conseil d'analyse économique

Évaluation des politiques publiques

Les notes du conseil d'analyse économique, n° 1, février 2013

L'évaluation des politiques publiques est un exercice difficile techniquement et institutionnellement. Techniquement, parce que de nombreux pièges attendent l'évaluateur : corrélation (entre une politique et ses résultats) ne vaut pas causalité et l'évaluateur doit tenir compte des causalités inverses et des interactions de la politique considérée avec de multiples autres facteurs ; il doit aussi être conscient du fait que le bénéficiaire final d'un dispositif n'est pas forcément celui qui est visé, ou que la politique peut avoir de multiples effets, parfois loin du champ initialement ciblé. Plusieurs techniques statistiques permettent de contourner ces problèmes, la clé étant d'être capable de reconstruire ce qui se serait passé en l'absence de la politique considérée. Lorsqu'une véritable expérimentation n'est pas possible, les chercheurs exploitent les discontinuités existantes des politiques publiques, soit que la politique soit mise en place par vagues successives, soit qu'elle s'applique avec des seuils (on compare alors les individus ou entreprises de part et d'autre du seuil).

L'évaluation est aussi difficile à mettre en œuvre institutionnellement car seul un protocole rigoureux, défini si possible avant la mise en place de la politique, permet d'obtenir une évaluation crédible. Ce protocole doit garantir l'indépendance des évaluateurs et leur accès aux données nécessaires à l'évaluation. Il doit aussi prévoir un temps de discussion contradictoire des hypothèses et des résultats, dans un cadre interdisciplinaire. Il doit, enfin, laisser les évaluateurs libres de publier leurs résultats et de les discuter avec d'autres experts, en France comme à l'étranger.

En pratique, l'évaluation d'une politique ne doit pas être menée par l'administration en charge de la mettre en œuvre. L'expertise administrative est un complément indispensable à l'expertise technique, en particulier pour comprendre les modalités d'application de la politique et les interactions avec d'autres dispositifs. Elle doit être combinée à l'expertise technique mais ne saurait s'y substituer. Les évaluateurs extérieurs doivent être nommés selon un processus transparent et extérieur lui aussi à l'administration en charge, en veillant à éviter toute relation de dépendance avec les commanditaires et à promouvoir une pluralité des approches. De leur côté, les évaluateurs doivent respecter strictement la confidentialité des données et être parfaitement transparents sur leurs éventuels conflits d'intérêt.

Finalement, une évaluation crédible devrait reposer sur un triptyque formé d'un coordonnateur (Parlement, Cour des Comptes, Inspection générale des finances…), des administrations concernées et d'experts indépendants. Ces éléments sont à la portée d'un gouvernement décidé à faire le tri dans ses politiques publiques.

Si une évaluation crédible prend du temps, un diagnostic fiable et indépendant permet ultérieurement de gagner du temps au cours du processus de décision.

Introduction[1]

La Modernisation de l'action publique annoncée le 18 décembre 2012 prévoit que « toutes les politiques publiques, sur l'ensemble du quinquennat, feront l'objet d'une évaluation »[2]. De fait, l'accumulation des dispositifs au fil des décennies rend peu lisible aujourd'hui l'action publique et dissimule probablement des politiques obsolètes (les objectifs initiaux ont été atteints), inefficaces (les objectifs sont mal atteints ou à un coût trop important), ou détournés (servant *de facto* d'autres buts que ceux affichés). Le tout est coûteux pour les finances publiques et manque de transparence démocratique. Il est donc légitime de vouloir évaluer chaque politique une à une.

L'évaluation des politiques publiques est un exercice difficile : de nombreuses embûches attendent les évaluateurs, pouvant fausser et décrédibiliser une évaluation ne respectant pas un protocole rigoureux. Pourtant, une bonne évaluation est à la portée d'un gouvernement déterminé à faire le tri dans ses politiques. Après avoir présenté les pièges classiques de l'évaluation, nous exposons les méthodes permettant de les contourner pour obtenir une évaluation crédible des politiques publiques, en précisant les besoins notamment en termes de données statistiques. Enfin, nous dessinons les contours d'une bonne évaluation, qui doit associer les différents niveaux d'expertise dans des protocoles d'évaluation assurant l'indépendance et la pluralité des évaluateurs, ainsi que la diffusion et la discussion de leurs hypothèses et de leurs résultats.

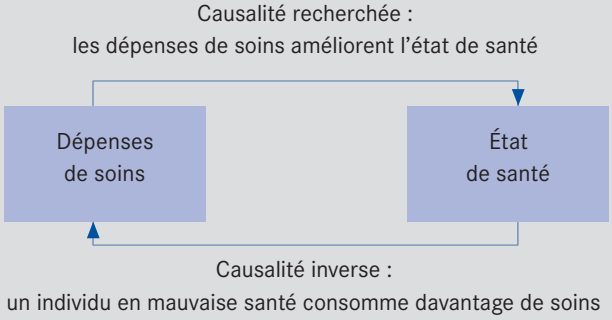
Les pièges de l'évaluation

Pour évaluer une politique publique, on ne peut se contenter d'observer l'évolution des indicateurs décrivant les objectifs visés par la politique. Nous présentons ici les pièges classiques de l'évaluation.

Identifier l'impact causal de la politique

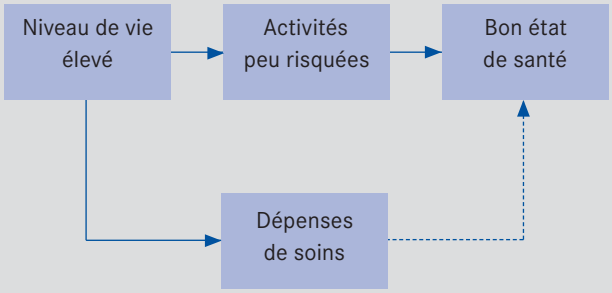
La première difficulté de l'évaluation est d'identifier une relation causale entre une politique et un résultat. Supposons, par exemple, qu'on souhaite évaluer l'impact des dépenses de soin sur l'état de santé d'une population. La simple corrélation, au sein de la population, entre dépenses de soins et état de santé est négative, car les individus qui dépensent le plus sont généralement les moins bien portants. Il s'agit ici d'une causalité inverse, de l'état de santé vers les dépenses, qui ne nous renseigne pas sur l'impact des dépenses (figure 1).

