desplatz
Marc Ferracci

FRANCE STRATÉGIE
ÉVALUER. ANTICIPER. DÉBATTRE. PROPOSER.

COMMENT ÉVALUER L'IMPACT DES POLITIQUES PUBLIQUES ?

Un guide à l'usage des décideurs et praticiens

Rozenn Desplatz
France Stratégie

Marc Ferracci
Université Panthéon Assas – CRED - CREST-ENSAE
Conseiller scientifique, France Stratégie

FRANCE STRATÉGIE

SEPTEMBRE 2016

SOMMAIRE

INTRODUCTION **5**

CHAPITRE 1 – MESURER L'IMPACT D'UNE POLITIQUE PUBLIQUE SUR SES BÉNÉFICAIRES **7**

1. Corrélation n'est pas causalité : la nécessité de construire un contrefactuel crédible 7
   - *Choisir des indicateurs pertinents* *8*
   - *Les mécanismes de sélection, facteurs de biais dans l'évaluation* *9*
2. Le choix de la méthode 11
   - *Des méthodes fondées sur des hypothèses et des données différentes* *12*
   - *Quelle méthode choisir ?* *21*
3. Le ciblage des politiques et des ressources publiques 22
   - *L'hétérogénéité des effets d'une politique publique* *23*

CHAPITRE 2 – GÉNÉRALISER LES RÉSULTATS DE L'ÉVALUATION **25**

1. Des effets des politiques sur les non-bénéficiaires 26
2. Des effets différents selon la population couverte 26

CHAPITRE 3 – EXPLIQUER L'ÉCHEC OU LE SUCCÈS D'UNE POLITIQUE **29**

1. L'approche structurelle : décomposer les effets et les quantifier 29
   - *Le principe : modéliser les mécanismes économiques* *29*
   - *La force de cette approche : comprendre les mécanismes économiques à l'œuvre* *32*
   - *Sa faiblesse : la sensibilité aux hypothèses du modèle théorique* *33*
   - *Une illustration : l'évaluation d'une mesure fiscale aux Pays-Bas* *35*

2. L'approche qualitative : connaître les perceptions et les pratiques des acteurs de terrain 37

Le principe : interroger et observer les acteurs *37*

La force de cette approche : comprendre les facteurs d'échec et de succès de la politique évaluée *39*

Sa faiblesse : de petits échantillons, la fiabilité des réponses et la difficulté à isoler les effets *41*

Une illustration : l'évaluation du revenu contractualisé d'autonomie *43*

CHAPITRE 4 – COMPARER LES EFFETS DE POLITIQUES DIFFÉRENTES 47

1. L'évaluation de traitements multiples visant un même objectif 47

Le principe : comparer deux à deux les dispositifs *48*

Une illustration : l'évaluation de dispositifs d'emploi pour les jeunes en France *48*

2. L'analyse coût-efficacité ou coût-bénéfice comme critères d'efficience 49

Le principe : comparer les coûts et les bénéfices *49*

Une llustration : l'évaluation du programme d'éducation Perry Preschool *51*

CHAPITRE 5 – DE L'ÉVALUATION À LA RECOMMANDATION DE POLITIQUE PUBLIQUE 55

S'appuyer sur les travaux disponibles *55*

Prendre garde aux limites de chaque méthode *56*

S'interroger sur la nature transposable ou non des résultats à une population plus large *57*

Favoriser la pluridisciplinarité des approches *58*

Sélectionner les résultats disponibles et les équipes *58*

CONCLUSION 61

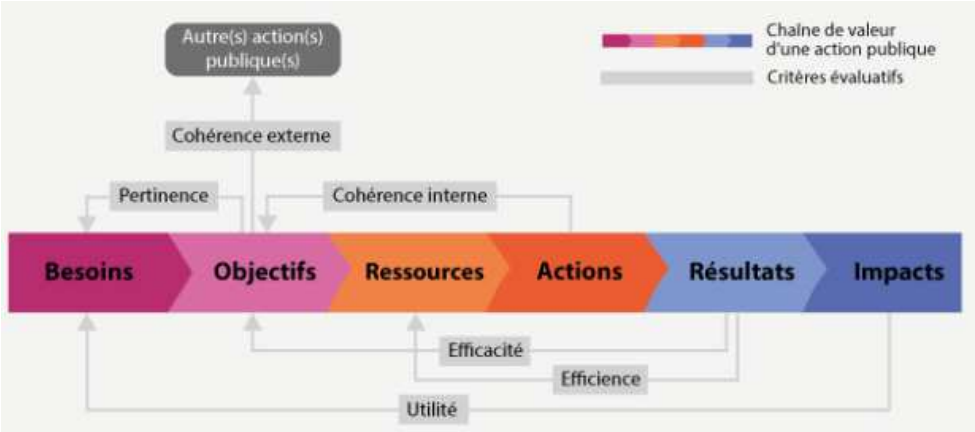
BIBLIOGRAPHIE 63

INTRODUCTION

L'efficacité et l'efficience des interventions publiques sont un enjeu majeur dans la plupart des pays développés où un niveau élevé des dépenses publiques se conjugue à une défiance croissante à l'égard des institutions politiques. C'est pourquoi une démarche objective d'évaluation de l'impact des politiques publiques est nécessaire.

Cette évaluation d'impact doit être distinguée d'autres pratiques, également légitimes mais ne visant pas les mêmes objectifs comme le contrôle, l'audit ou le conseil. **Elle doit être également distinguée de l'évaluation portant sur d'autres critères comme la pertinence,** la cohérence, la mise en œuvre ou l'utilité (schéma 1).

Schéma 1
Les critères d'évaluation d'une politique publique



Source : SG-MAP (2015)

Définie de façon simple, une **évaluation d'impact** réussie vise à établir la situation qu'aurait connue la société *en l'absence* de la politique évaluée. Cette situation fictive, aussi appelée *contrefactuelle*, permet, en la comparant à la situation effectivement observée, de déduire une relation de causalité entre l'intervention publique et un indicateur jugé pertinent (la santé, l'emploi, l'éducation, etc.).

Ce guide présente, de manière non technique, les différentes méthodes permettant d'atteindre cet objectif, en insistant sur les contraintes opérationnelles entourant la mise en œuvre des évaluations. Par souci d'homogénéité, elle se limitera à l'évaluation *ex post*, c'est-à-dire portant sur des dispositifs existants, pour lesquels des données sont disponibles[1].

De fait, les indicateurs à l'aune desquels on sera à même d'apprécier l'impact d'un dispositif public, donc les questions auxquelles il sera permis de répondre *in fine*, dépendent crucialement de l'accessibilité et de la qualité des données. Celles-ci conditionnent également le choix des méthodes à retenir, compte tenu de la double contrainte de limiter au maximum les biais et de réduire le coût de l'évaluation.

Ce guide expose d'abord le problème fondamental de l'évaluation d'impact et les différentes méthodes disponibles pour mesurer les effets d'un dispositif sur ses bénéficiaires. Puis il élargit la question à la mesure des effets indirects d'une politique et aux conséquences de son extension éventuelle. Il montre ensuite comment analyser les raisons du succès ou de l'échec d'une politique, avant d'aborder la mesure des effets de politiques multiples et leur comparaison. La conclusion propose un certain nombre de recommandations afin d'assurer le succès d'une évaluation d'impact.

.

[1] Par opposition, l'évaluation *ex ante* cherche à quantifier l'impact d'une politique qui n'est pas encore en œuvre, au moyen le plus souvent de simulations qui reposent sur des modèles représentant le fonctionnement de l'économie.

CHAPITRE 1

MESURER L'IMPACT D'UNE POLITIQUE PUBLIQUE SUR SES BÉNÉFICIAIRES

Sont présentées ici les principales méthodes permettant de mesurer l'impact d'une politique publique[2] sur les bénéficiaires directs de cette politique[3] [4]. Ces méthodes ne reposent pas sur une théorie établissant les mécanismes par lesquels l'action publique affecte le bien-être des citoyens. Elles sont dites non structurelles ou « athéoriques », et s'opposent en cela aux méthodes d'évaluation dites « structurelles », qui ont pour fondement un modèle théorique dont on cherche à estimer certains des paramètres représentant l'effet de la mesure publique.

Si les méthodes de l'évaluation peuvent parfois sembler techniques, elles sont en réalité guidées par des questions simples : la politique a-t-elle atteint son objectif ? A-t-elle engendré des effets non anticipés par le législateur ? Cela suppose de définir en amont et avec suffisamment de précision les objectifs poursuivis, puis d'identifier une relation de causalité entre la politique que l'on cherche évaluer et ces objectifs.

1. Corrélation n'est pas causalité : la nécessité de construire un contrefactuel crédible

La difficulté de l'évaluation causale est de ne pas s'arrêter à de simples relations de corrélation entre le « traitement » (i.e. la politique que l'on cherche à évaluer) et la variable de résultat. Par exemple, la corrélation entre les dépenses de soins des individus et leur santé pourra être négative parce que les individus qui dépensent le plus sont généralement les moins bien portants. Dans ce cas, la relation apparente

[2] Dans la suite de ce guide, les termes de « politique », de « dispositif », de « programme » ou de « mesure » seront utilisés de manière alternative pour désigner une intervention publique touchant des agents économiques clairement identifiés.

[3] Pour une présentation approfondie de ces méthodes, voir Givord (2015).

[4] Pour une présentation vulgarisée de ces méthodes, voir Ferracci et Wasmer (2011).

entre dépenses de soins et santé est la conséquence d'une causalité inverse : une santé dégradée accroît les dépenses. Un autre piège réside dans l'existence de facteurs susceptibles d'affecter à la fois le traitement et la variable de résultat, mais qui ne sont pas forcément observables par l'évaluateur. Ainsi, les individus les plus riches consentent une part plus importante de leur revenu aux dépenses de soins, mais ils exercent aussi en général des activités moins pénibles. Pour cette raison, la corrélation apparente entre dépenses de soins et état de santé peut souffrir d'un biais, dont il faut tenir compte pour ne pas l'interpréter comme un effet causal des dépenses sur la santé.

Le but de l'évaluation consiste donc à répondre de façon crédible à la question suivante : quelle situation la société aurait-elle connue en l'absence de la politique menée ? Cette situation hypothétique, aussi nommée situation *contrefactuelle*, est la clé d'une évaluation réussie. Un chômeur aurait-il retrouvé un emploi sans le stage de formation qu'il a suivi ? Un ménage aurait-il accédé à la propriété immobilière sans le dispositif d'aide à la pierre dont il a bénéficié ? Un élève aurait-il eu de meilleurs résultats scolaires dans une classe de taille plus petite ?

Le but de l'évaluation d'impact est de reconstituer ce contrefactuel, et d'établir ainsi une véritable relation de causalité entre la politique menée et les objectifs poursuivis. Ceci suppose au préalable de choisir des indicateurs pertinents pour l'évaluation, puis d'identifier les mécanismes qui contribuent à sélectionner certains individus dans le dispositif que l'on cherche à évaluer.

Choisir des indicateurs pertinents

Quelles informations l'évaluation doit-elle produire pour être utile à l'action publique ? Répondre à cette question suppose de définir au préalable la ou les variables sur lesquelles la politique considérée est susceptible d'avoir un impact. Dès lors que l'évaluation vise à quantifier les effets d'un dispositif, ces variables doivent évidemment être mesurables, mais aussi présenter un intérêt social. Ainsi, pour évaluer une politique d'allégement de cotisations sociales, on s'intéressera au nombre d'emplois créés ou sauvegardés. Dans le cadre d'une politique de réduction de la taille des classes, on cherchera à mesurer l'évolution des résultats scolaires. Pour évaluer l'impact de l'introduction d'une franchise médicale, on choisira par exemple un indicateur du taux d'accès aux soins.

En règle générale, de telles informations sont accessibles, plus ou moins aisément, au sein de fichiers administratifs ou statistiques existants. Cependant il est parfois nécessaire de les collecter pour permettre l'évaluation. Lorsque l'on évalue les effets

d'un dispositif, il est en effet souhaitable de disposer d'une variété d'indicateurs. Ceci permet de mesurer certains effets non anticipés du dispositif, qui présentent néanmoins un intérêt pour la collectivité. Ainsi une politique d'insertion pour les jeunes décrocheurs gagnera à voir mesurer son impact sur l'emploi des bénéficiaires, mais aussi sur leur santé, ou sur leur propension ultérieure à commettre des délits. Ceci suppose d'accéder à des données variées et donc de procéder fréquemment à des appariements de sources statistiques.

Les mécanismes de sélection, facteurs de biais dans l'évaluation

Comparer de façon « naïve » la situation des bénéficiaires et des non-bénéficiaires d'une mesure ne permet pas de mesurer l'effet causal de celle-ci. Il faut tenir compte du fait que certains individus ont pu participer à un dispositif car ils présentent des caractéristiques différentes de celles des non-bénéficiaires. Ainsi, l'évaluation d'un dispositif d'accompagnement renforcé pour les demandeurs d'emploi ne saurait reposer sur la simple comparaison des individus « traités » et « non traités ». Il est en effet probable que les agents du service public de l'emploi choisiront d'affecter les individus les moins employables à l'accompagnement renforcé. Dans ce cas la comparaison des individus bénéficiaires et non bénéficiaires produira une estimation biaisée de l'effet du dispositif.

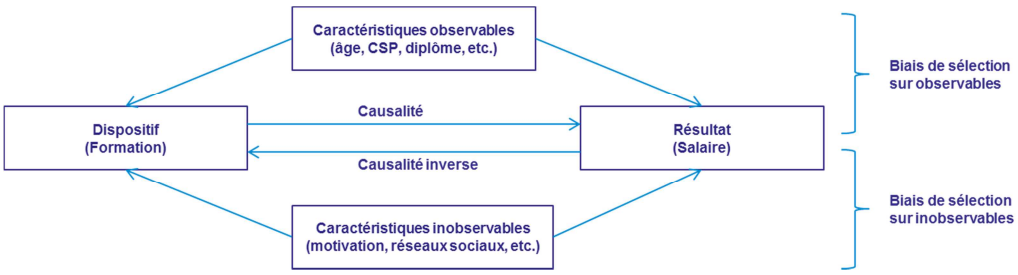
Évaluer l'effet spécifique d'une politique publique suppose donc de neutraliser ce type de biais de sélection, en comparant des agents économiques (individus, entreprises, communes, etc.) aux caractéristiques similaires. Celles-ci sont parfois observables par l'évaluateur car elles figurent dans les données existantes[5]. Mais dans d'autres cas il existe des facteurs inobservables (la motivation d'un individu, ses réseaux de relations par exemple) qui peuvent influencer conjointement la probabilité de recevoir le traitement et la variable de résultat.

Le schéma 2 résume ce problème fondamental de l'évaluation en l'appliquant à l'évaluation d'un programme de formation pour les salariés. On y distingue le dispositif (le traitement), la variable de résultat et des variables susceptibles d'engendrer un biais de sélection. Dans cet exemple le fait de passer en formation

[5] Parmi les caractéristiques observables qu'il convient de contrôler figurent l'ensemble des dispositifs auxquels l'agent économique a accès, et qui peuvent avoir une influence conjointe sur sa participation au dispositif évalué et sur la variable de résultat. Ainsi, dans le cas du Crédit Impôt Compétitivité Emploi (CICE) il est utile de disposer d'informations sur les autres aides qu'est susceptible de recevoir l'entreprise (comme le Crédit Impôt Recherche) pour évaluer l'impact sur les dépenses de R & D, afin de la comparer avec des firmes similaires.

peut avoir un effet causal sur le salaire. Mais le fait de recevoir un salaire élevé peut, à l'inverse, rendre plus probable le passage en formation, ce qui pose un problème de causalité inverse[6]. En outre certaines caractéristiques observables telles que l'âge ou le diplôme peuvent avoir une influence à la fois sur le traitement et sur la variable de résultat, de sorte que la comparaison des individus formés et non formés révèle l'effet du traitement, mais aussi l'effet de ces différences de caractéristiques entre les deux groupes. D'autres facteurs, inobservables ceux-là, peuvent également être dans ce cas.

Schéma 2
Les biais de sélection dans l'évaluation d'un programme



Source : Les auteurs

Ce problème fondamental posé à l'évaluation nécessite de comprendre les mécanismes qui engendrent de la sélection. Parfois celle-ci provient du comportement des agents publics, ou bien de règles juridiques déterminant de façon explicite les critères d'éligibilité à un dispositif – le revenu des parents pour l'obtention d'une bourse universitaire, par exemple. En d'autres circonstances, la participation au dispositif dépend du comportement et des caractéristiques des individus eux-mêmes. Par exemple, la propension d'un individu à déposer une demande pour une aide dépend du bénéfice que l'individu espère en tirer. Ne pas tenir compte de cet aspect des choses va évidemment conduire à surestimer l'effet de la mesure de politique publique s'il n'est évalué que pour les individus ayant fait la demande de l'aide, car ce sont eux qui *a priori* tirent le mieux parti de celle-ci. En présence de ce type de biais, le véritable effet causal de la mesure sur l'ensemble des personnes éligibles à cette dernière serait inférieur à l'effet évalué sur les seuls bénéficiaires. Or c'est cet

[6] C'est par exemple le cas si le rendement de la formation est plus élevé pour les individus les plus qualifiés. L'analyse des systèmes de formation confirme ce biais au profit des travailleurs les plus diplômés, qui accèdent plus fréquemment à la formation que les moins diplômés. Voir sur ce point Ferracci (2013).

effet sur la population éligible qu'il est utile de mesurer pour juger de la pertinence de la mesure et de son ciblage.