1. Un exemple de causalité inverse



L'identification d'une relation causale entre dépenses de soins et état de santé se heurte aussi à l'interférence de facteurs extérieurs, tels le niveau de vie, qui influent à la fois sur les dépenses et sur l'état de santé : les individus aisés dépensent plus pour leur santé et sont en général en meilleure santé, entre autres parce qu'ils exercent des activités moins risquées. La corrélation entre dépenses de soins et état de santé ne correspond alors à aucune relation causale (figure 2).

2. Un exemple de corrélation sans causalité



Pour évaluer une politique publique, on ne peut se contenter d'observer l'évolution des indicateurs décrivant les objectifs visés par la politique.

[1] Les auteurs remercient Clément Carbonnier qui a assuré le suivi de ce travail au sein de la cellule permanente du CAE.

[2] Déclaration de Jean-Marc Ayrault au Comité interministériel pour la modernisation de l'action publique, 18 décembre 2012, disponible sur http://www.gouvernement.fr/premier-ministre/declaration-de-jean-marc-ayrault-au-comite-interministeriel-pour-la-modernisation-d

Pour identifier l'impact des dépenses de soin sur l'état de santé des individus, il faudrait comparer non pas l'état de santé des « gros » consommateurs de soins à celui des « petits » consommateurs, mais les états de santé d'une même personne individuellement selon sa consommation de soins. Comme un même individu ne peut consommer à la fois beaucoup et peu de soins, il faut s'appuyer sur une multitude d'individus dont on contrôle finement toutes les caractéristiques pouvant influer sur l'état de santé, indépendamment des dépenses de soins. En estimant séparément les dépenses et les besoins de soins, Martin et al. (2008)[3] parviennent à mettre en évidence un impact causal positif : augmenter les dépenses de soins pour lutter contre le cancer et les maladies cardiovasculaires permet de sauver des vies. Les estimations montrent qu'en moyenne les soins contre le cancer peuvent faire gagner un an de vie pour 13 100 livres sterling, et ceux contre les maladies cardio-vasculaires un an de vie pour 8 000 livres.

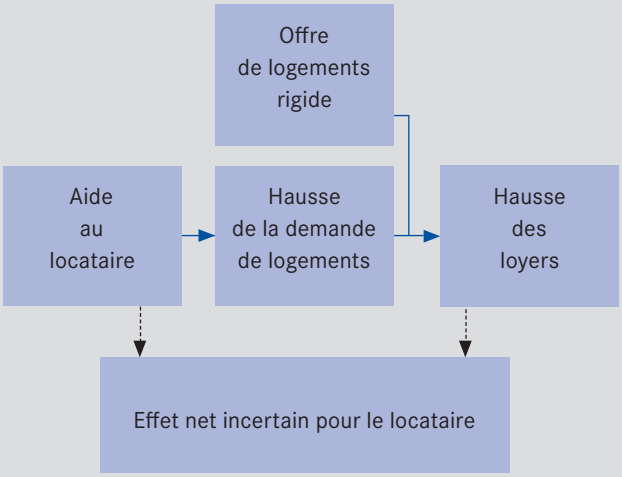
On se heurte aux mêmes types de problèmes lorsqu'on cherche à évaluer les politiques d'accompagnement au retour à l'emploi. Sans précaution, on peut trouver que ceux qui ont bénéficié de l'accompagnement ont mis plus de temps que les autres chômeurs à retrouver un emploi. Mais il faut tenir compte de ce que le personnel de Pôle emploi n'attribue pas forcément cet accompagnement au hasard : il peut cibler l'accompagnement sur les personnes les plus démunies en termes d'employabilité ou, à l'inverse, sur les personnes les plus proches de l'emploi (notamment si les employés de l'agence reçoivent une prime pour chaque chômeur ayant trouvé du travail). On dit qu'il y a un biais de sélection : les personnes accompagnées ne sont pas tirées au hasard dans la population des chômeurs ; de même, dans l'exemple précédent, les individus qui dépensent beaucoup pour leur santé ne sont pas tirés au hasard dans la population (ils sont généralement en plus mauvaise santé au départ).

La question de l'incidence

Le second problème auquel se heurte l'évaluation des politiques publiques est la question de l'incidence : le bénéficiaire final de la politique n'est pas forcément celui qui est visé. Ce second problème est fréquent lorsqu'il s'agit de taxation ou de subventions/transferts. La théorie de l'incidence fiscale montre ainsi que l'impôt ne pèse pas nécessairement, *in fine*, sur la personne qui rédige le chèque : les agents imposés peuvent transférer la charge de l'impôt sur d'autres ; à l'inverse, des individus non ciblés originellement par une subvention peuvent se retrouver indirectement les véritables bénéficiaires.

Une illustration est donnée par le cas des allocations logement. Ces aides aux locataires représentaient en 2009 un quart des prestations aux ménages. Si les ménages officiellement destinataires de ces aides en étaient les bénéficiaires réels, les allocations logement contribueraient pour plus d'un cinquième dans la réduction des inégalités de niveau de vie réalisée par le système socio-fiscal français[4]. Cette redistributivité est pourtant mise en doute par Fack (2005)[5] qui estime qu'entre 50 et 80 % de ces allocations ont en réalité profité aux bailleurs à travers des hausses de loyers par l'intermédiaire des mécanismes de marché (offre et demande de logement) : parce que l'offre locative évolue peu à court et moyen terme, la solvabilisation de la demande a fait monter les loyers, de sorte que la subvention pensée pour les locataires peu fortunés s'est retrouvée en partie captée par les propriétaires de logements à louer. Cet exemple montre qu'on ne peut évaluer une politique de ce type en se limitant au simple taux de recours ou au nombre de ménages bénéficiaires. Une analyse des allocations logement sans prise en compte de leur incidence sur les loyers conduirait à un diagnostic trop optimiste (figure 3).

3. La nécessaire prise en compte de l'incidence : l'exemple des aides au logement



[3] Martin S., N. Rice et P. Smith (2008) : « Does Health Care Spending Improve Health Outcomes? Evidence from English Programme Budgeting Data », *Journal of Health Economics,* n° 27, pp. 826-842.

[4] Chanchole M. et G. Lalanne (2011) : *Photographie du système socio-fiscal et de sa progressivité*, Rapport particulier pour le Conseil des prélèvements obligatoires.

[5] Fack G. (2005) : « Pourquoi les ménages à bas revenus paient-ils des loyers de plus en plus élevés ? L'incidence des aides au logement en France (1973-2002) », *Économie et Statistique*, n° 381-382.

1. Les effets multiples d'une politique

Dans les années 1970, un groupe de chercheurs a organisé en Californie une expérimentation en proposant des contrats d'assurance santé à un échantillon de population*. Ces assurances différaient entre elles notamment par le niveau de reste à charge et de franchise. En suivant les ménages pendant une période allant jusqu'à cinq ans, ils ont découvert que laisser un reste à charge à l'assuré était efficace pour réduire le problème de sur-médication et était une façon efficace de limiter les dépenses de santé inutiles.

L'étude a cependant mis en évidence les effets pervers possibles d'un tel dispositif, qui certes contenait les dépenses inutiles, mais réduisait également un certain nombre de dépenses utiles pour les plus démunis. Ainsi, en se limitant aux 20 % de ménages ayant les revenus les plus bas et sujets à de l'hypertension, les auteurs ont montré une divergence significative entre les tensions artérielles de ceux bénéficiant d'une couverture intégrale et ceux soumis à du reste à charge. Au vu de l'impact de l'hypertension artérielle sur la probabilité de survenue future de maladies cardiovasculaires graves, ce phénomène ne peut être ignoré au prétexte d'une réussite du dispositif de reste à charge quant à son objectif assigné, à savoir réduire les dépenses de santé inutiles.

* Newhouse J. (1993) : *Free For All? Lessons from the RAND Health Insurance Experiment*, Harvard University Press, Cambridge.

Par ailleurs, l'effet d'incidence peut détériorer la situation d'individus ne bénéficiant pas de la politique. C'est le cas ici des locataires ne recevant pas d'aides au logement (ils souffrent de la hausse des loyers) – sans parler des recettes fiscales supplémentaires qui sont nécessaires pour financer la politique.

La multiplicité des effets

Une troisième difficulté de l'évaluation réside dans la multiplicité des effets d'une même politique publique. Les effets multiples peuvent tout d'abord survenir dans le même champ d'action publique. Par exemple, introduire une franchise médicale peut être efficace pour lutter contre la sur-médication mais en même temps entraîner un non-recours aux soins préoccupant (encadré 1). Dans d'autre**s** cas, les effets multiples apparaissent dans des champs différents de l'action publique. Par exemple, une hausse de l'impôt sur le bénéfice des sociétés peut entraîner une baisse de la rémunération non pas du capital, mais du travail[6].

Lorsque l'effet indésirable d'une politique se produit hors du champ d'investigation de l'évaluateur ou hors des variables d'intérêt de son étude, voire de son champ disciplinaire, il risque fortement d'être négligé. C'est pourquoi le pluralisme des évaluateurs, en termes de discipline aussi bien que de sensibilité, est primordial pour comprendre l'ensemble des conséquences d'une politique publique sans se restreindre à celles étroitement liées à l'objectif initialement poursuivi.

2. Les paramètres opérationnels de la politique publique

Lors d'une expérimentation visant à améliorer le recours à la complémentaire santé par une majoration du chèque santé, il s'est avéré que le protocole avait une incidence sur les résultats. En particulier, l'invitation à une réunion d'information préliminaire avait été un facteur de découragement des participants. Ce résultat paradoxal et imprévu aurait pu fausser les conclusions de l'étude, si les chercheurs*, en collaboration avec la Caisse primaire d'assurance-maladie de Lille, n'avaient pas conçu leur protocole en faisant varier les conditions et en répartissant par tirage aléatoire les personnes concernées selon plusieurs groupes, avec et sans majoration du chèque santé, mais aussi avec et sans réunion d'information préalable. Il est apparu que l'invitation à une réunion d'information pouvait avoir des effets négatifs sur le recours à une assurance santé complémentaire. L'hypothèse des auteurs est que, bien que celle-ci soit facultative, les individus n'ayant pu assister à cette réunion alors qu'ils étaient invités se sont sentis illégitimes à recourir au chèque santé. Des mécanismes potentiellement vertueux peuvent ainsi être dénaturés par leur mise en œuvre concrète.

* Guthmuller S., F. Jusot et J. Wittwer (2011) : « Improving Take-up of Health Insurance Program: A Social Experiment in France », *Cahiers de la Chaire Santé*, n° 11, Université Paris-Dauphine.

[6] Voir Arulampalam W., M.P. Devereux et G. Maffini (2010) : « The Direct Incidence of Corporate Income Tax on Wages », *IZA Working Paper*, n° 5293.

La mise en œuvre concrète

Enfin, le problème de la mise en œuvre concrète peut influer fortement sur l'efficacité d'une politique. Dans certains cas, il pourrait être tentant de rejeter comme inefficace un dispositif qui a seulement été appliqué dans des conditions inopportunes. Ainsi, la tarification hospitalière à l'activité (T2A) n'a probablement pas donné les résultats escomptés en France, non en raison d'une conception erronée, mais du fait d'une classification excessivement fine des actes hospitaliers et d'une interaction défavorable avec d'autres dispositifs[7]. Pour résoudre ces problèmes, il pourrait s'avérer pertinent de réaliser, lorsque c'est possible, une expérience préliminaire à petite échelle pour mieux déterminer les paramètres opérationnels déterminants pour le succès d'une politique publique (voir l'encadré 2). Cette expérimentation n'entame en rien le fait qu'il faudra *a posteriori* évaluer précisément la réforme dans sa globalité. Au contraire, l'expérimentation préalable peut être propice à la construction d'un protocole afin d'évaluer ultérieurement les politiques. En particulier, il convient d'harmoniser les contraintes juridiques et administratives avec le respect des mécanismes économiques visés en faisant dialoguer les administrations avec les experts, et non, comme c'est le cas actuellement, en faisant intervenir séquentiellement les différentes étapes de la décision.