Dans le cas du Revenu de Solidarité Active (RSA), par exemple, la complexité des démarches administratives a pu dissuader certains actifs de demander l'aide financière, alors même qu'ils étaient éligibles à celle-ci[7]. Le fait de surmonter les obstacles administratifs et de bénéficier *in fine* du RSA est donc révélateur de caractéristiques individuelles (la motivation, le degré de familiarité avec l'administration…) qui peuvent avoir en elles-mêmes un effet sur le retour à l'emploi. Dans ce cas, l'écart de taux d'emploi parmi les éligibles entre les bénéficiaires et les non-bénéficiaires du RSA certes résulte de l'effet du dispositif – si cet effet existe –, mais aussi, plus simplement, de différences de caractéristiques entre les deux groupes.

Comprendre et identifier l'origine des mécanismes de sélection à l'œuvre dans les politiques publiques est crucial pour l'évaluation, car de cela dépendra le choix de la méthode d'évaluation la plus adaptée pour construire un contrefactuel crédible.

2. Le choix de la méthode

Plusieurs méthodes plus ou moins complexes peuvent être mobilisées pour évaluer le ou les effets causaux d'un programme. Le choix de la méthode est souvent tributaire du dispositif étudié et des données disponibles.

Concrètement, il est nécessaire de savoir : i) s'il existe des différences entre les groupes des individus traités et non traités ; ii) si ces différences sont susceptibles d'avoir une influence sur la variable de résultat (emploi, salaire, santé, etc.). Dans certains cas, ces différences résultent de facteurs observables par l'évaluateur, c'est-à-dire figurant dans les données disponibles (par exemple l'âge, ou le niveau de diplôme). Mais la plupart du temps, ces données sont trop limitées pour tenir compte de dimensions plus subtiles, et donc inobservables, telles que la motivation ou le degré d'information dont disposent les individus. De ce point de vue, toutes les méthodes ne présentent pas le même pouvoir de correction des biais de sélection, et donc pas la même rigueur. Certaines permettent de neutraliser les biais provenant de caractéristiques aussi bien observables qu'inobservables, tandis que d'autres, d'usage souvent plus aisé, ne contrôlent que la sélection provenant de caractéristiques observables.

[7] Voir Comité d'évaluation des expérimentations (2009).

Le choix de la méthode résultera donc d'un arbitrage entre la réduction des biais qu'elle permet (on parle à cet égard de *validité interne* des résultats de l'évaluation) et les contraintes opérationnelles qu'elle engendre.

Des méthodes fondées sur des hypothèses et des données différentes

Dans ce qui suit nous présentons les méthodes les plus courantes permettant d'évaluer l'impact causal d'un dispositif sur ses bénéficiaires. *L'expérimentation contrôlée*, ou *aléatoire*, constitue la méthode privilégiée par de nombreux évaluateurs, car elle présente l'avantage de construire la situation contrefactuelle de façon à la fois rigoureuse et intuitive. Elle permet ainsi d'identifier l'effet causal du traitement sur les bénéficiaires de ce dernier. Ses limites tiennent au coût et à la durée de telles expérimentations, ainsi qu'à la difficulté d'en généraliser les résultats à un public plus large que celui des individus effectivement traités (voir encadré 1).

Un exemple d'expérimentation récent vient illustrer cette démarche, l'initiative « Dix mille permis pour réussir » financée par le Fonds d'expérimentation pour la jeunesse à partir de 2009[8]. Le but de cette expérimentation consistait à mesurer l'impact de l'obtention du permis de conduire sur l'insertion professionnelle et sociale des jeunes. Un tirage au sort a été effectué au sein d'un groupe de candidats âgés de 18 à 25 ans. Dans le groupe test, 10 000 jeunes ont reçu une aide de 1 000 euros pour financer leurs leçons de conduite, tandis que les individus du groupe de contrôle ne recevaient pas d'aide. Rapportée au prix moyen du permis, de l'ordre de 1 500 euros, cette aide apparaissait extrêmement significative.

Encadré 1

Les méthodes d'expérimentation aléatoire

L'intérêt de ces méthodes d'évaluation réside dans une répartition par tirage au sort des individus dans les groupes de contrôle et de traitement. Ce principe d'affectation aléatoire garantit que les individus des deux groupes sont comparables au regard de leurs caractéristiques individuelles observables mais aussi inobservables. Dans ce cadre, l'effet causal du traitement (la politique publique) sera mesuré par l'écart entre la moyenne de la variable d'intérêt (par exemple, le salaire) au sein du groupe de traitement, et la moyenne de la même variable au sein du groupe de contrôle. Il est ici possible de parler d'effet causal, car cet écart entre les deux groupes ne saurait plus désormais résulter de

[8] Voir L'Horty *et al.* (2014).

différences entre les deux groupes et notamment de leurs caractéristiques individuelles, comme dans le cas de l'écart de taux d'emploi entre bénéficiaires et non bénéficiaires du RSA décrit plus haut.

Avantages

. forte validité interne des résultats (contrôle des caractéristiques observables et inobservables) ;

. caractère intuitif conférant un pouvoir de conviction aux décideurs politiques et à l'opinion publique.

Limites et points de vigilance méthodologiques

. validité externe potentiellement limitée : difficulté à généraliser les résultats à des publics plus larges ;

. sources de biais possibles : modification du comportement du groupe de contrôle du fait de l'expérimentation ; refus de certains membres du groupe test de se prêter à l'expérimentation (*non compliance*).

Contraintes opérationnelles

. coût (une certaine taille d'échantillon est nécessaire pour assurer la significativité des résultats) ;

. temporalité parfois trop longue comparée à l'exigence des décideurs politiques;

problèmes éthiques liés à l'introduction d'une discrimination entre individus traités et non traités.

Références méthodologiques

. Bouguen A. et Seban J. (2014), « L'assignation aléatoire comme méthode d'évaluation des politiques publiques », *Économie & Prévision*, (1), p. 119-143.

. Dupont-Courtade L., Chabé-Ferret S., Perbet M. et Treich N. (2015), *Évaluation des politiques publiques : expérimentation randomisée et méthodes quasi-expérimentales.* Document de travail de la DG Trésor, Numéro 2015/01, février.

. Glennerster R. et Takavarasha K. (2013), *Running Randomized Evaluations: A Practical Guide*, Princeton University Press.

Il faut par ailleurs noter que ce tirage a été réalisé au sein d'une population de jeunes ne se trouvant dans leur grande majorité ni dans le système scolaire, ni en emploi, ni en formation. Pour évaluer les effets du programme, la démarche a consisté à suivre les deux groupes, le groupe de traitement ayant bénéficié du dispositif entre janvier 2010 et septembre 2010.

À court terme, c'est-à-dire pendant la durée des cours de conduite, les résultats soulignent l'ampleur des effets d' « enfermement » liés au passage du permis de conduire, qui éloignent les jeunes du marché du travail. Le passage du permis

s'apparente en effet à une véritable formation professionnelle, intense et sélective qui limite d'autres investissements dans les sphères professionnelle et non professionnelle. Ces résultats sont toutefois contrebalancés par des effets positifs sur l'obtention du permis. En effet, deux ans après le début de l'expérimentation 45 % des individus du groupe test ont obtenu le permis, contre 30 % dans le groupe de contrôle, soit un égard de 15 points. En revanche, les effets du dispositif sur l'insertion professionnelle n'apparaissent pas clairement. Il n'existe en effet pas de différence significative entre les taux d'emploi et les niveaux de salaire du groupe test et ceux du groupe de contrôle deux ans après l'entrée dans le dispositif. Ces résultats peuvent paraître surprenants si on les rapporte à d'autres évaluations existantes, menées principalement sur données anglo-saxonnes, qui montrent un effet généralement très positif de l'obtention du permis sur l'emploi. Toutefois, dans le cadre de l'initiative « 10 000 permis pour réussir » l'expérimentation aléatoire s'est doublée d'une enquête qualitative qui fournit des clés d'interprétation utiles des résultats chiffrés[9]. Ceci illustre l'intérêt de mêler approche quantitative et qualitative de l'évaluation (voir *infra*, 3.2.).

Lorsque les expérimentations contrôlées ne sont pas possibles, d'autres méthodes permettent d'établir la causalité en contrôlant les biais provenant de caractéristiques inobservables par l'économètre. Ces méthodes ont comme point commun d'exploiter une variable déterminant l'exposition au traitement de façon exogène. C'est le cas par exemple lorsqu'une mesure de politique publique affecte seulement une partie de la population ou du territoire, pour des raisons sans lien avec la variable de résultat, telles qu'un évènement historique ou météorologique. On parle dans ce cas *d'expérience naturelle* – et non d'expérience aléatoire. Dans ce cas il est possible d'utiliser le reste de la population comme contrefactuel, en appliquant la méthode dite des « *différences de différences* » (encadré 2).

Encadré 2
La méthode des « différences de différences »

L'introduction d'une mesure nouvelle permet d'observer la variable d'intérêt pour quatre sous-populations distinctes. Les individus concernés par la mesure (ou « traités ») sont observés avant et après l'introduction de celle-ci. De la même

[9] En particulier, plusieurs réponses de l'enquête laissaient à penser que les jeunes qui ont été aidés sont plus exigeants dans leur recherche d'emploi. Ils refusent plus fréquemment les emplois qui ne correspondent pas à leurs compétences et concèdent des temps de trajet plus courts pour les emplois en CDD et/ou à temps partiel, que ceux qui n'ont pas été aidés.

manière, les individus « non traités » sont également observés avant et après l'introduction de la mesure. Comparer la situation des « traités » et des « non traités » après l'introduction de la mesure ne permettrait pas d'identifier l'effet causal de cette dernière, car ces deux populations n'ont pas été choisies de façon aléatoire, mais sur la base de critères administratifs ou politiques. L'écart entre ces deux groupes peut donc provenir de différences de caractéristiques individuelles, et pas seulement de l'effet du traitement. De la même manière, la comparaison de la situation des seuls individus « traités » avant et après l'introduction de la mesure est trompeuse, car elle peut résulter de modifications de l'environnement économique sans lien avec cette dernière.

L'évaluateur souhaite donc contrôler ces deux facteurs de biais : les différences de caractéristiques individuelles d'une part, les modifications de l'environnement d'autre part. La solution consiste à définir l'effet causal du traitement comme une « différence de différences » : la différence entre la situation du groupe de traitement après et avant la réforme, moins la différence entre la situation du groupe de contrôle après et avant la réforme.

Avantages

. forte validité interne des résultats (contrôle des caractéristiques observables et inobservables) ;

. estimation aisée au moyen de logiciels statistiques existants.

Limites et points de vigilance méthodologiques

. validité externe potentiellement limitée : difficulté à généraliser les résultats à des publics plus larges ;

. sources de biais possibles : absence de tendance commune entre le groupe de traitement et le groupe de contrôle ; anticipation de la mesure et modification de comportement chez les individus du groupe de traitement.

Contraintes opérationnelles

. nécessite l'existence d'un choc de politique affectant une partie de la population seulement.

Référence méthodologique

. Bertrand M., Duflo E. et Mullainathan S. (2004), "How Much Should We Trust Differences-In- Differences Estimates?", *Quarterly Journal of Economics*, 119.1.

L'intérêt de cette méthode est d'évaluer l'effet causal de la mesure non seulement en contrôlant des différences de caractéristiques entre agents économiques traités et non traités, mais aussi en neutralisant l'effet éventuel de l'évolution de la conjoncture au moment où est mise en place la mesure.

Une illustration de cette méthode est donnée dans l'article de Chemin et Wasmer[10]. Les auteurs cherchent à mesurer l'impact de l'indemnisation maladie sur l'absentéisme des salariés. Ils exploitent le fait qu'en Alsace-Moselle un certain nombre de dispositifs dérogatoires issus du droit germanique ont été conservés après la fin de la Première Guerre mondiale. Le régime de sécurité sociale y est notamment plus généreux : en matière de congé maladie, les employeurs doivent payer l'intégralité du salaire sur toute la durée de l'absence. Dans le reste de la France, depuis 1945, la sécurité sociale couvre seulement 50 % du salaire, et uniquement après le quatrième jour d'absence.

Néanmoins, une réforme de 1978 a obligé les employeurs à couvrir jusqu'à 90 % du salaire après le onzième jour d'absence pour les employés de plus de trois ans d'ancienneté. Si l'on veut connaître l'impact de cette réforme sur le nombre de jours d'absence, il est donc possible d'utiliser l'Alsace-Moselle comme contrefactuel, puisqu'elle n'a pas vu son régime modifié en 1978. En pratique la démarche consiste à mesurer la variation du nombre de jours d'absence après et avant la réforme en Alsace-Lorraine d'une part et dans le reste de la France d'autre part. L'impact de la réforme se mesure ensuite en faisant la différence entre ces deux variations, d'où la notion de « différence de différence »[11]. Les résultats obtenus par Chemin et Wasmer montrent que, alors que l'absentéisme a diminué après la réforme de 1978 en France, l'effet causal de cette politique sur le taux d'absentéisme est paradoxalement positif. Il y a bien eu une augmentation causale, certes faible, mais significative de l'absentéisme du fait de la réforme. Elle était masquée par les tendances globales du marché du travail[12].

Une autre méthode permettant d'établir la causalité statistique est celle de la *régression par discontinuité* (encadré 3). Comme dans le cas des expériences

[10] Chemin et Wasmer (2009).

[11] Il est possible d'estimer cette double différence au moyen d'une régression linéaire, ce qui a l'avantage de donner une indication sur la précision des effets estimés.

[12] La proportion d'individus employés en congé maladie en Alsace-Moselle et avec plus de trois ans d'ancienneté était de 5,11 % entre 1976 et 1978. Entre 1979 et 1989, cette proportion a diminué et est passée à 3,33 %. Cette évolution n'est pas imputable aux changements de législation des congés maladie, inchangée sur la période en Alsace-Moselle, mais aux évolutions du marché du travail, ainsi qu'à d'autres tendances sociétales. Dans le reste de la France, la proportion d'employés en congé maladie était de 4,26 % entre 1976 et 1978, et de 3,36 % entre 1979 et 1989. En Alsace-Moselle, la baisse *tendancielle* de l'absentéisme était donc de 1,78 point de pourcentage. Dans le reste de la France, cette baisse n'a été que de 0,90 point de pourcentage. La législation de 1978 a donc ralenti la baisse tendancielle.

naturelles, il s'agit d'utiliser un autre mécanisme que l'affectation aléatoire, mais susceptible de faire apparaître une situation contrefactuelle à la politique publique étudiée. Dans le cas de la régression par discontinuité, l'appartenance au groupe de traitement ou au groupe de contrôle dépendra d'une variable qui définit un seuil de participation au traitement.

Encadré 3

Les méthodes de régression par discontinuité

Cette méthode exploite le fait que la participation à la politique dépend d'un seuil qui définit la participation au traitement. Elle repose sur l'hypothèse que les individus proches du seuil sont statistiquement comparables et ne diffèrent que du fait de leur participation ou non au traitement. La méthode consiste alors à comparer les résultats des participants et des non participants localisés *au voisinage* du seuil de participation.

Avantages

. forte validité interne des résultats (contrôle des caractéristiques observables et inobservables).

Limites et points de vigilance méthodologiques

. validité externe limitée : les résultats sont valables seulement pour les individus situés à proximité du seuil ;
. source de biais possibles : manipulation possible du seuil par les bénéficiaires (ex : sous-déclaration de revenus pour bénéficier d'une aide).

Contrainte opérationnelle

. existence d'un seuil déterminant la participation au traitement.

Référence méthodologique

. Imbens G. W. et Lemieux T. (2008), Regression Discontinuity Designs: A Guide to Practice, *Journal of Econometrics*, 142(2), p. 615-635.

Imaginons par exemple que l'on cherche à évaluer l'impact sur la réussite universitaire d'une bourse octroyée sur critère de ressources familiales. Par hypothèse, seuls pourraient prétendre à la bourse les étudiants dont les parents ont un revenu inférieur à 15 000 euros par an. Au regard des critères d'attribution, l'étudiant dont les parents gagnent 14 999 euros par an aura des caractéristiques très proches de celui dont les parents gagnent 15 001 euros par an. Pourtant, le

premier aura droit à la bourse, tandis que le second devra y renoncer. Comparer les résultats aux examens des étudiants se trouvant juste en dessous du seuil d'éligibilité et ceux se situant juste au-dessus du seuil permettrait ici d'identifier l'effet causal de la bourse sur les performances universitaires. C'est cette méthode originale que quatre universitaires italiens[13] ont utilisée pour conclure que les droits de scolarité augmentaient les performances des étudiants, sans que ces différences de performance ne proviennent d'écarts socioéconomiques entre ces étudiants[14].