Les méthodes de l'évaluation

Une bonne évaluation se conçoit en principe *avant* la mise en place d'une politique publique, pour trois raisons. Premièrement, il est nécessaire de s'appuyer sur une anticipation des impacts attendus et, si possible, une expérimentation. Deuxièmement, il faut, comme on l'a mentionné ci-dessus, déterminer finement les modalités de mise en œuvre. Troisièmement, il faut fixer les modalités d'évaluation *a posteriori* de la politique. Cette dernière forme d'évaluation sera d'autant plus précise qu'elle aura été préparée en amont.

L'évaluation idéale consisterait à comparer la situation issue de la politique publique à une situation hypothétique qui serait intervenue si la politique n'avait pas vu le jour, tous les autres éléments de l'environnement socio-économique étant les mêmes – une situation hypothétique dite « contrefactuelle »[8]. Or, ceci est impossible. La situation issue de la politique publique est observable, ce qui n'est pas le cas du « contrefactuel ». La difficulté de l'évaluation est donc de reconstituer ce qui se serait passé en l'absence de la politique publique : il faut construire théoriquement ou empiriquement cette situation ou bien constituer un groupe de référence – le « contrefactuel ».

L'évaluation idéale consisterait à comparer la situation issue de la politique publique à une situation hypothétique qui serait intervenue si la politique n'avait pas vu le jour.

L'expérience aléatoire

Comme on l'a vu, une difficulté majeure de l'évaluation est liée au fait que les individus ou les entreprises visés par une politique publique ne sont pas tirés au hasard dans la population : ils sont, par exemple, en moins bonne santé, ou plus éloignés de l'emploi que la moyenne. Une manière de contourner ce problème est de réaliser une expérience aléatoire : on tire au sort un groupe d'individus ou d'entreprises qui se verra appliquer une politique, tandis qu'un autre groupe constituera le groupe de contrôle. Le tirage aléatoire sur une population suffisamment importante permet de s'assurer que les groupes de contrôle et de traitement sont comparables : ce ne sont pas des caractéristiques individuelles qui ont permis d'obtenir le traitement.

L'expérimentation citée en encadré 1 est une expérience aléatoire : les expérimentateurs ont proposé une assurance gratuite à un large panel de ménages californiens. Les participants n'ont pas eu le loisir de choisir le type du contrat d'assurance qui leur était offert. Celui-ci, et en particulier le fait qu'il s'agissait d'une assurance totale ou laissant un reste à charge, était tiré au sort.

Une expérimentation aléatoire peut s'avérer coûteuse, même si la précision des résultats et les économies budgétaires qu'ils peuvent permettre en font souvent un investissement rentable. Le coût et la complexité de l'expérience aléatoire sont à la hauteur des résultats qu'elle peut apporter. C'est ainsi que l'expérience aléatoire mentionnée en encadré 1, réalisée dans les années 1970, sert encore aujourd'hui de référence alors même que les comportements se sont modifiés depuis cette époque.

Par ailleurs, l'expérience aléatoire soulève dans certains cas des problèmes éthiques[9]. Certains domaines sont peu propices à l'expérience aléatoire pour des raisons d'équité il est par exemple inconcevable (et anticonstitutionnel) d'étudier l'effet d'une réforme fiscale en soumettant au hasard divers contribuables à des impôts différents, d'autres parce qu'ils peuvent exposer des sujets vulnérables.

[7] Voir Saint-Paul G. (2012) : *Réflexions sur l'organisation du système de santé*, Rapport du CAE, n° 103, la Documentation française et le « Commentaire » de B. Dormont dans le même volume.

[8] Cet idéal est celui de la médecine où l'on teste une thérapie à partir de deux groupes d'individus comparables dont un seul se voit appliquer la thérapie.

[9] Voir sur ce sujet l'avis du Comité d'éthique du CNRS (COMETS) sur l'expérimentation sociale : http://www.cnrs.fr/comets/IMG/pdf/07-experimentation-sociale-20100119-2.pdf

3. Les méthodes économétriques dans le cadre des expériences « naturelles »

La double différence

Ne pouvant comparer des individus identiques dans deux mondes différents (avec et sans la politique publique), il faut se contenter de comparer les individus traités avant et après leur traitement (on est alors sujet aux biais de conjoncture) ou les individus traités et non traités après le traitement des premiers (on est alors sujet aux biais de sélection). Le principe de l'évaluation en double différence consiste à associer les deux approches. On rassemble les individus dans un groupe de traitement (ceux dont la situation est censée avoir été modifiée par la politique publique) et un groupe de contrôle (ceux dont la situation n'a pas été modifiée). On compare ensuite l'évolution de ces deux groupes, le groupe de contrôle servant de contrefactuel au groupe de traitement.

Le relèvement de plafond de la réduction d'impôt pour l'emploi à domicile décidé en 2002 peut servir d'exemple. Une simple comparaison des déclarations d'emplois à domicile avant et après ce relèvement pourrait faire croire à une efficacité substantielle de la mesure. Cependant, celle-ci est intervenue en plein développement de ces services et le relèvement de plafond a été concomitant d'autres mesures d'incitations (baisses de cotisations sociales), de simplifications administratives (chèque emploi service simplifié) et de l'entrée des entreprises dans un marché alors quasi exclusivement composé de travailleurs individuels. La prise en compte d'un contrefactuel (en l'occurrence les ménages non touchés parce que situés précédemment en dessous de l'ancien plafond ou au-dessus du nouveau plafond) permet d'isoler l'effet spécifique de la mesure puisque les membres du groupe de contrôle sont tout autant que ceux du groupe de traitement touchés par les autres mesures incitatives. On trouve alors que le relèvement de plafond a effectivement augmenté la demande de services à domicile mais n'est responsable que marginalement du développement de ce secteur[a].

Des méthodes d'assortiment peuvent améliorer encore l'évaluation. Elles consistent à repérer des individus semblables à l'intérieur des groupes de contrôle et de traitement. La comparaison des évolutions des variables d'intérêt ne se fait alors plus globalement entre les groupes de contrôle et de traitement, mais individuellement entre les sous-ensembles assortis tirés de ces groupes.