Expériences naturelles et régressions par discontinuité constituent des cas particuliers dans lesquels une variable sans lien direct avec la variable de résultat influence la participation au traitement (un choc de politique économique dans le premier cas, une variable déterminant le seuil de participation dans le second). De façon plus large, il est possible d'établir la causalité statistique dès lors qu'existe une variable ayant un effet sur la participation au traitement, mais pas d'effet sur la variable de résultat. On parle dans ce cas de *variable instrumentale* ou simplement d'instrument (encadré 4).

Denis Fougère, Francis Kramarz et Julien Pouget sont par exemple parvenus à identifier à l'échelle départementale la relation de causalité entre chômage des jeunes et criminalité au moyen de cette méthode[15]. Établir une telle relation est difficile, car il est possible que dans les départements où existe un fort taux de criminalité, les entreprises soient dissuadées d'investir, ce qui contribue à y dégrader la situation du marché du travail.

Encadré 4

Les méthodes de variables instrumentales

Cette méthode exploite une variable qui est statistiquement corrélée à la variable de traitement, mais non corrélée à celle de résultat. Lorsque c'est le cas, les variations du traitement induites par l'instrument permettent d'identifier l'effet du traitement sur la variable de résultat. La démarche d'évaluation consiste alors à régresser dans un premier temps la variable de traitement sur l'instrument puis à

[13] Voir Garibaldi *et al.* (2012).

[14] Pour la France, une étude de Thomas Piketty et Mathieu Valdenaire (2006) exploite les discontinuités liées au franchissement des seuils d'ouverture et de fermeture des classes, et montre que la réduction de la taille des classes tend à améliorer les résultats scolaires des élèves de CE1 et CE2.

[15] Voir Fougère *et al.* (2009).

utiliser l'estimateur obtenu dans l'équation reliant la variable de résultat au traitement.

Avantages

. forte validité interne des résultats (contrôle des caractéristiques observables et inobservables) ;
. estimation aisée au moyen des logiciels statistiques existants.

Limites et points de vigilance méthodologiques

. validité externe limitée : les résultats sont valables seulement pour les individus dont la probabilité d'être traité est influencée par l'instrument ;
. difficulté à justifier la condition d'exclusion (i.e. l'absence d'effet direct de l'instrument sur la variable de résultat).

Contrainte opérationnelle

. existence d'une ou de plusieurs variables corrélées à la variable de traitement mais pas à la variable de résultat.

Référence méthodologique

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

Dans ce cas, c'est bien la criminalité qui est à l'origine du chômage et non l'inverse. En d'autres termes, l'existence d'une simple corrélation positive entre les deux variables ne permet pas de conclure que le chômage cause effectivement les comportements délictuels. Pour contrôler le caractère potentiellement endogène de la criminalité, les auteurs utilisent comme instrument la répartition sectorielle de l'emploi dans chaque département. Cette variable explique correctement le taux de chômage des jeunes, car les départements où les activités de services sont importantes donnent plus de perspectives d'embauche à ces derniers, contrairement aux départements où l'industrie est dominante. La répartition sectorielle de l'emploi n'a pas d'impact direct sur la criminalité. En revanche, elle constitue une source de variation exogène du taux de chômage, qui permet d'identifier l'effet causal de ce dernier sur la criminalité. Fougère *et al.* montrent que cet effet est très significatif. Sur des données portant sur la période 1990-2000, leurs résultats indiquent qu'une variation du taux de chômage des jeunes de cinq points de pourcentage multiplie par deux le taux de vols avec violence dans le département considéré.

Lorsqu'il s'agit d'évaluer l'impact d'un dispositif sur ses bénéficiaires, les méthodes précédentes présentent un intérêt important dans la mesure où elles contrôlent les biais de sélection provenant tout à la fois de caractéristiques individuelles

observables et inobservables. Cependant les expérimentations aléatoires sont parfois difficiles à mettre en œuvre en pratique, et les données statistiques disponibles par ailleurs ne proposent pas toujours de sources de variation exogène dans la participation au traitement. Dans cette situation, les évaluateurs se tournent vers des méthodes moins ambitieuses, en ce qu'elles ne contrôlent que les biais provenant de caractéristiques observables, c'est-à-dire accessibles dans les bases de données existantes. C'est le cas de la méthode dite de l'appariement statistique qui consiste à associer à chaque individu traité, un ou des individus non traités présentant des caractéristiques similaires et susceptibles d'avoir une influence à la fois sur la participation au traitement et sur la variable de résultat (voir encadré 5).

Ainsi, dans l'évaluation d'un programme de formation pour les chômeurs on associera à un individu formé un ou plusieurs individu(s) ayant une probabilité similaire – aussi appelée score de propension – de passer par la formation au regard de ses (leurs) caractéristiques (l'âge, le diplôme, l'ancienneté au chômage par exemple), mais ne bénéficiant pas de la formation[16].

Encadré 5

Les méthodes d'appariement

Cette méthode consiste à constituer des « paires » entre des individus traités et non traités présentant des caractéristiques observables similaires et susceptibles d'influencer conjointement la participation au traitement et la variable de résultat (l'âge ou le diplôme dans le cas d'une évaluation de l'impact de la formation sur les salaires, par exemple). En pratique, l'appariement suppose d'abord d'estimer la probabilité de traitement de chaque individu conditionnellement à ses caractéristiques individuelles.

À cette fin on régresse la variable de traitement sur ces caractéristiques, puis on calcule pour chaque individu, traité ou non, une probabilité conditionnelle de traitement, aussi appelée score de propension. L'appariement consiste ensuite à

[16] Fitzenberger et Speckesser (2008) évaluent l'impact de la formation sur le taux d'emploi des chômeurs et de salariés menacés de perte d'emploi en Allemagne. En utilisant la méthode de l'appariement par le score de propension, ils trouvent que le taux d'emploi du groupe de contrôle est supérieur à celui des individus formés dans les semaines qui suivent le début du programme. Ceci traduit un effet d' « enfermement » (« lock-in effect ») : durant la période de formation, l'intensité de la recherche d'emploi diminue, et le taux de retour à l'emploi également. Pour une évaluation plus récente des contrats aidés en France au moyen des techniques d'appariement, voir : Benoteau (2015).

associer à chaque individu traité un ou plusieurs individus non traités ayant un score de propension similaire. Le groupe des non traités ainsi constitué représente la situation contrefactuelle, et l'effet causal se définit comme la différence entre les moyennes de la variable de résultat dans le groupe des traités et dans ce groupe des « jumeaux ».

Avantages

. possibilité d'estimer l'effet d'un traitement pour des sous-populations (les jeunes, les moins qualifiés par exemple), afin d'améliorer le ciblage des politiques ;
. estimation aisée au moyen des logiciels statistiques existants.

Limites et points de vigilance méthodologiques

. validité interne limitée : les résultats sont valables seulement s'il n'existe pas de variables inobservables susceptibles d'affecter conjointement la participation au traitement et la variable de résultat[17] ;
. nécessité d'un « support commun », c'est-à-dire de valeurs du score de propension similaires entre individus traités et non traités.

Contraintes opérationnelles

. faibles dès lors que les données individuelles sont suffisamment riches pour contrôler la sélectivité.

Référence méthodologique

. Caliendo M. et Kopeinig S. (2008), "Some Practical Guidance for the Implementation of Propensity Score Matching", *Journal of Economic Surveys*, 22(1), p. 31-72.

Le groupe des individus non traités présentant un score similaire à celui des traités constitue alors le contrefactuel. L'effet causal de la mesure s'obtient alors en calculant simplement la différence entre la moyenne de la variable d'intérêt – ici, le taux de retour à l'emploi des chômeurs – dans le groupe de traitement et dans ce groupe de contrôle.

Quelle méthode choisir ?

Les méthodes précédentes présentent toutes des avantages et des limites, et l'arbitrage dépend de plusieurs facteurs : la nécessité de réduire les biais, les contraintes opérationnelles et la disponibilité des données permettant l'évaluation.

[17] La difficulté tient ici au fait que cette hypothèse n'est pas testable empiriquement, sauf si l'on dispose de données expérimentales permettant de comparer les résultats issus de l'expérience aléatoire et de l'appariement. Dans certains cas, ces résultats sont très voisins comme l'ont montré Dehejia et Wahba (1999) en évaluant l'effet de programmes de formation aux États-Unis.

Le critère essentiel reste l'obtention de résultats décrivant le véritable effet causal de la mesure évaluée, ce qui invite à privilégier les méthodes permettant de contrôler la sélectivité provenant à la fois de facteurs observables et inobservables. Dans ce cadre une bonne démarche pour le choix de la méthode d'évaluation pourrait être la suivante[18].

✓ Appliquer les expériences aléatoires contrôlées (EAC) lorsque c'est possible, et si possible en amont de la mise en œuvre de la politique.

✓ Si les contraintes liées aux EAC sont trop importantes et si les données le permettent, appliquer dans l'ordre de priorité suivant (c'est-à-dire en tenant compte de la portée plus ou moins générale des résultats) :

- les différences de différences (résultats valables pour l'ensemble de la population soumise au traitement) ;
- les variables instrumentales (résultats valables seulement pour les individus dont le traitement est sensible à l'instrument) ;
- les régressions par discontinuité (résultats valables seulement pour les individus à proximité du seuil déterminant la participation au traitement).

Si aucune de ces méthodes n'est applicable, appliquer les méthodes d'appariement, qui présentent l'inconvénient de ne contrôler qu'une partie des biais de sélection (ceux résultant de caractéristiques observables par l'économètre).

3. Le ciblage des politiques et des ressources publiques

Bien souvent, les évaluations disponibles se contentent d'étudier les effets de dispositifs en moyenne sur une population donnée. Or, dans certains cas, les politiques peuvent avoir des effets très hétérogènes selon les catégories de la population concernée, les territoires ou l'orientation de la conjoncture. Pour évaluer l'effet d'une politique publique, il est souvent pertinent d'aller au-delà des effets moyens de celle-ci. Il peut ainsi être socialement souhaitable de mettre en œuvre une politique éducative qui permette de réduire la proportion d'élèves en grande difficulté, même si elle n'a qu'un effet négligeable sur le niveau moyen de l'ensemble des élèves. Ceci milite en faveur d'évaluations proposant des résultats différenciés suivant le type d'agents économiques que l'on considère.

[18] Cette démarche est inspirée de Givord, *op. cit*.

C'est ce que montrent par exemple Bonnal, Fougère et Sérandon[19] qui évaluent différentes mesures d'emploi destinées aux jeunes, telles que les contrats en alternance, les travaux d'utilité collective (TUC) et d'autres stages de formation. Ces auteurs mettent en évidence que les programmes qui allient formation et emploi marchand permettent une meilleure insertion à court terme, mais que cet impact est plus fort chez les individus peu ou pas diplômés, et peu significatif chez les diplômés.

Certaines évaluations illustrent que des dispositifs d'emploi peuvent même s'avérer contreproductifs pour certains individus, alors qu'ils apportent un bénéfice à d'autres. Ainsi, il a été montré que l'accompagnement proposé par le service public de l'emploi à des chômeurs autonomes dans leur recherche pouvait nuire à ces derniers. En les obligeant à substituer des actes formels, comme la rencontre avec des conseillers, à une recherche basée sur l'exploitation de leurs propres réseaux, l'accompagnement du service public de l'emploi (SPE) contribue à réduire le taux de retour à l'emploi de ce type de chômeurs par rapport à la situation contrefactuelle[20].

De façon plus large, il est souhaitable de mener des évaluations en stratifiant les résultats suivant des dimensions territoriale ou temporelle. Un dispositif d'aide à l'accession à la propriété aura par exemple un effet plus faible dans un département urbain où les possibilités de construction seront limitées. L'impact d'un programme de formation professionnelle pour les chômeurs pourra ainsi varier dans le temps, en ayant un impact négatif à court terme – lié au fait que les individus en formation cherchent moins activement en emploi, mais positif à plus long terme – en permettant de retrouver des emplois plus stables[21].

L'hétérogénéité des effets d'une politique publique

Le développement de techniques économétriques, comme la régression par quantiles, permet d'estimer l'impact d'un dispositif selon les catégories de la population concernée. Le principe de cette méthode est de fournir une estimation de l'impact d'une variable explicative sur l'ensemble des quantiles de la distribution de la variable d'intérêt. Ceci permet d'aller plus loin que l'estimation d'un simple effet moyen qui peut, pour certains individus, s'avérer trompeur. Il est par exemple possible d'évaluer une mesure en distinguant son effet sur les individus à haut et bas revenus.

[19] Bonnal *et al*. (1997).
[20] Van den Berg *et al.* (2004).
[21] Voir sur données françaises, Crépon *et al*. (2012).

Dans une introduction pratique aux régressions quantiles Givord et d'Haultfoeuille (2014)[22] appliquent cette méthode à un programme de formation pour les chômeurs appelé le « Job Training Partnership Act (JTPA) », mis en place à partir de 1983 aux États-Unis. Il s'agit d'un ensemble de programmes de formation et d'assistance destinés aux jeunes défavorisés, pour lequel un processus d'expérimentation aléatoire permet de contrôler les biais de sélection à l'entrée du programme[23]. Les régressions quantiles menées pour évaluer l'impact du JTPA indiquent que la moyenne masque de grandes disparités dans l'effet du programme. Pour les femmes, l'effet moyen du traitement, c'est-à-dire la différence entre la moyenne espérée des revenus en l'absence du programme et celle des revenus avec le programme, est de 1 825 dollars. Mais il n'augmente en fait que de 390 dollars le premier quartile de la distribution de revenus, alors que l'augmentation atteint 2 800 dollars pour le dernier quartile. Ceci suggère que l'impact du programme est plus important en valeur absolue pour les individus dont les revenus sont initialement les plus élevés.

On le voit, les méthodes d'évaluation non structurelles offrent une vaste gamme d'outils permettant d'évaluer l'impact causal d'un dispositif sur ses bénéficiaires. Toutefois ces méthodes présentent des limites : la difficulté à interpréter les résultats d'abord, car ces méthodes ne donnent en général pas d'information sur les mécanismes par lesquels la politique produit ses effets ; la difficulté à étendre les résultats obtenus à une population plus large ensuite. Pour ces raisons, il est souvent utile de compléter l'analyse par une démarche d'évaluation structurelle.

[22] d'Haultfœuille et Givord (2014).

[23] L'évaluation de l'efficacité de ce type de programme est souvent rendue difficile par les effets d'auto-sélection : en général, ce sont les personnes qui peuvent en retirer le plus grand bénéfice qui choisissent de rentrer dans le dispositif. Une expérimentation a cependant été mise en place entre 1987 et 1989 dans 16 structures locales auprès d'un échantillon initial de 20 000 jeunes environ. Les programmes de formation correspondant au JTPA n'ont été proposés qu'à deux tiers de ces jeunes, tirés aléatoirement. Des données ont ensuite été collectées sur la séquence de revenus de l'ensemble des jeunes de l'échantillon initial.

CHAPITRE 2

GÉNÉRALISER LES RÉSULTATS DE L'ÉVALUATION

Les méthodes présentées à la section précédente permettent des évaluations *ex-post,* qui mesurent l'impact de politiques déjà mises en œuvre sur les agents qui sont les bénéficiaires directs de ces politiques. Les chercheurs les qualifient de méthodes non structurelles, dans la mesure où elles s'appuient uniquement sur la variabilité contenue dans les données pour établir la causalité statistique, sans faire d'hypothèse sur les mécanismes économiques par lesquels la politique produit ses effets.

La principale limite de l'ensemble de ces méthodes réside dans la difficulté à généraliser leurs résultats à une population plus vaste, dans le cas où la politique serait étendue. En effet une telle extension peut mener à changer les effets estimés de la politique, pour plusieurs raisons. D'une part, parce que la population à laquelle la politique est étendue peut avoir une sensibilité différente au dispositif évalué ; d'autre part, parce que le changement d'échelle de ce dispositif peut engendrer des effets indirects susceptibles de modifier l'impact que la politique a sur l'ensemble de l'économie.

Par conséquent, on s'intéresse maintenant à l'évaluation d'une politique publique qui serait étendue à des populations non concernées jusqu'alors. La question qui se pose est alors la capacité à prévoir correctement les effets d'une politique donnée dans un contexte différent. Cette démarche nécessite souvent le recours à une approche structurelle[24], c'est-à-dire à une représentation explicite de l'économie fondée sur des mécanismes bien identifiés par le chercheur et matérialisés par des paramètres structurels invariants aux politiques économiques mises en œuvre.

[24] Pour une comparaison approfondie des approches structurelle et non structurelle en matière d'évaluation le lecteur se reportera avec profit à Roux (2015).

1. Des effets sur les non-bénéficiaires

L'extension d'une politique à une population plus large peut produire des effets différents de ceux évalués sur la seule population des bénéficiaires concernés jusqu'alors en raison de l'hétérogénéité des effets selon les publics (cf. *infra*). Mais au-delà de cette hétérogénéité, l'extension de la politique peut avoir des effets indirects sur les non-bénéficiaires au travers de certaines externalités qui ne se manifestent pas à petite échelle lorsqu'un petit nombre de bénéficiaires est concerné. Les effets de « déplacement » en sont un exemple, et ont pu être mis en évidence dans le cadre de l'évaluation de dispositifs d'accompagnement renforcé pour les demandeurs d'emploi. Lorsque les offres d'emploi sont en nombre limité ce type de dispositif peut en effet améliorer le taux d'emploi de ses bénéficiaires au détriment des autres chômeurs, qui se voient évincés dans la file d'attente pour les postes. Crépon *et al.*[25] parviennent à ce résultat en faisant varier de façon aléatoire la proportion de chômeurs bénéficiant de l'accompagnement renforcé d'une agence d'emploi à l'autre[26]. Ils trouvent que lorsque cette proportion est forte les chômeurs bénéficiaires d'un accompagnement classique au sein de la même agence ont un taux de retour à l'emploi inférieur, toutes choses égales par ailleurs, à ceux qui bénéficient du même accompagnement dans des agences où l'accompagnement renforcé est inexistant. Ceci suggère que l'accompagnement renforcé peut avoir des effets indirects négatifs sur les non-bénéficiaires. La conséquence en est que l'impact de cette mesure évalué sur l'ensemble de la population serait différent de celui évalué sur les seuls bénéficiaires[27].

2. Des effets différents selon la population couverte

D'autres mécanismes peuvent aboutir à ce que la généralisation d'une politique produise des effets différents de ceux évalués à petite échelle. C'est le cas par exemple lorsque la politique engendre une modification de l'offre ou de la demande sur un marché donné. Une illustration de ce phénomène est donnée par le

[25] Crépon *et al*. (2013).

[26] En pratique la proportion d'individus traités était de 0 %, 25 %, 50 %, 75 % ou 100 % selon les agences.

[27] L'étude de Crépon *et al*. (2013) est remarquable dans la mesure où elle s'appuie sur une démarche d'évaluation non structurelle, c'est-à-dire ne reposant pas sur une représentation théorique de l'économie, mais sur une démarche d'expérimentation aléatoire originale. Dans la majorité des cas, néanmoins, la mise en évidence des effets indirects d'une politique publique réclame de recourir à un modèle structurel.

développement du programme Medicare aux États-Unis. La question est ici de savoir si une politique d'assurance santé à grande échelle telle que Medicare est susceptible d'avoir un effet différent sur les dépenses de santé de celui d'un dispositif d'assurance au périmètre plus modeste. Ainsi ont été lancées au début des années 1970 des expérimentations aléatoires visant à évaluer l'effet sur les dépenses de santé de l'introduction d'un dispositif d'assurance. Ces évaluations, plus connues sous le nom de *Rand Health Insurance Experiment* (RHIE), montrent que l'impact du dispositif sur les dépenses de santé est relativement faible. Toutefois, les bénéficiaires du RHIE constituaient une population restreinte, de sorte que l'expérimentation ne permettait pas de mesurer les effets d'une réforme à plus grande échelle, telle que l'introduction de Medicare.

L'économiste Amy Finkelstein a de son côté évalué l'effet de l'introduction du Medicare sur les dépenses de santé à l'échelle macroéconomique[28]. Elle montre que cette réforme est associée à une augmentation des dépenses de santé environ six fois plus importante que celle mesurée dans le cadre restreint du RHIE. Selon Finkelstein, l'explication de cette différence tient au fait que, par son ampleur, Medicare a pu faire évoluer les pratiques médicales à l'échelle du pays. En particulier l'afflux de ressources financières nouvelles a engendré la construction d'hôpitaux et a suscité des efforts accrus de recherche médicale. Cette amélioration de l'offre de soins a contribué à stimuler les dépenses de santé d'individus qui n'étaient pas bénéficiaires de Medicare et que ne permettait pas de mesurer une expérimentation aléatoire telle que le RHIE.

Un autre exemple de l'importance que peuvent avoir les effets indirects d'une politique publique est donné par l'article d'Heckman, Lochner et Taber[29]. À l'aide de leur modèle, Heckman *et. al.* évaluent l'effet sur le niveau de la scolarisation et sur les inégalités de salaires d'une subvention forfaitaire de 500 dollars pour tous les élèves au niveau du collège. Ils mesurent d'abord l'effet de cette aide dans un cadre microéconomique, sans tenir compte des effets d'équilibre par lesquels la mesure est susceptible d'influencer le fonctionnement du marché du travail. Ils trouvent ainsi que la subvention accroît le niveau de la scolarisation de 5,3 % en moyenne. À l'équilibre général en revanche, l'effet estimé se réduit à 0,46 %. Dans ce cadre, en effet, la hausse de la scolarisation conduirait à une hausse massive du nombre de diplômés sur le marché du travail. Cette hausse ne peut pas rester sans effet : les auteurs estiment qu'elle produirait une baisse des salaires des diplômés relativement aux

[28] Finkelstein (2005).
[29] Heckman *et al.* (1998).

non-diplômés, en raison de la moindre rareté des diplômés. Cette baisse du rendement salarial de l'éducation conduirait alors les agents à ajuster leur comportement, en s'éduquant moins. Au moyen de leur modèle structurel, les auteurs montrent que cette réponse globale contribue à atténuer fortement l'effet partiel de la mesure, dans lequel les effets d'équilibre ne sont pas pris en compte. Cet exemple montre que l'introduction d'un dispositif peut engendrer des phénomènes parfois difficiles à anticiper. De façon plus générale, cette étude illustre l'intérêt de prendre en compte les effets d'équilibre général dans toute évaluation d'une politique à grande échelle.

Lorsque cela est possible, il est ainsi très utile d'associer à l'évaluation portant sur les bénéficiaires d'une politique une démarche visant à mesurer les effets de l'extension de cette dernière.

C'est ce que font Olivier Bargain et Karina Doorley[30], en analysant l'impact potentiel d'une extension du RSA aux jeunes de moins de 25 ans. L'originalité de leur approche consiste à exploiter la discontinuité que produit la règle qui réserve l'accès du Revenu Minimum d'Insertion (RMI) aux jeunes de plus de 25 ans. Cette discontinuité leur permet d'identifier un paramètre de comportement : la réaction de l'offre de travail des individus à l'existence d'une allocation sociale telle que le RSA-socle, qui a pris en 2009 la suite du RMI. L'identification de ce paramètre de comportement leur permet ensuite de simuler au moyen d'un modèle structurel l'effet de l'extension du RSA à la population des moins de 25 ans. Leurs résultats indiquent que le fait de devenir éligible au RMI conduit à une baisse du taux d'emploi de l'ordre de 5 % à 9 %. En revanche l'introduction du RSA-activité qui, contrairement au RMI, comporte une partie d'aide conditionnée à l'activité, contribue à restaurer une certaine incitation à l'emploi. Bargain et Doorley concluent que l'extension du RSA à la population des moins de 25 ans ne contribuerait pas à réduire les incitations au travail.

Les cas cités ci-dessus illustrent que les effets indirects associés à l'extension d'une politique ne doivent pas être négligés dans le processus d'évaluation, sous peine de commettre des erreurs coûteuses, aussi bien pour les citoyens que pour les finances publiques. À cet égard, les approches non structurelles développées *supra* présentent des limites car, en règle générale, elles ne permettent d'estimer l'impact d'un dispositif que sur les bénéficiaires effectifs de ce dispositif..

[30] Bargain et Doorley (2013).

CHAPITRE 3

EXPLIQUER L'ÉCHEC OU LE SUCCÈS D'UNE POLITIQUE

Les évaluations quantitatives empiriques du type de celles qui sont décrites ci-dessus permettent d'évaluer l'impact d'un dispositif sur un ou plusieurs critères mais pas les raisons de son échec ou de son succès, ce qui est pourtant essentiel pour juger notamment de la reproductibilité dans un autre contexte. Le recours à d'autres approches est particulièrement utile pour compléter les résultats. Nous présentons ici le principe, les forces et les faiblesses des approches structurelles et qualitatives, comparativement à l'approche athéorique.

1. L'approche structurelle : décomposer les effets et les quantifier

Les débats sont vifs dans la communauté scientifique entre les partisans de l'approche structurelle et ceux de l'approche non structurelle pour l'évaluation des politiques publiques. Ils ont été alimentés par plusieurs contributions récentes. Certains économistes cherchent néanmoins à articuler ces deux approches en montrant comment elles se complètent, à l'instar de Sébastien Roux dans le cas français[31] [32]. En raison de sa forte technicité, l'approche structurelle est mise en œuvre essentiellement par le monde académique.

Le principe : modéliser les mécanismes économiques

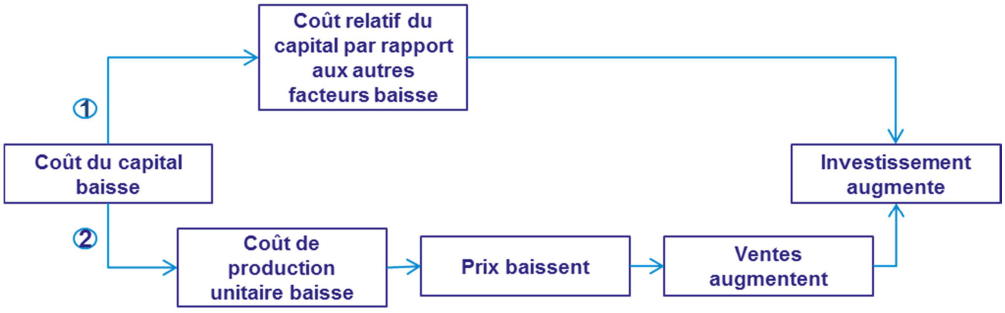
L'approche structurelle propose une représentation de la réalité économique, fondée sur une idée *a priori* des mécanismes économiques à l'œuvre (stratégies des acteurs, imperfections de marchés, canaux de transmission de la politique publique

[31] Voir Roux, *op. cit*.
[32] Voir aussi Heckman (2010).

par lesquels se manifestent les effets). Elle explicite les comportements d'un ou plusieurs agents économiques (entreprises, ménages, etc.) intervenant sur un ou plusieurs marchés (marchés des biens, du travail, du capital, etc.). Ces comportements décrivent comment les agents économiques réagissent à des modifications qui se produisent dans leur environnement, en modifiant les offres ou les demandes qu'ils expriment sur les différents marchés. Ils peuvent avoir des fondements microéconomiques ou macro-économiques.

À titre d'exemple, dans un modèle microéconomique structurel de demande de facteurs de production, le coût du capital agit au niveau individuel sur la demande de capital *via* deux canaux (schéma 3). Le premier canal de transmission correspond à une modification de la combinaison productive qui conduit les entreprises à substituer le capital aux autres facteurs de production quand le coût du capital baisse. La production devient plus intensive en capital : cette hausse est d'autant plus forte que les possibilités de substitution entre les facteurs de production sont importantes *(effet de substitution)*. Le deuxième canal par lequel une baisse du coût du capital agit sur la demande de capital transite par une baisse du coût de production unitaire qui, si elle est répercutée dans les prix, se traduit par une hausse de la demande adressée à l'entreprise. Cette dernière est d'autant plus forte que la demande est sensible aux prix *(effet de profitabilité ou de volume)*.

Schéma 3
Mécanismes économiques attendus d'une baisse du coût du capital



Source : Les auteurs

En définitive, dans ce modèle élémentaire en équilibre partiel, une baisse du coût du capital induit une hausse de la demande de capital et de la demande adressée à l'entreprise, et elle a un effet à la hausse ou à la baisse sur les autres facteurs de

production, selon les possibilités de substitution entre les facteurs de production et la sensibilité aux prix de la demande adressée aux entreprises[33].

Ces « relations de comportement » sont ensuite estimées par des techniques économétriques à partir de données statistiques, individuelles ou agrégées[34]. L'estimation permet de connaître la valeur des paramètres dits structurels (technologies, préférences, dotations et structures de marché), supposés invariants à des modifications de politiques publiques. Une fois estimées, ces relations de comportement permettent de construire les situations « contrefactuelles » que l'on aurait observées sans la politique ou avec une autre politique (encadré 6).

Encadré 6

Les six étapes de l'approche structurelle

L'approche structurelle peut se décomposer schématiquement en six étapes :

1er étape : identifier les principaux mécanismes économiques à l'œuvre.

2eétape : recenser les données disponibles pour l'estimation du modèle. Le choix dépendra mécanismes considérés.

3e étape : spécifier le modèle économique, explicitant les relations de comportement et le processus générateur des données (hypothèses distributionnelles des aléas).

4e étape : estimer les paramètres structurels du modèle à partir de données microéconomiques (au niveau des individus ou des entreprises) ou macroéconomiques. Certains paramètres peuvent aussi être « calibrés » (c'est-à-dire se voir attribuer une valeur numérique fixe) à partir de résultats généralement issus d'approches non structurelles disponibles dans la littérature économique.

5e étape : vérifier que les estimations obtenues sont crédibles et robustes.

6e étape : utiliser le modèle pour réaliser des exercices contrefactuels correspondant à des simulations que l'on aurait observées en l'absence de politique publique ou avec des politiques publiques alternatives.

[33] Voir par exemple Crépon et Gianella (2001) pour la spécification et l'estimation de ce modèle sur données d'entreprises.

[34] Ces estimations sont généralement réalisées par la méthode du maximum de vraisemblance ou par celle des moments généralisés (GMM).

Dans l'exemple précédent, le modèle microéconomique de demandeurs de facteurs de production repose sur quatre paramètres clés que l'économètre va chercher à estimer : l'élasticité de substitution entre les facteurs de production, l'élasticité aux prix de la demande de biens, les rendements d'échelle et le taux de marge (voir ci-dessous les hypothèses du modèle). Les valeurs estimées pourront servir à évaluer les effets sur l'emploi, l'investissement et la production d'une réforme de la fiscalité pesant sur les entreprises ou les ménages. Ces effets seront obtenus en faisant varier le paramètre de la politique que l'on étudie (par exemple, le taux d'impôt sur les sociétés ou encore le taux de cotisations sociales employeurs).

Dans l'approche non structurelle, le modèle théorique permettant d'obtenir les équations estimées n'est pas explicité (il s'agit de « formes réduites »). Les paramètres estimés ne peuvent donc pas être interprétés. En reprenant l'exemple précédent, il ne serait pas possible, à l'aide de l'approche athéorique, de décomposer les effets obtenus en un effet de substitution et un effet de volume.

La force de cette approche : comprendre les mécanismes économiques à l'œuvre

L'intérêt de l'approche structurelle est, au-delà de la seule mesure des effets, de comprendre les causes des effets de changements de politiques. En effet, l'approche structurelle permet d'interpréter les résultats en considérant les mécanismes économiques à l'œuvre, à la différence de l'approche athéorique qui se concentre sur l'identification des effets.

L'approche structurelle a deux autres avantages par rapport à l'approche non structurelle : elle est utile pour prévoir les impacts de politiques publiques déjà mises en œuvre dans des contextes donnés, dans d'autres contextes ; elle est également utile pour prévoir les impacts de politiques publiques encore jamais mises en œuvre. L'approche non structurelle seule n'est pas appropriée dans ces deux cas de figure car comme nous l'avons vu, elle s'appuie sur la variabilité contenue dans les données pour évaluer les effets d'une politique publique dans un environnement précis[35] (cf. *infra*).

[35] L'approche non structurelle d'évaluation d'impact peut toutefois contribuer à l'évaluation *ex ante*. Les résultats obtenus par l'approche athéorique pour un dispositif déjà appliqué peuvent être utilisés pour l'évaluation *ex ante* d'un dispositif similaire, de manière directe ou indirecte (pour calibrer les paramètres d'un modèle structurel).

Enfin, l'approche structurelle permet, au moins en théorie, de modéliser l'ensemble des distorsions que connaissent les différents marchés, mais aussi l'ensemble des réponses déjà existantes à ces distorsions. Prenons l'exemple d'un marché du logement qui connaîtrait une demande limitée par le rationnement du crédit dont font preuve les banques à l'égard des ménages modestes. Face à cette situation, le gouvernement décide de mettre en place une aide à l'accession à la propriété, sous forme d'un prêt à taux zéro pour lesdits ménages. Cette aide s'avère avoir a des effets non anticipés, notamment une hausse des prix de vente artificiellement entretenue par le soutien à la demande que constitue l'aide publique. Dans ce contexte, une autre mesure est alors mise en œuvre, que des chercheurs souhaitent évaluer : un dispositif d'assurance par lequel l'État garantit les loyers des locataires impécunieux, afin de lever les réticences des propriétaires à louer leur bien. À l'évidence, l'évaluation de l'impact de cette garantie sur l'accès au logement doit tenir compte de l'existence du premier dispositif, le prêt à taux zéro. Celui-ci, en élevant potentiellement les prix des biens, peut en effet contraindre certains ménages à louer plutôt qu'à acheter. L'évaluation de la garantie des loyers doit dans ce cas se faire dans un contexte où la demande pour la location est plus importante que si le prêt à taux zéro n'existait pas.

Cet exemple montre la difficulté à évaluer des réformes globales, comportant pour un même champ ministériel plusieurs mesures concomitantes, et plus encore la difficulté à évaluer des réformes visant à agir conjointement sur plusieurs marchés[36]. L'évaluation doit au maximum tenir compte de ces effets de contexte engendrés par la complexité de l'action politique. En pratique, il n'est généralement pas possible de le faire en envisageant tous les dispositifs existants. Mais l'approche structurelle fournit néanmoins un cadre particulièrement utile lorsque l'on cherche à évaluer une réforme globale qui consiste à substituer certains dispositifs à d'autres (par exemple, le remplacement du Crédit Impôt Compétitivité Emploi par un allègement de cotisations sociales).