La régression autour d'une discontinuité

Une autre méthode consiste à repérer une discontinuité dans le droit au traitement et à ne réaliser l'évaluation qu'à ce niveau : c'est le principe de régression autour d'une discontinuité. Fack et Grenet (2010) ont utilisé cette méthode pour estimer la disposition à payer pour l'éducation, à partir d'une discontinuité de la carte scolaire[b]. Le prix d'un logement au mètre carré dépend du quartier et de la qualité du logement ; il est relativement stable à localisation et qualité égales. Ainsi, en appariant des appartements de qualité identique, de part et d'autre d'une même rue – donc dans des quartiers identiques – mais dont l'adresse envoie les enfants dans des écoles différentes du fait de la carte scolaire, ils parviennent à mesurer le supplément de prix que sont prêts à payer les parents pour envoyer leurs enfants dans une école plutôt que dans une autre.

Le principe de la régression autour d'une discontinuité consiste ainsi à comparer non plus l'ensemble des individus traités ou non, mais seulement ceux très proches du seuil décidant de l'assignation entre les deux groupes. En supposant que les caractéristiques des individus sont continues, les individus tout proches du seuil d'un côté (et ainsi non traités) sont identiques et donc comparables aux individus très proches du seuil de l'autre côté (et donc traités).

Un autre exemple est l'usage qui a été fait par Piketty et Valdenaire (2006) des seuils d'ouverture de classes pour estimer l'impact de la taille des classes sur la réussite scolaire[c]. La taille des classes n'est pas déterminée au hasard : elle

[a] Carbonnier C. (2010) : « Réduction et crédit d'impôt pour l'emploi d'un salarié à domicile, conséquences incitatives et redistributives », *Économie et Statistique*, n° 427-428, pp. 67-100.

[b] Fack G. et J. Grenet (2010) : « When do Better Schools Raise Housing Prices? Evidence from Paris Public and Private Schools », *Journal of Public Economics*, n° 94, pp. 59-77.

[c] Piketty T. et M. Valdenaire (2006) : « L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français. Estimations à partir du panel primaire 1997 et du panel secondaire 1995 », *Les Dossiers Enseignement Scolaire*, n° 173, ministère de l'Éducation nationale.

n'est pas la même en ville et en zone rurale, et au sein d'une même école les enfants ne sont pas répartis au hasard dans les classes. Ainsi il est difficile de déterminer l'impact de la taille des classes sur la réussite scolaire. Les deux auteurs exploitent la règle selon laquelle une classe de CE1 ne peut dépasser 30 élèves : lorsqu'un nouvel élève arrive (événement aléatoire) dans une cohorte de 30 élèves, il se crée une classe supplémentaire, et les écoliers apprennent donc dans des classes de 15 ou 16 élèves. Cet événement crée une discontinuité qui peut être exploitée pour mesurer l'impact de la taille des classes sur les résultats scolaires.

La méthode des variables instrumentales

Une dernière méthode consiste à trouver une variable dite « instrumentale » pour délimiter les groupes de contrôle et de traitement. Il s'agit d'une variable fortement corrélée avec le fait d'être « traité » (par la politique publique), mais sans influence directe sur le paramètre d'intérêt (le résultat de la politique) et non manipulable par les individus. Cette méthode a été utilisée pour estimer l'impact de la maternité sur la participation au marché du travail, ce qui est utile pour calibrer des politiques d'encouragement à l'activité des mères. Le problème ici est que le choix du nombre d'enfants est influencé par le statut de la mère sur le marché du travail (emploi, chômage, inactivité). Pour contourner ce problème, Angrist et Evans (1998)[d] ont séparé un groupe homogène de femmes ayant au moins deux enfants selon que les deux premiers enfants sont de même sexe ou de sexes différents. Il n'y a *a priori* aucune influence directe de cette variable « instrumentale » binaire (même sexe, sexes différents) sur la participation des femmes au marché du travail. En revanche, les femmes dont les deux premiers enfants sont de même sexe ont plus souvent que les autres un troisième enfant, et ce pour des raisons exogènes et non du fait de caractéristiques individuelles différentes ni de leur statut sur le marché du travail. Les auteurs ont alors observé que les femmes dont les deux premiers enfants sont de même sexe participent significativement moins au marché du travail que les femmes avec deux enfants de sexes différents, ce qu'ils ont interprété comme l'effet causal du fait d'avoir un troisième enfant.

[d] Angrist J. et W. Evans (1998) : « Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size », *The American Economic Review*, n° 88, pp. 450-477.

Si l'expérience aléatoire pure peut se révéler difficile et parfois coûteuse, des formes approchées donnent aussi des résultats très satisfaisants. Il s'agit, au moment où une politique publique est décidée, de ne pas la mettre en place d'un seul bloc sur tout le territoire mais d'échelonner sa mise en œuvre en plusieurs vagues, par exemple par groupes de départements. Il convient alors de choisir les départements de chaque vague de manière à ce que chacune soit la plus comparable possible aux autres.

Ce type d'expérimentation avait été prévu pour le remplacement du Revenu minimum d'insertion (RMI) par le Revenu de solidarité active (RSA). Le dispositif a d'abord été appliqué dans l'Eure, puis dans 25, 34 et enfin 40 départements avant d'être généralisé à l'ensemble du territoire. Cependant, l'expérimentation n'a pas été effectuée avec suffisamment de constance et les évaluations n'ont pas été à la hauteur des attentes.

L'expérience aléatoire peut donner des résultats très fiables à condition d'avoir été préparée à l'avance, soit par la définition de groupes tests, soit par la mise en place séquentielle de la politique. Lorsque cela n'a pas été le cas, il faut envisager d'autres méthodes d'évaluation.

L'expérience naturelle

L'expérience « naturelle »[10] consiste à comparer des groupes d'individus (ou d'entreprises) qui se trouvent séparés de manière non intentionnelle en termes d'accès à la politique considérée. Les individus exclus du bénéfice de la politique publique servent de « contrefactuel » aux bénéficiaires effectifs.