Sa faiblesse : la sensibilité aux hypothèses du modèle théorique

Les modèles structurels sont élaborés à partir de nombreuses hypothèses sur les comportements présupposés des agents ou sur des éléments plus techniques relatifs notamment aux formes fonctionnelles reliant les différentes variables entre elles (supposées linéaires le plus souvent) ou aux distributions des aléas du modèle

[36] La loi Croissance et activité votée en 2015 qui comportait des dispositions relatives aux marchés du travail, de l'immobilier, des biens et services, illustre ce point.

(supposés tirés dans des lois normales le plus souvent). Pour illustrer la nature de ces hypothèses, on peut repartir de l'exemple donné précédemment. Comme nous l'avons vu, ce modèle élémentaire considère deux canaux par lesquels le prix d'un facteur de production se transmet aux demandes de facteurs (effets de substitution et de volume). Par ailleurs, sa définition requiert de faire les hypothèses que la technologie de production est à élasticité de substitution constante entre les facteurs de production et à rendements d'échelle constants, que la demande adressée à l'entreprise a une élasticité par rapport aux prix de vente constante et que les entreprises appliquent un taux de marge constant sur leur coût de production. Ces hypothèses contraignent fortement et fragilisent de la sorte les estimations des paramètres. Si elles ne sont pas vérifiées, en particulier si les mécanismes économiques considérés ne sont pas les bons ou ne sont pas les plus importants, « le risque est de se tromper lourdement dans l'évaluation » (voir Sébastien Roux, 2015).

Les méthodes structurelles sont également plus complexes à appréhender pour un décideur public. Les modèles structurels sont en réalité extrêmement variés, et peuvent être plus ou moins élaborés. La simplicité se fait cependant au prix d'une représentation assez « limitée » du fonctionnement de l'économie qui laisse de côté certains mécanismes économiques. Dès lors que l'on souhaite mieux appréhender la réalité économique, on est conduit à rajouter nombre d'hypothèses qui complexifient la structure des modèles, les rendent plus difficiles à estimer et accroissent la fragilité des résultats.

Enfin, certains économètres structurels portent parfois une attention insuffisante à l'information contenue dans les données, en particulier à leur variabilité qui doit être suffisante pour permettre l'identification des paramètres. L'économètre structurel ne doit pas seulement se préoccuper de la spécification de son modèle mais aussi s'interroger sur les questions d'identification (et notamment rechercher des sources exogènes de variation dans les données).

En résumé, l'approche structurelle requiert de s'interroger à la fois sur la pertinence du modèle structurel qu'elle utilise et sur la robustesse des résultats.

Une illustration : l'évaluation d'une mesure fiscale aux Pays-Bas

Lokshin et Mohnen[37] évaluent les effets d'une mesure fiscale en faveur de la R & D aux Pays-Bas à partir de données individuelles d'entreprises. Créé en 1994, le dispositif appelé WBSO[38] est une réduction fiscale s'appliquant aux coûts salariaux liés aux activités de recherche et développement menées dans l'entreprise.

Les auteurs spécifient un modèle microéconomique structurel de comportement d'investissement en R & D en fonction notamment d'un coût d'usage du capital en R & D de l'entreprise[39], qui dépend lui-même des caractéristiques du dispositif WBSO. Ils en estiment les paramètres de long terme et la dynamique d'ajustement sur un échantillon d'entreprises présentes sur la période 1996-2004 (suivant la mise en place du dispositif). L'évaluation *ex post* de l'impact du dispositif est ensuite obtenue en simulant les effets sur la R & D, sans le dispositif.

Les auteurs utilisent un modèle élémentaire de demande de facteurs de production où, à l'équilibre de long terme, la productivité marginale du capital de R & D est égale au coût d'usage du capital de recherche[40]. La demande de capital d'équilibre dépend alors notamment du coût d'usage du capital et de la production de l'entreprise. Le coût du capital affecte négativement le stock de capital optimal avec une élasticité qui est égale à l'élasticité de substitution entre ce capital et les autres facteurs de production (un indice composite). L'effet du coût du capital sur la R & D est donc d'autant plus faible que la substitution entre le capital de recherche et les autres facteurs de production est faible. La production de l'entreprise a également un effet positif sur le stock de capital de R & D.

---

[37] Lokshin et Mohnen (2012).

[38] En 2004, le taux de crédit d'impôt est de 42 % pour des coûts salariaux liés à la R & D en deçà de 110 millions d'euros et de 14 % au-delà, dans la limite d'un plafond égal à 7,9 millions d'euros. Il bénéficie essentiellement aux PME (97 % des bénéficiaires ; 73 % du budget de la mesure). En 2005, son coût pour les finances publiques était de 500 millions d'euros.

[39] Le coût d'usage du capital est un coût individualisé, qui intègre la structure du bilan et le taux d'intérêt bancaire, la fiscalité pesant sur les sociétés et les détenteurs d'actions, l'inflation et les amortissements.

[40] Les équations de demande des facteurs de production sont obtenues en résolvant le programme de maximisation du profit par les producteurs. Elles sont explicitées en spécifiant une fonction de production à élasticité de substitution constante (CES) entre le capital et les autres facteurs de production.

Dans ce modèle, le coût du capital agit sur la demande de capital *via* un effet de substitution seulement. La prise en compte d'un effet volume nécessiterait l'estimation d'un modèle plus complet déterminant à la fois les demandes de facteurs et le volume d'activité des entreprises (cf. *infra*).

Plusieurs spécifications sont testées qui permettent de faire une distinction entre les effets de court et de long terme du coût d'usage du capital sur le capital de R & D. Elles sont estimées par la méthode des variables instrumentales pour tenir compte du caractère endogène possible des variables (coût d'usage du capital, production). Les auteurs trouvent des élasticités de court terme au coût d'usage du capital significatives allant de -0,2 à -0,5 selon les spécifications considérées et des élasticités de long terme allant de -0,5 à -0,8.

À partir de ces estimations, les auteurs simulent un scénario contrefactuel dans lequel ils calculent à chaque date le capital de R & D sans la mesure fiscale (sur une durée de quinze ans depuis la date de suppression de la mesure). Les résultats sont exprimés sous la forme d'un ratio appelé « Bang for the Buck » (BFTB), qui mesure les montants investis en R & D additionnelle (en euros) par euro d'avantage fiscal reçu. Ils obtiennent un ratio de 0,5 en moyenne pour l'ensemble des entreprises, mais qui varie sensiblement selon la taille des entreprises (il est 1,2 pour les petites et moyennes entreprises de moins de 200 salariés et de 0,4 pour les entreprises de plus de 200 salariés).

Ce type de modèle peut également être utilisé pour mener des évaluations *ex ante*. Par exemple, Mulkay et Mairesse[41] recourent à un modèle similaire à celui de Lokshin et Mohnen pour évaluer *ex ante* l'effet de la réforme de 2008 du Crédit d'Impôt Recherche (CIR) en France[42]. Le modèle est estimé sur un échantillon d'entreprises suivies pendant les années antérieures à la réforme (2004-2007). Les estimations obtenues servent ensuite à simuler le comportement des entreprises en termes de R & D avec et sans la réforme de 2008. Les auteurs trouvent que la

[41] Voir Mulkay et Mairesse (2013).

[42] Depuis sa création en 1983, le crédit d'impôt recherche (CIR) a connu de nombreuses évolutions : de 1983 à 2003, il a bénéficié aux entreprises qui accroissaient leurs dépenses en R & D ; de 2004 à 2007, il a intégré une part calculée sur l'augmentation des dépenses en R & D et une part calculée sur le niveau des dépenses ; depuis 2008, le dispositif est assis uniquement sur le volume des dépenses en R & D effectuées par les entreprises. Le taux de crédit d'impôt est de 30 % pour les dépenses inférieures à 100 millions d'euros et de 5 % au-delà (voir Commission nationale d'évaluation des politiques d'innovation, 2016).

réforme de 2008 devrait à long terme permettre d'accroître de 20 % le capital investi en R & D.

En résumé, l'approche structurelle apparaît particulièrement utile à la fois pour quantifier les effets et en comprendre les causes. Toutefois, elle repose sur des hypothèses fortes notamment sur les mécanismes économiques qui, si elles ne sont pas valides, peuvent complètement fausser les résultats.

2. L'approche qualitative : connaître les perceptions et les pratiques des acteurs de terrain

Les méthodes qualitatives permettent des investigations approfondies sur de petits échantillons. De plus en plus de dispositifs d'évaluation mêlent aujourd'hui à la fois des travaux qualitatifs et quantitatifs[43]. Les vertus d'une telle démarche mixte ont été soulignées dans divers travaux[44] [45].

Les méthodes qualitatives ne sont pas l'apanage des scientifiques. Elles sont également très utilisées et promues par les consultants du secteur privé et les autres acteurs de l'évaluation des politiques publiques, comme les corps d'inspection, le SG-MAP, la Cour des Comptes ou encore la Société française d'évaluation (SFE).

Le principe : interroger et observer les acteurs

L'approche qualitative permet des investigations approfondies sur le contexte local et l'environnement institutionnel ; la construction et la mise en œuvre d'un dispositif ainsi que les liens et la coordination entre les différents acteurs ; la perception et l'appropriation du dispositif par les différentes catégories d'acteurs. Elle est aussi utile pour mesurer la satisfaction des bénéficiaires, et apprécier, de manière subjective, les retombées du dispositif sur les bénéficiaires finaux à partir des opinions des bénéficiaires eux-mêmes ou de tiers. Elle permet également de révéler les conceptions et les représentations que se font les personnes sur une intervention publique, les attentes et les besoins des acteurs, souvent à l'origine de leurs stratégies, positions et comportements.

[43] En France, le Laboratoire Interdisciplinaire d'Évaluation des Politiques Publiques (LIEPP), rattaché à Sciences Po, promeut l'interdisciplinarité pour l'évaluation des politiques publiques.
[44] Rao et Woolcock (2003).
[45] Wasmer (2014).

Les outils mobilisés par cette approche sont ceux des sciences sociales, comme les entretiens des acteurs, l'observation *in situ* de l'évaluateur, les études de cas ou monographies (encadré 7). L'approche qualitative combine souvent plusieurs de ces outils, et croise fréquemment les informations « subjectives » ainsi obtenues avec des données « objectives ». Les données collectées par ces méthodes et auprès de différents acteurs sont retranscrites, regroupées et comparées pour dégager les avis communs et divergents avant d'être analysées et interprétées par l'évaluateur. Cette « triangulation » des données et des méthodes (i.e. le fait de les combiner) permet d'accroître la fiabilité des résultats.

Comme toute approche scientifique, cette démarche qualitative doit reposer sur un protocole d'analyse rigoureux et s'appuyer sur des outils clairement définis (guides d'entretien précisant les thèmes, grilles d'observation, etc.). Elle doit faire appel à des enquêteurs bien formés aux outils de l'analyse qualitative, en capacité d'animer les discussions tout en restant neutres afin de ne pas orienter les réponses des enquêtés.

L'approche qualitative est menée à partir d'échantillons de petite taille mais reflétant une diversité de situations (diversité de profils d'acteurs, diversité de contextes). Les informations collectées sont mises en regard avec les caractéristiques disponibles sur les bénéficiaires. L'approche qualitative permet notamment de construire des typologies de populations, et d'isoler les facteurs explicatifs déterminants pour chaque catégorie de populations[46].

Cette démarche s'utilise seule ou en combinaison avec une approche quantitative. Dans ce dernier cas, les travaux quantitatifs et qualitatifs peuvent être menés en parallèle sans réelle interaction ou bien de manière séquentielle, les uns se nourrissant des autres.

[46] Dans la restitution faite des résultats, elle intègre souvent les citations textuelles (*verbatim*) les plus marquantes des personnes interrogées, pour rendre palpables les réalités vécues et les représentations des uns et des autres.

Encadré 7

Les principaux outils de l'analyse qualitative

Les entretiens des acteurs prennent différentes formes. Il peut s'agir d'entretiens directifs, semi-directifs ou ouverts ; d'entretiens individuels ou collectifs (dont les « focus groupes ») ; d'entretiens répétés ou non ; d'entretiens en face à face ou téléphoniques. Par rapport aux entretiens individuels, les entretiens collectifs permettent de recueillir plus d'informations en un temps plus court et d'identifier des points de consensus et de désaccords entre les acteurs par la confrontation de leurs points de vue. L'interrogation utilise des questions ouvertes, propices aux échanges entre l'enquêteur et les acteurs interrogés. Elles favorisent la formulation de nouvelles questions auxquelles l'enquêteur n'avait pas pensé initialement ainsi que l'approfondissement de certaines questions.

L'observation directe de l'enquêteur permet à ce dernier de s'immerger dans l'organisation, en qualité d'observateur et sans intervention directe. L'enquêteur peut ainsi observer les acteurs en situation, leurs pratiques et comment ils interagissent. Cette méthode peut parfois aboutir à des constats qui contredisent ce que les acteurs affirment lorsqu'ils sont interrogés.

Les études de cas ou monographies reposent sur l'étude approfondie de cas particuliers (par exemple divers territoires d'application d'une intervention publique) pour en donner une vision la plus complète possible (déroulement de l'intervention publique dans le temps, interrogation du système d'acteurs). Elle s'appuie sur toutes les données disponibles (documentation, entretiens, etc.). Les informations sont collectées sous couvert de confidentialité et d'anonymat.

Référence méthodologique

Conseil scientifique du Fonds d'expérimentation pour la jeunesse (2010), *Guide méthodologique pour l'évaluation des expérimentations sociales*. http://experimentation.jeunes.gouv.fr/5-evaluation-mode-d-emploi.html

La force de cette approche : comprendre les facteurs d'échec et de succès de la politique évaluée

L'intérêt de l'approche qualitative est de considérer de nombreuses dimensions de l'évaluation non couvertes par l'approche quantitative. L'approche qualitative est utile pour évaluer la pertinence[47] ou l'utilité[48] d'une intervention publique ou bien encore

[47] La pertinence mesure l'adéquation des objectifs avec le problème posé au départ.
[48] L'utilité mesure l'adéquation des résultats avec les besoins et les besoins d'aujourd'hui.

pour évaluer sa mise en œuvre et identifier des dysfonctionnements et des obstacles rencontrés par les acteurs. Ces dimensions ne relèvent pas de l'évaluation d'impact mais peuvent fournir des clés de compréhension des résultats portant sur les effets de l'intervention publique considérée. Par exemple, certains dispositifs souffrent d'un taux de non recours important. L'analyse qualitative permet dans ce cas d'en comprendre les raisons et d'expliquer ainsi leur faible incidence.

Pour l'évaluation d'impact, l'approche qualitative, lorsqu'elle est utilisée en complément d'une approche quantitative, permet d'interpréter les résultats issus de l'analyse quantitative. Elle est particulièrement utile quand les explications reposent sur des éléments difficiles à mesurer mais qui sont fondamentaux pour la compréhension des résultats comme les croyances, les conceptions ou les représentations sociales[49]. Elle peut aussi permettre de détecter des erreurs de mesure lorsque des questions de l'enquête (quantitative) ont été mal comprises par les enquêtés. Elle est également importante pour illustrer la diversité des situations, et l'hétérogénéité des pratiques et des comportements.

En amont de l'évaluation d'impact, l'approche qualitative est souvent utilisée pour tester des hypothèses et repérer des effets qui devront ensuite être mesurés. Cette phase exploratoire est importante quand les dispositifs sont innovants ou quand la population cible est mal connue et que les comportements peuvent être inattendus. Elle précède alors souvent la réalisation d'enquêtes par questionnaire exhaustives ou sur des échantillons relativement importants d'individus représentatifs de la population à étudier[50].

L'approche qualitative a aussi l'avantage d'être facilement compréhensible pour un public non spécialiste, même si elle requiert des compétences fortes pour bien maîtriser les outils qualitatifs (cf. *infra*).

[49] Par exemple, dans le cas des politiques en faveur de la réinsertion professionnelle des personnes détenues, une enquête qualitative auprès des dirigeants d'entreprises est utile pour comprendre leurs croyances et leurs représentations à l'origine de leurs comportements.

[50] Ainsi, les enquêtes monographiques menées en 2010-2011 auprès de structures de l'insertion par l'activité économique et de salariés en insertion ont été utiles pour la conception des questionnaires des enquêtes statistiques menées en 2012 par la Dares auprès de ces structures de l'insertion et des salariés passés par de telles structures.

Sa faiblesse : de petits échantillons, la fiabilité des réponses et la difficulté à isoler les effets

La première limite de l'approche qualitative est qu'elle repose sur des petits échantillons qui ne permettent pas d'avoir des résultats statistiquement représentatifs d'une population plus large. Cette faiblesse des échantillons est due au coût important de ces méthodes, lui-même lié à celui de la collecte des informations, en particulier pour les études de cas ou les monographies qui recueillent de nombreuses informations. Les difficultés fréquentes d'accès au terrain (capacité à joindre les personnes et à entrer dans les entreprises) restreignent également les possibilités d'étendre le nombre d'observations.

Une deuxième limite de l'approche qualitative est qu'elle fait davantage appel à la subjectivité des personnes interrogées et à celle des évaluateurs. Les répondants sont ainsi généralement favorables à un dispositif quand ils en sont bénéficiaires. Mais, ils peuvent aussi volontairement ne pas dire la vérité ou tout simplement ne pas savoir répondre. Certains d'entre eux peuvent en effet avoir une retenue à exprimer un point de vue critique sur un dispositif (malgré les principes de confidentialité et d'anonymat des entretiens).

C'est ce qui semble s'être passé pour les jeunes bénéficiaires du Revenu Contractualisé d'Autonomie (RCA) interrogés dans le cadre d'une évaluation qualitative réalisée par le CREDOC[51]. « Leurs réponses sont souvent apparues stéréotypées, notamment en ce qui concerne le bon usage de l'allocation. […] Les modalités de passation de l'entretien, réalisées de façon peu directive et dans un cadre et sur un ton peu formel, ont pu tout de même apparaître pour certains jeunes comme une forme de « contrôle » des usages et du comportement des jeunes »[52]. D'autres peuvent ignorer les réponses quand les dispositifs sont complexes et influent de manière indirecte.

[51] Voir CREDOC (2012).

[52] Un autre exemple est celui des entreprises bénéficiaires d'un dispositif fiscal comme celui en faveur de la R & D. Dans une évaluation qualitative, celles-ci peuvent avoir intérêt à amplifier ses effets pour des raisons stratégiques en se disant que leurs réponses influenceront la décision politique de maintenir ou d'amplifier le dispositif. Mais elles peuvent aussi atténuer volontairement ses effets si l'innovation constitue pour elles un argument marketing et commercial fort. Voir Lentile et Mairesse (2009).

En définitive, l'approche qualitative ne peut capter que ce qui est identifiable à partir des témoignages des acteurs. La qualité des informations dépend de manière cruciale du niveau de connaissances des personnes interrogées et de leur disposition à bien vouloir répondre aux questions. De son côté, l'évaluateur doit analyser et interpréter des discours et du « texte » et non des réponses à des questions fermées pouvant être facilement codifiées et traitées informatiquement. Il doit être sûr de bien saisir le sens des propos des enquêtés et pouvoir recouper les informations pertinentes.

Enfin, une troisième limite est que l'approche qualitative peut difficilement être utilisée pour mesurer les effets d'une intervention publique car elle ne peut pas indiquer de manière fiable ce qui se serait passé sans le programme. L'évaluation subjective faite par les bénéficiaires eux-mêmes n'est pas un critère suffisant d'appréciation d'un dispositif, de même que la satisfaction des bénéficiaires à l'égard d'un dispositif. En effet, dans l'approche qualitative, les effets d'un dispositif sont appréhendés à partir de ce que les bénéficiaires disent des changements opérés dans leur vie grâce au dispositif. Or il est extrêmement difficile pour eux de démêler dans les changements ayant affecté leur vie ceux qui sont imputables au programme et ceux qui ne le sont pas. Par ailleurs, la comparaison avant et après la mise en place du dispositif n'est pas suffisante pour identifier des effets. À tout le moins il faudrait pouvoir recourir à un groupe de comparaison composé d'individus similaires du point de vue de leurs caractéristiques mais n'ayant pas bénéficié du dispositif. De même, ce n'est pas parce que les personnes se déclarent satisfaites d'un dispositif que celui-ci remplit ses objectifs. En effet, comme nous l'avons vu, les individus ont généralement tendance à se déclarer satisfaits d'un dispositif dès lors qu'il leur octroie un supplément d'aides (subvention, prêt à taux réduit, réduction de la fiscalité, etc.).

Une illustration : l'évaluation du revenu contractualisé d'autonomie

Le revenu contractualisé d'autonomie (RCA) est un exemple de dispositif qui a fait l'objet d'une démarche d'évaluation mixte. Ce programme expérimental d'aide à l'insertion professionnelle des jeunes faiblement qualifiés[53] mis en œuvre en 2011 a été évalué par deux méthodes : une évaluation aléatoire contrôlée[54] et des enquêtes de terrain à la fois auprès des acteurs institutionnels et des jeunes accompagnés[55].

L'évaluation quantitative visait à évaluer l'impact du dispositif sur l'insertion professionnelle des jeunes. Le protocole expérimental, conforme aux standards scientifiques internationaux, a permis la comparaison de deux groupes de jeunes – bénéficiaires du RCA (groupe test) et du CIVIS (groupe témoin) – parfaitement comparables en moyenne[56]. L'évaluation a porté sur environ 5 500 jeunes s'inscrivant en CIVIS pour la première fois en février ou mars 2011 dans les missions locales participant à l'expérimentation. Le suivi statistique des jeunes des deux groupes a reposé sur des enquêtes téléphoniques complétées par des données de gestion issues du système d'information sur les missions locales[57].

Les résultats montrent que les jeunes bénéficiaires du RCA ont été plus nombreux à se rendre à la mission locale, ont bénéficié de davantage de rendez-vous avec leur conseiller et se sont vus proposer davantage de formations, ateliers et offres

---

[53] Mis en œuvre à titre expérimental dans les missions locales, le RCA s'adresse aux jeunes éligibles au Contrat d'Insertion dans la Vie Sociale (CIVIS). Il a été lancé en 2011 dans 82 missions locales. Il octroie une allocation mensuelle de 250 euros pendant un an, qui décroît par trimestre au cours de la deuxième année (240 euros/mois au 1er trimestre ; 180 euros/mois au 2e trimestre ; 120 euros/mois au 3e trimestre et 60 euros/mois au 4e trimestre). Le montant de l'allocation est modulé en fonction des revenus d'activité (il devient nul quand les revenus d'activité sont au moins égaux à un SMIC à temps complet). Les jeunes bénéficient en parallèle d'un accompagnement par les conseillers des missions locales, qui est le même que celui offert aux bénéficiaires d'un CIVIS.

[54] Voir Aeberhardt *et al*., (2014) ; Bérard et Valdenaire (2014).

[55] Voir Gilles et Guisse (2014).

[56] Les deux groupes, le groupe test composé de jeunes bénéficiaires du RCA comme le groupe témoin composé de jeunes bénéficiaidréres du CIVIS, ont été constitués par tirage au sort parmi les missions locales volontaires.

[57] Les jeunes suivis ont répondu à trois enquêtes réalisées en avril 2011, 2012 et 2013. Elles permettent de mesurer de nombreux indicateurs sur l'accompagnement reçu, les emplois occupés, les efforts de recherche d'emploi, la formation, les ressources et les dépenses des jeunes et leur intégration sociale. Ces informations ont été complétées par des données provenant du système d'information des missions locales (appelé Parcours 3) sur l'éventuel arrêt du programme, le nombre d'entretiens avec le référent, le nombre de formations et d'ateliers suivis, etc.

d'emploi[58]. En revanche, aucune différence statistiquement significative n'a été observée entre les bénéficiaires du RCA et du CIVIS sur le taux d'emploi, la qualité des emplois (CDI, temps complet), la formation ou bien encore les efforts des jeunes pour la recherche d'emploi.

Enfin, l'allocation a eu comme contrepartie une baisse des autres sources de revenus (en particulier des aides parentales), si bien que les ressources des jeunes bénéficiaires du RCA ont faiblement progressé en moyenne[59]. Ces résultats semblent suggérer que le surcroît d'accompagnement octroyé aux bénéficiaires du RCA (suivi plus régulier, plus riche et plus durable) n'a pas été suffisant pour faire la différence avec les bénéficiaires du CIVIS. De même, l'aide financière accordée était probablement insuffisante pour modifier les comportements des jeunes.

L'enquête qualitative auprès des jeunes a été réalisée en parallèle dans le cadre du CREDOC auprès de 70 jeunes bénéficiaires du RCA suivis dans plusieurs missions locales. Elle a reposé sur des entretiens individuels, menés en face à face (mode privilégié) et par téléphone (par défaut), en deux temps : une première vague, un an après l'entrée dans le dispositif (auprès de 34 jeunes) et une deuxième vague, deux ans après l'entrée dans le dispositif (auprès de 35 jeunes). Cette enquête a complété une première enquête qualitative réalisée auprès de 80 acteurs (conseillers et membres de l'équipe de direction) dans huit missions locales, au moment du démarrage du dispositif. Cette dernière enquête visait notamment à analyser les conditions de mise en œuvre du dispositif et les pratiques professionnelles des conseillers ainsi qu'à identifier les effets du RCA sur les parcours des jeunes du point de vue des acteurs.

Les objectifs de l'évaluation qualitative auprès des jeunes étaient de quatre ordres :
- appréhender les perceptions et opinions des jeunes à l'égard du dispositif (ce qu'ils pensaient de son principe et de son fonctionnement) ;

[58] Cela tient notamment au fait que les jeunes en RCA doivent obligatoirement se rendre chaque mois à la mission locale pour pouvoir bénéficier de l'allocation et que le RCA est un programme d'accompagnement sur deux ans contre un an pour le CIVIS.

[59] De fortes disparités ont cependant été observées entre les jeunes ayant les ressources les plus faibles qui ont connu une nette progression de leurs ressources et ceux ayant le plus de ressources qui ont vu leurs ressources inchangées.

- retracer les principales évolutions dans leur situation et leur parcours depuis l'obtention du RCA[60] et déterminer si elles étaient liées au RCA ;
- identifier l'usage et l'utilité du RCA ;
- analyser le point de vue des bénéficiaires sur la suite de leurs parcours : la fin de l'allocation et leur vision de l'avenir (voir le guide d'entretien des bénéficiaires)[61].

L'enquête qualitative auprès des jeunes a montré que le dispositif a été globalement bien perçu par ceux-ci : l'allocation mensuelle semble avoir généré chez la majorité d'entre eux un sentiment de responsabilité, de devoir ou de contrepartie vis-à-vis de la société ; les contacts mensuels constituant une obligation contractuelle ont pu constituer un précieux soutien moral. En revanche, le principe de la dégressivité de l'allocation souvent méconnu par les jeunes n'a pas été anticipé.

L'enquête qualitative illustre également l'hétérogénéité des effets du RCA sur les parcours d'insertion selon la situation des jeunes. Elle identifie cinq types de parcours, des « dynamisés » par le dispositif pour qui le RCA a permis la définition et la mise en place d'un projet d'insertion jusqu'aux « désaffiliés » en rupture avec les institutions et qui ne souhaitent pas s'inscrire dans un parcours d'insertion professionnelle. Enfin, l'enquête qualitative montre que le RCA a permis de sécuriser les jeunes et de les rendre plus autonomes, que cette autonomie soit réelle (autonomie résidentielle, mobilité) ou davantage symbolique (affirmation de soi).

Les enseignements de ces évaluations ont été utilisés pour la définition du nouveau dispositif expérimental « Garantie jeunes »[62]. En particulier, la Garantie jeunes vise un public très désocialisé, pas ou peu diplômé, qui n'est ni en emploi, ni en études, ni en formation, alors que le RCA visait des jeunes déjà inscrits en CIVIS, donc déjà en

[60] Les situations sont considérées par rapport à l'emploi, la formation, le logement, la vie familiale, la situation budgétaire des jeunes interrogés.

[61] Pour plus de détails sur les questions posées, se reporter au guide d'entretien figurant en annexe III (pages 56 à 60) du rapport du CREDOC (Gilles et Guisse, 2014).

[62] Mise en place à titre expérimental depuis l'automne 2013, la Garantie jeunes a concerné 72 départements (dix départements en 2013, 62 départements en 2015) et 46 000 jeunes. Elle sera élargie à 19 départements supplémentaires. Elle propose un accompagnement intensif mettant l'accent sur l'accès à l'emploi *via* les périodes de mise en situation professionnelle, et offre une allocation forfaitaire mensuelle de 452,21 euros (depuis le 1er janvier 2015), dégressive à partir de 301 euros de revenus d'activité nets et s'annulant à 80 % du Smic brut. Le public est composé de jeunes en situation de grande précarité financière (91 % des jeunes étaient ni étudiants, ni en emploi, ni en formation (Neets) au 31 décembre 2015).

contact avec les missions locales. Elle propose également un accompagnement plus intensif que dans le cadre du RCA (suivi très régulier) comprenant une dimension collective importante au début du parcours (mise en place de groupes de travail composés de jeunes). Enfin, l'aide financière a été presque doublée (450 euros environ contre 250 euros auparavant).

Ce dispositif fait actuellement l'objet d'une évaluation sous l'égide d'un comité scientifique qui a pour mission :

- de quantifier et caractériser le public éligible à la Garantie jeunes ;
- décrire les parcours des jeunes et le rôle des contrats d'engagement réciproque ;
- analyser le fonctionnement des partenariats locaux et leur évolution ;
- identifier et comprendre les changements sur les pratiques d'accompagnement des missions locales ;
- mesurer l'impact de la Garantie jeunes sur les parcours d'insertion sociale et professionnelle et les conditions de vie.

Comme pour le RCA, l'évaluation de la Garantie jeunes comprend trois volets confiés à des équipes distinctes : une évaluation statistique ; deux évaluations qualitatives, l'une centrée sur les acteurs institutionnels et les autres sur les jeunes.

En résumé, les méthodes qualitatives sont particulièrement utiles lorsqu'il s'agit d'évaluer la pertinence ou l'utilité d'une intervention publique ou encore sa mise en œuvre. Elles sont beaucoup moins performantes quand elles sont utilisées seules pour étudier son efficacité ou son efficience. En revanche, combinées avec une approche quantitative de l'évaluation, elles offrent des clés utiles pour expliquer les résultats issus de l'analyse quantitative (comme les éléments de contexte, sources d'efficacité ou d'inefficacité) et illustrer la diversité des situations, des pratiques et des comportements. Comme l'illustrent les exemples précédents, la combinaison des méthodes quantitatives et qualitatives peut être particulièrement utile.

CHAPITRE 4

COMPARER LES EFFETS DE POLITIQUES DIFFÉRENTES

Jusqu'à présent, on s'est intéressé à l'évaluation de l'impact d'un programme par rapport à la situation qui prévaudrait en son absence. Le décideur peut cependant souhaiter comparer les effets de différents dispositifs (par exemple, plusieurs types de formation pour les chômeurs) pour juger de leur efficacité relative, et mieux allouer les fonds publics. Les méthodes présentées ci-dessus ont été étendues au cas de plusieurs « traitements » servant le même objectif[63]. L'analyse coût-efficacité ou coût-bénéfice permet de mesurer l'efficience d'une intervention, c'est-à-dire de mettre en regard les résultats et les coûts.

1. L'évaluation de traitements multiples visant un même objectif

Les méthodes d'évaluation empiriques passées en revue dans la première partie de cette guide peuvent être étendues au cas où il existe plusieurs dispositifs (appelés traitements multiples[64]. Pour ne pas alourdir la présentation, nous illustrons ce point seulement dans le cas du matching sur le score de propension.

[63] Dans le point 1. de ce chapitre, on évoque le cas où le traitement est discret c'est-à-dire lorsqu'il prend un ensemble fini de valeurs. Néanmoins, les méthodes passées en revue *supra* ont également été étendues dans le cas d'un traitement continu prenant un continuum de valeurs. Par exemple, pour la politique d'allégements de charges sur les bas salaires, le barème d'exonérations est dégressif entre 1 et 1,6 fois le SMIC. La méthode consiste alors à comparer les individus selon qu'ils sont plus ou moins traités. Dans le cas du *matching* sur variable continue, voir par exemple Hirano et Imbens (2004). Pour une application à la formation professionnelle, voir Ferracci *et al.*, (2014).

[64] Frölich (2004).

Le principe : comparer deux à deux les dispositifs

Dans le cas de l'appariement sur le score de propension (appariement), l'étude de l'effet d'un dispositif *j* par rapport à un autre dispositif *k* se fait en se plaçant sur le sous-échantillon des individus ayant suivi l'un ou l'autre de ces deux dispositifs et en appariant le résultat pour chaque bénéficiaire du dispositif *j* à la moyenne pondérée des résultats des bénéficiaires du traitement *k* sur la base du score conditionnel[65].

Si le principe est simple, les résultats sont cependant plus difficiles à présenter et à interpréter : les résultats s'interprètent en effet en termes d'efficacité relative du traitement *j* par rapport au traitement *k* et non en termes d'efficacité absolue ; ils peuvent différer en pratique selon que l'on s'intéresse à l'effet moyen du dispositif *j* par rapport au dispositif *k* (pour les individus ayant reçu le dispositif *j*) ou à l'effet moyen du dispositif *k* par rapport au dispositif *j* (pour les individus ayant reçu le dispositif *k*).

Une illustration : l'évaluation de dispositifs d'emploi pour les jeunes en France

Brodaty, Crépon et Fougère[66] étudient ainsi l'impact de plusieurs dispositifs d'emploi destinés aux jeunes chômeurs sur leur taux de transition du chômage vers l'emploi. Leur travail couvre deux périodes distinctes : 1986-1988 et 1995-1998. Différents dispositifs d'emploi sont étudiés : le contrat d'apprentissage, le contrat de qualification, le contrat d'adaptation, les stages d'initiation à la vie professionnelle, les travaux d'utilité collective, les stages pour les 16-25 ans.