La difficulté de l'évaluation en expérience naturelle est la validité du « contrefactuel », c'est-à-dire la comparabilité des groupes de traitement et de contrôle : une similitude apparente des groupes comparés n'exclut pas la présence de biais dans l'évaluation. De nombreuses techniques économétriques ont de ce fait été développées pour s'assurer de la comparabilité des groupes de traitement et de contrôle. L'encadré 3 présente certaines de ces techniques.

Les limites des expériences aléatoires et naturelles

Les techniques décrites plus haut sont relativement récentes et leurs capacités explicatives sont robustes. En revanche, leurs qualités prédictives sont contestées au motif que les comportements sont dépendants de l'environnement socio-

[10] Le qualificatif « naturelle » ne désigne pas nécessairement un lien avec la nature mais un caractère simplement non intentionnel.

économique sans cesse mouvant. Ainsi, les comportements peuvent différer entre la réaction à une expérimentation à petite échelle et la réaction à la mise en place réelle d'une politique, du fait que cette dernière modifie plus profondément le cadre économique. Par exemple, l'expérience aléatoire de la RAND *corporation* (encadré 1) examine l'effet d'une assurance sur les consommations de soins d'un faible nombre de sujets. Si l'on considère une assurance publique de plus grande ampleur comme Medicare[11], on peut trouver une augmentation des dépenses de soins bien plus importante : le fait que l'assurance touche plus de personnes et augmente substantiellement les ressources financières du secteur induit une réaction du côté de l'offre de soins, la construction de nouveaux hôpitaux et une augmentation de la recherche médicale. Des méthodes mêlant l'estimation en expérience naturelle et des modélisations plus globales sont en cours de développement pour corriger ce handicap[12].

La nécessité de données fiables

Pour réaliser les évaluations, il est impératif de disposer de bases de données à la fois exhaustives et fiables. Les chercheurs ont rarement les moyens de réaliser eux-mêmes des enquêtes. Heureusement, les données nécessaires existent déjà pour la plupart dans différentes bases administratives. Il est donc important de mettre en place des institutions et des procédures afin que les chercheurs puissent exploiter ces données tout en préservant les droits des individus et des entreprises dont les informations sont consignées dans ces bases. Des protocoles sûrs existent déjà, comme l'accès sécurisé aux données (ASD), nécessitant un accord par étude de la part de la Commission du secret statistique. Cependant, deux lacunes se font encore cruellement sentir.

La première est que de nombreuses données restent inaccessibles, notamment celles d'assurance-maladie et les données fiscales. Or, celles-ci sont primordiales à nombre d'évaluations, et ce pour plusieurs raisons. Tout d'abord, les politiques fiscales sont nombreuses et encore peu évaluées par manque de données. Un accès aux données fiscales permettrait un grand progrès concernant ce type de politiques. De plus, du fait de leur richesse, les données fiscales peuvent être mobilisées pour évaluer des politiques non fiscales. Il faut ainsi permettre le travail des évaluateurs sur ces bases tout en garantissant le secret fiscal aux contribuables. Cela passe par des protections telles que l'ASD et par l'anonymisation des bases, tout en préservant un codage des observations pour garder la possibilité de constituer des panels. Ceci est aisé techniquement et peu coûteux.

La deuxième lacune concerne la possibilité d'apparier les fichiers administratifs ou d'enquêtes. En effet, mêmes si l'ensemble des informations contenues dans les bases de données accessibles aux évaluateurs s'élargissait, celles-ci ne seraient pas toujours utilisables. Pour étudier le comportement des mères en termes d'offre de travail, par exemple, il est nécessaire d'avoir des informations sur les enfants d'une femme, ce que l'on trouve dans une base de données, et sur sa participation au marché du travail, ce que l'on trouve dans une autre. Si l'on n'est pas capable d'apparier les bases, les informations contenues dans chaque base se révéleront inutiles. Il existe des moyens simples, fiables et peu coûteux de réaliser les appariements tout en respectant l'anonymat des données.

Comment comparer différentes politiques publiques ?

L'évaluation d'une politique publique peut conduire à une conclusion tranchée : la politique est inefficace au regard de l'objectif assigné, voire elle est contre-productive. Souvent, cependant, le jugement est plus nuancé : la politique est efficace, mais elle semble coûteuse au regard des résultats obtenus. Pour arbitrer avec d'autres politiques, notamment celles qui agissent dans des champs différents de l'action publique, il faut alors convertir ses bénéfices dans une métrique qui les rende comparable à la fois aux coûts et aux bénéfices des autres politiques publiques. En pratique, il faut affecter une valeur monétaire à des bénéfices non monétaires comme la qualité de l'air, la longévité ou la santé.

Ceci peut choquer de prime abord, mais constitue le seul moyen de rendre explicites les critères utilisés pour la décision publique. Ces valeurs monétaires peuvent être définies de manière tutélaire, comme cela a souvent été le cas en matière de sécurité routière. Toutefois, il est préférable de chercher à repérer les préférences des individus à partir des enquêtes où ils expriment leur disposition à payer pour une amélioration de la qualité de l'eau par exemple[13].

Les structures de l'évaluation

L'évaluation des politiques publiques n'est pas qu'une affaire de données et d'expertise technique. Les politiques qui sont

[11] Finkelstein A. (2007) : « The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare », *The Quarterly Journal of Economics*, n° 122, pp. 1-37.

[12] Attanasio O., C. Meghir et A. Santiago (2012) : « Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA », *Review of Economic Studies*, n° 79, pp. 37-66.

[13] Un point très important est la façon dont on synthétise les dispositions à payer des individus. De nombreux travaux, en environnement, mais aussi en santé, montrent que les décisions publiques peuvent être fortement modifiées selon les pondérations retenues. Voir Anthoff D., C. Hepburn, R.S.J. Tol (2009) : « Equity Weighting and the Marginal Damage Costs of Climate Change », *Ecological Economics*, n° 68, pp. 836-849 ; Fleurbaey M., S. Luchini, E. Schokkaert et C. Van de Voorde (2013), « Evaluation des politiques de santé : pour une prise en compte équitable des intérêts des populations », *Économie et Statistique*, à paraître.

évaluées sont souvent complexes et elles opèrent généralement une redistribution au sein de la société. Ces caractéristiques imposent d'organiser l'évaluation avec beaucoup de rigueur aux différents niveaux d'expertise.