Ces auteurs utilisent des estimateurs par appariement adaptés à de tels traitements multiples Leur étude repose sur l'exploitation de l'enquête « Suivi des chômeurs » réalisée par l'Insee, indiquant chaque mois l'état des jeunes chômeurs sur le marché du travail, entre novembre 1986 et mai 1988 (en emploi à durée indéterminée, en emploi à durée déterminée, inactif, dans l'un des dispositifs étudiés). L'échantillon étudié comprend 3 160 jeunes chômeurs âgés de moins de 27 ans en août 1986.

Les auteurs trouvent que, dans le secteur marchand, l'impact de l'alternance sur la probabilité d'accéder à un emploi stable est supérieur à celui des emplois aidés. Ceci suggère que l'expérience acquise dans l'emploi ne suffit pas à améliorer l'insertion de manière durable, et qu'il est nécessaire de lui associer des actions d'enseignement.

[65] Lechner (2002).
[66] Brodaty *et al.* (2001).

L'étude montre également que les programmes d'apprentissage sont plus efficaces pour les individus dont la durée en chômage est inférieure à un an au début du programme, que pour les chômeurs de longue durée.

2. L'analyse coût-efficacité ou coût-bénéfice comme critères d'efficience

L'efficience d'une intervention publique repose sur la combinaison des informations relatives aux coûts et aux bénéfices induits par l'intervention publique. Elle peut se fonder sur l'analyse coût-efficacité qui définit un ratio entre le coût et le niveau d'efficacité ou bien sur l'analyse coût-bénéfice qui calcule la différence entre les bénéfices et les coûts exprimés en termes monétaires. Cette deuxième méthode fait appel au calcul économique public, appelé aussi calcul socioéconomique, prenant en compte tous les agents concernés et tous les impacts[67]. Cette méthode est très fréquemment appliquée pour évaluer la rentabilité économique des infrastructures de transport, mais elle tend à se diffuser aujourd'hui plus largement à d'autres leviers (programmes) et domaines d'action possibles (énergie, santé, éducation, etc.) grâce aux progrès méthodologiques[68] [69].

Le principe : comparer les coûts et les bénéfices

Comparer les dispositifs entre eux pour déterminer les plus efficients requiert de mettre en regard leurs bénéfices et leurs coûts respectifs. Les bénéfices correspondent à l'ensemble des effets directs et indirects (c'est-à-dire les externalités qui affectent l'ensemble des acteurs), marchands et non marchands. Les coûts sont quant à eux ceux de l'investissement initial, de l'entretien, du fonctionnement.

Les bénéfices et les coûts sont mesurés dans deux situations, l'une avec le dispositif et l'autre sans le dispositif (situation dite de référence). Ils sont également mesurés dans la durée (à court, moyen et long termes), comparés et agrégés à l'aide d'un taux d'actualisation qui reflète les arbitrages et les préférences des acteurs économiques vis-à-vis du temps (la valeur accordée aux bénéfices ou aux coûts aujourd'hui par rapport à ceux de demain). Enfin, ils sont mesurés pour différentes catégories d'acteurs, comme les bénéficiaires du dispositif et la société.

[67] Le calcul économique public est plus large que le calcul économique privé qui ne considère que les impacts et les coûts qui affectent directement l'entrepreneur privé.
[68] Voir Quinet (2013), rapport publié à l'issue de la mission présidée par E. Quinet.
[69] Mission d'évaluation des politiques publiques (2008)..

Deux méthodes principales existent : l'analyse coût-efficacité et l'analyse coût-bénéfice. Elles diffèrent par la manière dont elles mesurent les bénéfices, les coûts étant mesurés de la même façon dans les deux méthodes (encadré 8). L'analyse coût-efficacité est utile quand on doit comparer des dispositifs qui visent un seul et même objectif (stimuler l'emploi ou bien réduire le nombre d'accidents de la route). Elle nécessite en effet de se centrer sur un résultat unique ou principal d'un dispositif. L'intérêt de l'analyse coût-bénéfice est de considérer une large gamme d'effets dans les bénéfices. Elle permet de comparer aussi bien des dispositifs visant plusieurs objectifs simultanément que des dispositifs variés, dans des domaines très différents (santé, emploi, éducation, etc.). Au-delà de son utilité pour ces comparaisons de dispositifs, elle peut être utilisée pour déterminer si un dispositif est en lui-même rentable et vaut la peine d'être financé ou poursuivi d'un point de vue économique (ce sera le cas si l'ensemble des bénéfices excède la totalité des coûts).

Encadré 8

Les méthodes coût-efficacité et coût-analyse

Dans l'analyse **coût-efficacité**, les bénéfices d'un dispositif sont exprimés dans les unités de mesure des résultats principaux (comme le nombre d'emplois créés s'il s'agit d'un dispositif d'emploi ou le nombre d'accidents de la route évités s'il s'agit d'un dispositif de prévention routière). Les résultats apparaissent alors sous la forme d'un ratio représentant le coût pour produire une unité de bénéfice (par exemple, le coût pour créer un emploi ou le coût pour prévenir un accident de la route).

Dans l'analyse **coût-bénéfice**, les bénéfices sont exprimés en monnaie (dollars, euros, etc.). Les résultats apparaissent sous la forme d'un bénéfice net égal à la différence entre les bénéfices et les coûts ou sous la forme d'un ratio entre les coûts et les bénéfices ou bien encore sous la forme d'un taux de rentabilité interne (TRI) égal à la valeur du taux d'actualisation pour lequel les bénéfices sont égaux aux coûts. Elle inclut tous types de bénéfices et coûts, y compris ceux non marchands (par exemple, les effets environnementaux et sociaux comme la perte de vies humaines ou les accidents, les effets sur la santé, les effets psychologiques comme le stress, la pollution des sols ou de l'air, etc.). Elle requiert donc de faire des hypothèses fortes sur la valorisation monétaire des différents bénéfices, lesquelles peuvent largement différer selon les catégories d'acteurs et les intérêts qu'ils représentent. Elle est donc plus souvent controversée.

La qualité de ces analyses dépend de plusieurs facteurs :

- la prise en compte de l'ensemble des coûts et des bénéfices ;
- la complétude et la qualité des données pour les mesurer ;
- la qualité des méthodes utilisées pour évaluer les effets des dispositifs ;
- la fiabilité des résultats par rapport à des changements dans les hypothèses posées (notamment par rapport au taux d'actualisation).

L'application des résultats issus de différents contextes à un contexte donné (la France par exemple à partir de résultats provenant de différents pays) n'est cependant pas immédiate, et requiert une discussion approfondie des effets et des coûts dans le contexte considéré.

Une illustration : l'évaluation du programme d'éducation Perry Preschool

Heckman *et al.*[70] ont évalué le rendement socio-économique du programme éducatif préscolaire intensif appelé *Perry Preschool* mis en place aux États-Unis dans les années soixante[71].

Les coûts initiaux du programme par enfant ont été comparés aux effets du programme au moment où il atteignait l'adolescence et l'âge adulte sur les salaires, les impôts payés, le recours à l'aide sociale et les frais déboursés en matière de justice criminelle pour chacun. L'étude ne considère pas les effets sur la santé faute de données dans l'étude de suivi des cohortes d'enfants[72]. Elle ne considère pas non plus les modifications liées au comportement parental, mais seulement celles liées aux enfants bénéficiaires.

Le programme de la Perry Preschool s'est accompagné d'un dispositif d'évaluation rigoureux reposant sur la comparaison de deux groupes constitués par tirage au sort – le groupe d'enfants traités ayant bénéficié du programme et le groupe d'enfants de contrôle n'en ayant pas bénéficié. Les enfants des deux groupes ont été suivis sur

[70] Heckman *et al.* (2010).

[71] La Perry Preschool est un programme éducatif préscolaire intensif mis en place dans la commune d'Ypsilanti dans le Michigan aux États-Unis entre 1962 et 1967. Il s'adresse à des enfants afro-américains issus de milieux défavorisés âgés de trois à quatre ans pendant le programme. D'une durée de un à deux ans, il consiste en des cours de deux heures et demi tous les jours de la semaine et des visites à domicile d'une heure et demi une fois par semaine.

[72] Une étude spécifique sur la santé a toutefois été menée auprès des participants inscrits à la Perry Preschool à l'âge de 40 ans.

plusieurs décennies jusqu'à l'âge de 40 ans[73]. Au total, 123 enfants ont été suivis, dont 58 appartenaient au groupe des traités et 65 au groupe de contrôle. Les variables mesurées peuvent être classées en trois grandes catégories : la réussite scolaire[74] ; la réussite socio-économique[75] et les risques sociaux[76].

L'étude de Heckman et *al.* repose sur les enquêtes de suivi des individus appartenant aux groupes de traitement et de contrôle jusqu'à l'âge de 40 ans. Les différents effets sont projetés au-delà de 40 ans jusqu'à l'âge de 65 ans. Comme les études précédemment menées sur la base des suivis réalisés à des âges antérieurs[77], les auteurs trouvent un impact important et durable de ce programme sur la réussite des enfants d'origine modeste, qu'elle soit scolaire ou professionnelle et sur les risques sociaux.

Heckman *et al*. montrent que la dépense initiale nécessaire pour financer le programme, pourtant élevée, est très inférieure aux bénéfices retirés par la société durant les décennies suivantes. Le taux de rentabilité interne (TRI, voir encadré 8) du Perry Preschool serait compris entre 7 et 10 % annuels selon les hypothèses retenues. Exprimé en termes de ratio entre les coûts et les bénéfices chaque dollar investi dans le programme produirait en retour entre 7 et 12 dollars de bénéfices collectifs pour la société (avec un taux d'actualisation de 3 %).

Ces résultats sont nettement plus faibles que ceux issus d'études précédentes, notamment celles de Belfield *et al.*[78] et de Rolnick et Grunewald[79] (tableau 1). En

[73] Des enquêtes de suivi ont été réalisées aux âges de 15, 19, 27 et 40 ans. Elles ont été complétées par des données administratives issues des établissements scolaires, des tribunaux, de la police et de l'aide sociale.

[74] La réussite scolaire est mesurée en termes de notation académique, d'orientation scolaire vers des programmes d'éducation spécialisée, d'obtention de diplômes secondaires ou universitaires, d'allongement de la durée des études.

[75] La réussite socio-professionnelle est mesurée en termes de taux d'emploi, de salaire moyen, de taux d'accès à la propriété.

[76] Les risques sociaux considérés concernent les taux de délinquance et de criminalité, de dépendance envers les aides publiques, de taux de grossesses précoces.

[77] Un des apports de l'étude d'Heckman *et al*. est d'avoir corrigé plusieurs problèmes méthodologiques rencontrés dans diverses études antérieures et qui jetaient un doute sur la robustesse des résultats.

[78] Belfield *et al.* (2006).

[79] Rolnick et Grunewald (2003).

particulier, les bénéfices pour la société sont sensiblement plus faibles en termes de criminalité. Ces différences sont liées à la façon dont les auteurs évaluent les coûts sociaux de la criminalité. Elles illustrent combien les résultats peuvent être sensibles aux hypothèses faites sur les effets et les coûts (alors même qu'ils reposent sur le même suivi longitudinal des bénéficiaires et des non-bénéficiaires). Il n'en demeure pas moins que les résultats sont impressionnants : ce programme dont le coût initial est élevé génère des gains économiques considérables.

La transposition directe des résultats à d'autres contextes n'est cependant pas immédiate. Dans le cas français, il n'est pas du tout certain que la mise en place d'un tel programme aujourd'hui aboutirait à des résultats d'ampleur comparable. D'une part, l'expérimentation est ancienne (elle date des années soixante) et il est possible que les comportements en réaction à ce type de programme soient aujourd'hui différents (aux États-Unis comme en France). D'autre part, la France n'est pas dans la même situation que les États-Unis : elle ne connaît pas les mêmes taux de criminalité et de pauvreté radicalisée.

Ce débat autour de l'ampleur des effets ne remet cependant pas en cause l'efficacité ou l'efficience de ce type d'intervention précoce auprès des jeunes enfants, aujourd'hui largement reconnues dans la communauté académique internationale.

Tableau 1
Comparaison des résultats de l'analyse coût-bénéfice du programme Perry Preschool
(en dollars 2006 ; taux d'actualisation : 3 %)

	Heckman *et al.* (2010)	**Belfield *et al.*** (2006)	**Rolnick et Grunewald** (2003)
Éducation	4 325	14 382	9 034
Revenus	78 010	68 429	43 583
Criminalité	66 780	386 985	101 132
Aide sociale	3 698	3 118	381
Total des bénéfices	152 813	472 914	154 130
Coût initial du programme	17 759	17 759	17 759
Ratio entre les coûts et les bénéfices[a]	**8,6** *(3,9)*	**26,6**	**8,7**
Taux de rendement interne[d]	**8,6** *(2,6)*	**21,0**	**16,0**

Notes : Résultats issus de Heckman *et al.*, tableau 8, page 126. Pour comparer leurs résultats avec ceux des deux autres études, ces auteurs retiennent une méthode similaire à celles de ces deux autres études (pour interpoler les valeurs manquantes avant 40 ans et pour extrapoler au-delà de 40 ans).
a et d : Les chiffres entre parenthèses sont les écarts-types estimés.

Source : Heckman et al. (2010)

CHAPITRE 5

DE L'ÉVALUATION À LA RECOMMANDATION DE POLITIQUE PUBLIQUE

Le panorama qui précède montre que les évaluations d'impact peuvent se heurter à des problèmes de nature aussi bien scientifique qu'opérationnelle. Pour parvenir à des résultats rigoureux mais également compréhensibles par les décideurs et l'opinion publique, un certain nombre de principes doivent être appliqués de façon systématique dans une démarche d'évaluation, que celle-ci soit interne à l'administration, ou bien externalisée à destination de chercheurs académiques ou d'experts de la sphère privée[80].

S'appuyer sur les travaux disponibles

Avant de collecter des données nouvelles ou de lancer un protocole d'exploitation des données existantes, une démarche vertueuse consiste à établir une revue de la littérature sur les évaluations d'impact similaires menées en France ou à l'étranger. Si les résultats obtenus dans des contextes institutionnels différents ne sont pas immédiatement transposables, il est néanmoins parfois possible d'en dégager des conclusions probantes lorsque les travaux convergent. Ceci peut parfois permettre d'exclure ou au contraire d'adopter un dispositif sans évaluation préalable sur données françaises[81].

[80] Pour une analyse plus complète des enjeux institutionnels de l'évaluation des politiques publiques, voir Ferracci et Wasmer (2011).

[81] Le cas des interventions sur les très jeunes enfants issus de milieux défavorisés, développé *supra*, illustre bien ce point. Le consensus international est aujourd'hui tel sur l'efficacité de ce type d'interventions que le principe politique en est désormais assumé en France, sans qu'une évaluation randomisée ait été menée à ce jour dans le contexte hexagonal. Pour une revue des études, voir Burgess (2016).

Ce travail de revue de la littérature peut être parfois difficilement conclusif compte tenu de la diversité des résultats. C'est le cas lorsqu'ils portent sur des dispositifs mis en œuvre dans différents pays, sur différentes périodes et qu'ils sont obtenus à l'aide de méthodes et de bases de données également différentes. La réconciliation des résultats peut se faire grâce à une méta-analyse : elle consiste à rassembler les résultats d'un grand nombre d'évaluations portant sur une même variable de résultat et à traiter les résultats de ces évaluations comme les unités d'observation d'une base de données[82].

Prendre garde aux limites de chaque méthode

Une bonne évaluation d'impact doit être capable d'identifier correctement l'effet causal d'un dispositif sur les bénéficiaires. Cette validité interne des résultats dépend de façon cruciale de la méthode employée au regard des données disponibles.

Les méthodes athéoriques d'évaluation des politiques publiques ont été développées pour répondre spécifiquement aux questions d'identification des effets[83]. Parmi elles, les méthodes randomisées sont particulièrement performantes en matière de validité interne, mais ne sont pas toujours applicables pour des raisons politiques ou opérationnelles. L'évaluateur doit alors se tourner vers les méthodes quasi-expérimentales mobilisant les sources de données existantes. Les méthodes des variables instrumentales, les différences de différences ou les régressions par discontinuité offrent alors à l'évaluateur un « second best ». Sous réserve de disposer dans les données d'une source de variation dans l'accès au dispositif qui soit parfaitement indépendante de la variable de résultat, l'évaluateur est alors en mesure de corriger les biais de sélection provenant à la fois de variables observables et inobservables. Dans le cas contraire, un « third best » consiste à mettre en œuvre les méthodes d'appariement (« matching »). Pour être convaincantes, elles doivent reposer sur des données extrêmement riches quant aux caractéristiques des individus, mais aussi à leur environnement économique et sociétal. Ainsi l'impact d'une mesure d'aide à l'embauche aura un effet différent sur deux individus pourtant

[82] La méta-analyse de Card, Kluve et Weber (2015) consacrée aux politiques d'emploi offre une illustration éclairante de cette démarche mais aussi des problèmes méthodologiques qu'elle pose. En particulier, il est possible que les travaux exploités dans la méta-analyse souffrent d'un « biais de publication positive », résultant du fait que les articles qui établissent un impact significatif d'un dispositif ont souvent plus de chances d'être publiés. Voir également la méta-analyse réalisée par Lentile et Mairesse (*op. cit.*) sur les mesures fiscales en faveur de la R & D.