Expertise technique et expertise administrative

Si l'expertise technique est indispensable pour déjouer les pièges de l'évaluation mentionnés plus haut, on ne saurait se passer de l'expertise administrative sur la mise en œuvre concrète des politiques et le fonctionnement des établissements publics ou administrations qui les gèrent. L'expertise administrative permet non seulement de construire la stratégie d'identification et les scénarios contrefactuels, mais également de discerner, dans les résultats, ce qui relève du principe général de la politique et ce qui est le fait de sa mise en œuvre concrète.

Ces deux types d'expertises – technique et administrative – doivent collaborer non seulement au cours de la phase d'évaluation proprement dite, mais également et surtout en amont, si possible avant la mise en œuvre de la politique évaluée. Cela doit permettre d'adapter certaines modalités de la loi pour la rendre plus précisément évaluable. C'est également au cours de cette phase de coordination *a priori* que peut se décider, suivant les possibilités légales, une mise en œuvre échelonnée de la mesure afin de construire dès avant l'évaluation un contrefactuel pertinent. Enfin, cette coordination *a priori* pourrait permettre d'éviter les défauts de mise en œuvre susceptibles de rendre inefficace une politique publique en principe bénéfique.

L'indépendance des évaluateurs

Aussi rigoureuse soit-elle, une évaluation reste sujette à incertitude scientifique : les résultats sont conditionnels à la validité des méthodes (choix du contrefactuel, généralisation des résultats…). Or, pour que l'évaluation des politiques publiques soit utile, il importe que ses résultats soient crédibles : que les hypothèses soient présentées de manière transparente, sans que l'on puisse soupçonner que certaines ont été cachées. La transparence et la crédibilité nécessitent l'indépendance des évaluateurs. La difficulté est alors de faire collaborer les partenaires institutionnels et scientifiques tout en préservant l'indépendance de l'évaluation. Il existe des conflits d'intérêts évidents lorsque l'évaluation est réalisée par les administrations, ministères, directions ou établissements publics en charge de concevoir ou d'appliquer une politique publique. Une même institution ne peut être à la fois juge et partie. Ce n'est cependant pas le seul problème d'indépendance et il faut veiller à ne pas créer une dépendance de fait au cours du processus de désignation des évaluateurs, ou en bloquant la publication des résultats.

Le temps de l'évaluation n'est pas le temps du politique. Plusieurs raisons imposent cette divergence de calendrier. Tout d'abord, la plupart des méthodes d'évaluation nécessitent des données longitudinales. Il faut donc attendre de réunir suffisamment de données pour s'assurer de la validité de l'évaluation. De plus, l'évaluation elle-même prend du temps, de la nomination des évaluateurs à la discussion des résultats, en passant par les choix méthodologiques et le travail statistique. Dans ce domaine, la précipitation est bien souvent l'ennemie de la précision et de l'exhaustivité. En particulier, les hypothèses doivent être soumises à une critique approfondie des pairs. Cette méthode de validation *a posteriori* des résultats, qui est la méthode proprement scientifique, ne doit pas être négligée dans le cas de l'évaluation des politiques publiques[14]. Précipiter une évaluation revient à en réduire la crédibilité, ce qui rend plus difficile (et potentiellement plus longue) la décision publique.

La diffusion des résultats

La liberté de diffusion des résultats est une condition clé de l'indépendance des évaluateurs. En particulier, l'accès aux données ne doit pas être conditionné à un droit de regard par l'administration dépositaire des données. Toute pratique allant à l'encontre de cette liberté de diffusion, en mettant une pression sur l'évaluateur quant à ses résultats, contreviendrait à la nécessité d'indépendance de celui-ci, outre le fait qu'elle exclurait tout débat scientifique sur la méthode et les résultats.

La liberté de diffusion des résultats est une condition clé de l'indépendance des évaluateurs.

La diffusion des résultats doit s'accompagner d'une confrontation avec d'autres évaluations issues, le cas échéant, d'autres champs disciplinaires (cf. infra). Cette confrontation doit s'opérer à la fois par la publication des différents résultats et de leurs critiques, mais également par l'organisation de débats, voire de conférences de consensus. Ceci, afin de permettre une meilleure information des citoyens et une meilleure compréhension des divers effets d'une politique publique. À cette fin, des méthodes de hiérarchisation de la robustesse des résultats peuvent être mobilisées (preuve scientifiquement établie, présomption scientifique, faible niveau de preuve…).

[14] Cette phase de discussion par les pairs garantit l'indépendance du choix de la méthode ; elle facilite aussi une séparation claire entre l'évaluation et la décision politique. Voir l'avis du Comité d'éthique du CNRS, *op. cit*.

4. Exemples d'organisation de l'évaluation à l'étranger

Différentes structures ont été mises en place pour coordonner l'évaluation des politiques publiques à l'étranger. Une comparaison éclairante est celle de l'Institute for Fiscal Studies (IFS) au Royaume-Uni et du Government Accountability Office (GAO) aux États-Unis[a]. Le GAO n'est pas indépendant par nature, ni porté initialement vers l'évaluation. Il dépend directement de l'État fédéral et sa mission à sa création en 1921 était l'audit des finances des agences gouvernementales. L'audit est très différent de l'évaluation des politiques publiques, mais au cours du temps, cette institution a vu ses compétences s'élargir. De nombreux chercheurs en sciences sociales, ainsi que des collaborations avec des universitaires, sont venus compléter les capacités principalement juridiques du GAO. Son objectif est d'informer le Congrès et les citoyens sur les actions du gouvernement, afin de permettre au Congrès d'effectuer au mieux son rôle législatif, et de pouvoir si besoin s'opposer de manière éclairée au pouvoir exécutif. Il le fait notamment en contrôlant les évaluations ministérielles selon des critères scientifiques (validité des contrefactuels) et institutionnels (séparation du commanditaire et de l'évaluateur, indépendance de ce dernier, publication automatique des résultats). Afin d'assurer son indépendance, le GAO est dirigé par le « contrôleur général des États-Unis » dont le mandat est long (15 ans), incompressible et non renouvelable.