[83] Voir *supra*, le point 2., chapitre 1.

similaires, selon qu'ils se trouvent dans des bassins d'emploi plus ou moins dynamiques économiquement. Dans ce contexte il est nécessaire de disposer d'informations précises sur les métiers en tensions et la dynamique des créations d'emploi dans chaque bassin d'emploi.

Les méthodes structurelles permettent également de mesurer les effets d'une politique publique et d'en comprendre les causes, en explicitant les mécanismes par lesquels une politique produit ses effets. Les résultats sont cependant très sensibles aux hypothèses faites dans les modèles sur les mécanismes économiques considérés, les formes fonctionnelles, ou aux hypothèses émises sur la distribution des aléas.

La qualité de l'évaluation d'impact tient souvent à l'ingéniosité de l'évaluateur et à sa connaissance du cadre institutionnel, des méthodes et des bases de données existantes. Ces caractéristiques lui permettent de trouver la méthode qui constitue le meilleur compromis entre la réduction des biais et les contraintes de mise en œuvre techniques (accès aux données, procédures d'estimation). Lorsque le dispositif étudié et les données le permettent, il est par ailleurs très utile de leur appliquer des méthodes différentes afin de vérifier que les résultats ne sont pas déterminés par les hypothèses parfois restrictives inhérentes à chaque méthode.

S'interroger sur la nature transposable ou non des résultats à une population plus large

Les effets obtenus sur une population donnée ne sont pas forcément directement applicables à une autre population, composée d'individus différents (la validité externe des résultats). En d'autres termes, la validité interne d'une évaluation n'implique pas forcément sa validité externe. En revanche, il ne peut y avoir de validité externe sans validité interne. Ainsi les méthodes d'évaluation précédentes gagnent-elles à être appliquées à des sous-populations, afin de dépasser les enseignements parfois trompeurs que peut recéler l'effet moyen d'un traitement. Indépendamment de l'hétérogénéité des publics, il est possible que l'extension de la politique à une population plus large en modifie substantiellement les effets.

Pour s'assurer de la validité externe des résultats, recourir à une approche structurelle permet d'anticiper, au moyen de simulations préalables, les effets d'une

généralisation éventuelle de la politique publique[84]. Par ailleurs, certains protocoles expérimentaux permettent de quantifier les externalités inhérentes à la généralisation d'une mesure sans recourir à l'analyse structurelle. Ce type de protocole reste cependant rare et est en cours de développement méthodologique dans le milieu académique[85].

Favoriser la pluridisciplinarité des approches

Établir des résultats qui montrent l'impact d'un traitement ne renseigne pas nécessairement sur les mécanismes par lesquels ce dispositif agit. Ceci peut rendre difficile le passage de l'évaluation à la décision de politique publique, surtout lorsque le dispositif évalué comporte plusieurs dimensions dont l'évaluation ne parvient pas à dissocier les effets[86]. Afin de traduire les résultats d'une évaluation en mesures opérationnelles, il est ainsi souvent utile de compléter l'approche quantitative par une approche qualitative. Si les coûts de coordination associés au dialogue entre sociologues et économistes ne sont pas négligeables, ils ne sont pas pour autant insurmontables, et les bénéfices à attendre de cette interdisciplinarité en matière d'efficacité des politiques publiques sont importants[87].

Sélectionner les résultats disponibles et les équipes

Les constats précédents invitent les commanditaires d'une évaluation d'impact à se montrer sélectifs dans l'exploitation des résultats d'évaluation disponibles, ou dans le choix des équipes susceptibles de mener à bien une démarche d'évaluation.

Ainsi les contraintes pratiques associées aux protocoles d'expérimentation aléatoire imposent de maîtriser les différentes méthodes d'estimation associées à ces

[84] Voir *supra*, chapitres 2 et 3.

[85] Voir sur ce point l'article de Crépon *et al.* (2013, *op. cit.*) portant sur l'évaluation de l'accompagnement renforcé des chômeurs.

[86] Les évaluations menées sur les lois Aubry ayant ramené la durée légale du travail de 39 h. à 35 h. illustrent ce point. Alors que ces lois prévoyaient des allègements de charges sociales (ACS) en contrepartie de la création d'emplois dans le cadre de la réduction du temps de travail (RTT), les évaluations se contentaient d'établir l'impact des dispositifs Aubry, sans qu'il soit possible de conclure à l'efficacité respective des ACS et de la RTT.

[87] Voir *supra* chapitre 3 de ce guide. Voir aussi Wasmer (*op. cit.*).

protocoles[88]. S'agissant des méthodes non expérimentales, elles connaissent un grand nombre de variantes qui leur permettent de s'adapter aux données existantes et de maximiser leur potentiel d'identification de l'effet causal du traitement[89].

Les méthodes structurelles requièrent quant à elles des compétences à la fois en modélisation et en économétrie, tandis que les méthodes qualitatives exigent des évaluateurs une bonne connaissance des outils issus des sciences sociales.

Les trois premières méthodes, par nature très techniques sont généralement appliquées par des chercheurs expérimentés appartenant à la sphère académique. Les travaux réalisés sont susceptibles d'être publiés dans des revues scientifiques à comité de lecture, gage que le travail réalisé est conforme aux standards académiques les plus exigeants[90] [91].

La dernière méthode (méthode qualitative) est généralement appliquée à la fois par des chercheurs du milieu académique et des consultants du secteur privé. Les enjeux pour les équipes ne sont alors pas les mêmes en termes de publication dans des revues académiques. La qualité des travaux réalisés doit donc être garantie par d'autres processus (mise en débat des résultats).

[88] C'est le cas, par exemple, lorsque le tirage au sort dans le cadre d'une expérimentation n'est pas réalisé sur des individus, mais sur des ensembles, économiques ou géographiques (clusters). Dans ce cas les écarts-types associés aux effets moyens du traitement doivent eux-mêmes être estimés au niveau du cluster et non au niveau individuel, alors que les variables sont mesurées au niveau individuel.

[89] Pour réduire les biais, il est par exemple possible d'associer à la méthode des différences de différences un appariement préalable entre individus traités et non traités sur la base de leur score de propension. Pour une référence utile sur l'application de cette méthode hybride, voir Chabbé-Ferret (2014).

[90] Voir sur ce point, voir Ferracci et Wasmer, *op. cit*.

[91] Il peut cependant exister de réelles difficultés pour trouver des chercheurs acceptant de répondre aux appels à projet lancés par la puissance publique. En France, le milieu des chercheurs formés aux méthodes empiriques de l'évaluation d'impact des politiques publiques demeure assez concentré. Par ailleurs, les intérêts et les contraintes des chercheurs et des commanditaires ne sont pas toujours compatibles : pour avoir des chances de publier dans de bonnes revues, les premiers recherchent l'innovation dans les dimensions étudiées ou les méthodes appliquées ; *a contrario* les seconds souhaitent avant tout avoir des résultats robustes aux questions qu'ils se posent et dans le respect des délais impartis (ce qui peut impliquer de répliquer des travaux existants, réalisés dans d'autres contextes, lesquels sont plus difficilement valorisables dans des revues scientifiques).

CONCLUSION

Pour conclure, les conditions de succès d'une évaluation résident d'abord dans la compétence des évaluateurs, c'est-à-dire aussi bien leur expertise technique (de la politique étudiée et des méthodes) que leur capacité à interpréter les résultats à l'aune du contexte institutionnel et social dans lequel ils se trouvent.

L'indépendance des évaluateurs ou des commanditaires des évaluations est également un élément primordial de la crédibilité des résultats. L'indépendance du commanditaire est rarement absolue, mais son impartialité peut être assurée par différents moyens (mise en place de commissions d'évaluation plurielles dans leur composition et indépendantes dans l'élaboration de leur programme de travail et le choix des méthodes de travail ; transparence et publication des résultats, etc.).

Enfin, une condition essentielle au succès des évaluations est la possibilité pour les équipes d'accéder à des données microéconomiques et de pouvoir les apparier dans des délais raisonnables afin de transmettre les résultats aux décideurs en temps voulu.

Du point de vue opérationnel, ceci implique de réfléchir à une meilleure prise en compte des contraintes spécifiques à l'évaluation des politiques publiques dans les procédures d'accès aux données (mise à disposition accélérée de fichiers provisoires par les administrations, procédures d'accès facilitées au CNIS, etc.). La loi, lorsqu'elle instaure un nouveau dispositif, pourrait à cet égard prévoir à la fois l'évaluation de ce dispositif ainsi que les moyens nécessaires à sa bonne réalisation, notamment en termes de données et de leur mise à disposition aux évaluateurs par les services producteurs.

BIBLIOGRAPHIE

Aeberhardt R., Crépon B., Chiodi V., Gaini M. et Vicard A. (2014), « Du revenu contractualisé d'autonomie à la Garantie jeunes : les enseignements d'une expérimentation », In J. Bérard et M. Valdenaire, La Documentation française.

Bargain O. et Doorley K. (2013), "Putting Structure on the R&D Design: Social Transfers and Youth Inactivity in France", IZA Discussion Paper n°7508.

Belfield C.R., Nores M., Barnett, S. et Schweinhart L. (2006), "The High/Scope Perry Preschool Program Cost–Benefit Analysis Using Data from the Age-40 Followup", *Journal of Human Resources*, 41(1), p. 162-190.

Benoteau I. (2015), « Quels effets du recrutement en contrat aidé sur la trajectoire professionnelle? Une évaluation à partir du Panel 2008 », *Économie et Statistique*, n°477(1), p. 85-129.

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

Bonnal L., Fougere D. et Sérandon A. (1997), "Evaluating the Impact of French Employment Policies on Individual Labour Market Histories", *The Review of Economic Studies*, 64(4), p. 683-713.

Brodaty T., Crépon B. et Fougère D. (2001), "Using Matching Estimators to Evaluate Alternative Youth Employment Programs: Evidence from France, 1986–1988", In *Econometric Evaluation of Labour Market Policies*, p. 85-123, Physica-Verlag HD.

Burgess S.M. (2016), "Human Capital and Education: The State of the Art in the Economics of Education", IZA Discussion Paper n°9885, avril.

Card D., Kluve J. et Weber A. (2015), "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations", NBER Working Paper n°21431, National Bureau of Economic Research.

Chabbé-Ferret S. (2014), "Why Does Difference in Difference Matching Work?", miméo, août.

Chemin M. et Wasmer É. (2009), "Regional Difference-in-Differences in France Using the German Annexation of Alsace-Moselle in 1870-1918", In *NBER International Seminar on Macroeconomics* 2008 (p. 285-305), University of Chicago Press.

Commission nationale d'évaluation des politiques d'innovation (2016), *Quinze ans de politiques d'innovation en France*, rapport, France Stratégie, janvier.

Comité d'évaluation des expérimentations (2009), *Rapport final sur l'évaluation des expérimentations du RSA*, mai, La Documentation française.

Centre de Recherche pour l'Étude et l'Observation des Conditions de Vie (CREDOC) (2012), *Enquête qualitative auprès des jeunes bénéficiaires, un an après leur entrée dans le dispositif*, rapport final, décembre.

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

Crépon B., Ferracci M. et Fougère D. (2012), Training the Unemployed in France: How Does It Affect Unemployment Duration and Recurrence?, *Annals of Economics and Statistics*, p. 175-199.

Crépon B. et Gianella C. (2001), « Fiscalité et coût d'usage du capital : incidences sur l'investissement, l'activité et l'emploi », *Économie et Statistique*, n°341(1), p. 107-128.

Dehejia R.H. et Wahba S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating The Evaluation of Training Programs", *Journal of the American Statistical Association*, 94(448), p. 1053-1062.

D'Haultfœuille X. et Givord P. (2014), « La régression quantile en pratique », *Économie et Statistique*, n°471(1), p. 85-111.

Ferracci M. (2013), *Évaluer la formation professionnelle*, Presses de Sciences Po.

Ferracci M., Jolivet G. et van den Berg G.J. (2014), "Evidence of Treatment Spillovers within Markets", *Review of Economics and Statistics*, 96(5), p. 812-823.

Ferracci M. et Wasmer É. (2011), *État moderne, État efficace : évaluer les dépenses publiques pour sauvegarder le modèle français*, Éditions Odile Jacob.

Finkelstein A. (2005), “The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare”, NBER Working Paper n°11619, National Bureau of Economic Research.

Fitzenberger B. et Speckesser S. (2008), “Employment Effects of the Provision of Specific Professional Skills and Techniques in Germany”, In *The Education and Training* (p. 331-375), Part of the series Studies in Empirical Economics, Physica-Verlag HD.

Fougère F., Kramarz F. et Pouget J. (2009), “Youth Unemployment and Crime in France”, *Journal of the European Economic Association*, 7, p. 909-938.

Frölich M. (2004), “Programme Evaluation with Multiple Treatments”, *Journal of Economic Surveys*, vol. 18(2), p. 181-224.

Garibaldi P., Giavazzi F., Ichino A. et Rettore E. (2012), “College Cost and Time to Complete a Degree: Evidence from Tuition Discontinuities”, *Review of Economics and Statistics*, 94(3), p. 699-711.

Gilles L. et Guisse N. (2014), *Évaluation qualitative du revenu contractualisé d’autonomie (RCA)*, étude réalisée pour le ministère des Droits des femmes, de la ville, de la jeunesse et des sports, juillet, collection des rapports du CREDOC, n°309.

Givord P. (2015), « Méthodes économétriques pour l’évaluation de politiques publiques », *Économie et Prévision* n°204-205, p. 1-28.

Grunewald R. et Rolnick A.J. (2003), “Early Childhood Development: Economic Development with a High Public Return”, *The Region*, 17(4), supplement, p. 6-12.

Heckman J.J. (2010), “Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy”, *Journal of Economic Literature*, 48(2), p. 356-398.

Heckman J.J., Lochner L. et Taber C (1998), “General Equilibrium Treatment Effects: A Study of Tuition Policy”, *American Economic Review*, vol. 88(2).

Heckman J.J., Moon S.H., Pinto R., Savelyev P.A. et Yavitz A. (2010), “The Rate of Return to The Highscope Perry Preschool Program”, *Journal of Public Economics*, 94(1), p. 114-128.

Hirano K. et Imbens G. (2004), “The Propensity Score with Continuous Treatments”, in A. Gelman et X.-L. Meng, eds., *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perpectives*, New York: Wiley.

Quinet É. (2013), *Évaluation socioéconomique des investissements publics*, collection rapports & documents, Commissariat général à la stratégie et à la prospective, septembre.

Lechner M. (2002), "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies", *Review of Economics and Statistics*, 84(2), p. 205-220.

Lentile D. et Mairesse J. (2009), "A Policy to Boost R&D: Does the R&D Tax Credit Work?", *European Investment Bank Paper*, vol. 14, n°1, décembre.

L'Horty *et al*. (2014), « Les effets inattendus des aides au permis de conduire », In J. Bérard et M. Valdenaire, La Documentation française.

Lokshin B. et Mohnen P. (2012), "How Effective Are Level-Based R&D Tax Credits? Evidence from the Netherlands", *Applied Economics*, 44(12), p. 1527-1538.

Mission d'évaluation des politiques publiques (2008), « Calculer pour décider », dossier, *Cahiers de l'évaluation*, vol. 2, p. 9-11.

Mulkay B. et Mairesse J. (2013), "The R&D Tax Credit in France: Assessment and ex-ante Evaluation of the 2008 Reform", *Oxford Economic Papers*, vol. 65(3), p. 746-766, Oxford University Press.

Piketty T. et Valdenaire M. (2006), « L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français », *Les dossiers évaluations et statistiques* n°173, mars, ministère de l'Éducation nationale.

Rao V. et Woolcock M. (2003), "Integrating Qualitative and Quantitative Approaches in Program Evaluation", In F. Bourguignon & L. Pereira da Silva (Eds.), *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools* (chap. 8, p. 165-190), Oxford University Press.

Roux S. (2015), « Approches structurelles et non structurelles en micro-économétrie de l'évaluation des politiques publiques », *Revue française d'économie*, n°30 (1), p. 13-65.

SG-MAP (2015), « Mémo de l'évaluation des politiques publiques : principes, processus et méthode ».

Van den Berg G.J., Van der Klaauw B. et Van Ours J.C. (2004), “Punitive Sanctions and the Transition Rate from Welfare to Work”, *Journal of Labor Economics*, 22(1), p. 211-241.

Wasmer É. (2014), « Évaluation des politiques publiques : faut-il de l’interdisciplinarité ? », *Économie et Prévision* n°204-205, p. 193-209.

Directeur de la publication :
Jean Pisani-Ferry, commissaire général
Directeur de la rédaction :
Fabrice Lenglart, commissaire général adjoint
Contact presse :
Jean-Michel Roullé, directeur du service Édition-Communication-Évènements
01 42 75 61 37, jean-michel.roulle@strategie.gouv.fr