La transposition au cas français ne serait toutefois pas aisée, et rien n'assure qu'un tel dispositif serait tout autant indépendant. Le Royaume-Uni, qui possède comme la France un exécutif fort s'appuyant sur une puissante administration, a mis en place un tout autre système. Ainsi l'IFS, qui a un statut d'association non gouvernementale, est indépendant par nature. Afin d'éviter la dépendance envers des intérêts constitués, le financement repose sur une multiplicité de subventions d'institutions et d'entreprises. La seule subvention essentielle provient de l'Economic and Social Research Council (ESRC), l'agence publique chargée de financer la recherche en sciences sociales dans le pays. La compétence scientifique est assurée par le recrutement de chercheurs en sciences sociales et des collaborations de long terme avec des universitaires. Les missions sont principalement centrées sur l'évaluation des politiques publiques et l'explication des dispositifs dont la complexité nuit à la transparence. Les résultats servent à conseiller les membres du Parlement, voire du Gouvernement, ainsi que divers groupes de la société civile. Enfin, l'IFS s'impose une mission importante vis-à-vis du grand public, avec des publications didactiques dans les médias.

L'Australie présente également un cas riche d'enseignements. Regrettant que le contrôle des dépenses prenne le pas sur l'évaluation des performances, le gouvernement australien a tenu à inculquer à ses ministères une véritable culture de l'évaluation à partir de la fin des années 1980. Chaque ministère devait remettre au ministère des Finances un plan d'évaluation annuel, permettant d'évaluer l'intégralité de ses politiques tous les trois à cinq ans. Les résultats étaient rendus publics. Le ministère des finances, et surtout l'Australian National Audit Office (ANAO) contrôlaient ces pratiques d'évaluations. En particulier, l'ANAO servait à la fois d'organe de conseil en évaluation et évaluait lui-même la qualité des évaluations. Il en a résulté une évaluation effective des politiques et une importante prise en compte des résultats dans les propositions de politiques nouvelles. Cependant, l'ANAO a constaté, dans son rapport de 1997[b], que la communication autour des méthodes et des résultats des évaluations menées par les ministères en charge des politiques était insuffisante.

[a] Pour une discussion approfondie, voir Ferracci M. et É. Wasmer (2011) : *État moderne, État efficace*, Odile Jacob.

[b] ANAO (Australian National Audit Office) (1997-1998) : « Program Evaluation in the Australian Public Service » *AGPS, Performance Audit Report*, n° 3.

Pluralité des évaluateurs et interdisciplinarité

Une même politique publique ayant souvent des effets multiples, il est nécessaire de disposer de plusieurs évaluations correspondant à différentes approches, disciplines ou sensibilités. Le Crédit d'impôt compétitivité emploi (CICE, loi n° 2012-1510 du 29 décembre 2012) fournit un exemple dans ce sens. Quand il s'agira d'évaluer ce dispositif, de nombreuses variables d'intérêt seront envisageables. En premier lieu, une étude d'impact sur la position commerciale de la France vis-à-vis de l'étranger paraît l'évidence, tout comme une étude de l'impact sur l'emploi en termes quantitatifs. Mais d'autres évaluateurs pourront s'intéresser à d'autres problématiques, comme l'impact de la politique sur la structure des qualifications, sur les carrières au sein de l'industrie ou sur les conditions de travail. On pourra aussi évaluer l'impact sur le financement de la protection sociale d'un point de vue plus général, et sur son acceptabilité.

Cet exemple illustre l'importance du pluralisme non seulement dans l'évaluation *a posteriori*, mais également dans sa préparation *a priori*. La phase de coordination en amont doit

permettre de définir les variables d'intérêt dont on souhaite mesurer l'évolution du fait de la politique étudiée, et permettre que chacune des conséquences du dispositif soit évaluée. Il importe donc que cette coordination *ex ante* soit pluraliste en termes de méthodes, de disciplines et de sensibilités. Ce pluralisme *ex ante* est plus complexe à mettre en place que le pluralisme *ex post*, notamment parce que ce dernier peut être obtenu par la juxtaposition d'évaluations émanant de groupes d'experts différents. La phase de coordination initiale étant forcément unique, l'attention pour que le pluralisme existe doit être d'autant plus soutenue dès cette étape.

Le triptyque de l'évaluation

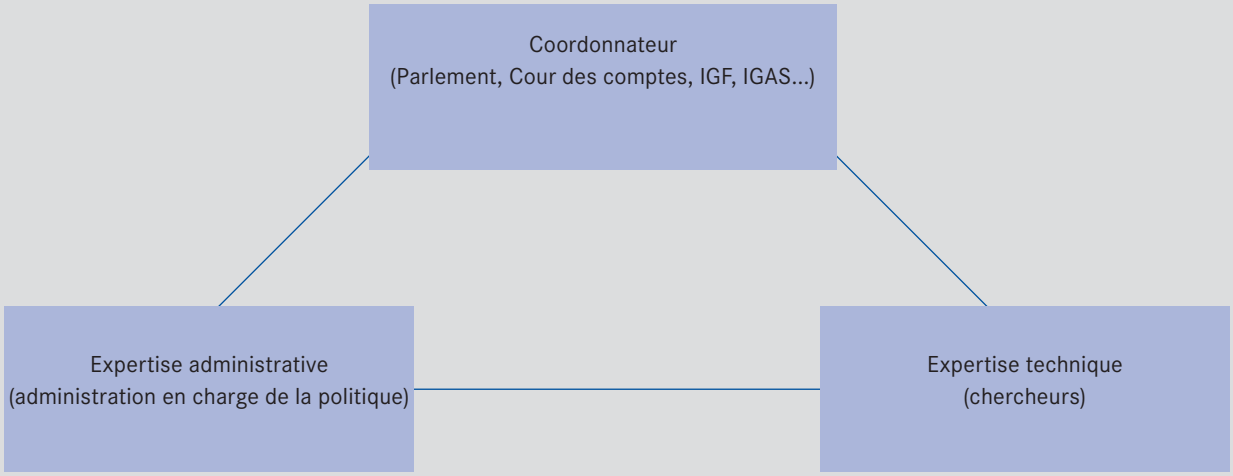
Contrairement à d'autres pays développés (voir l'encadré 4), la France a peu d'expérience en matière d'évaluation des politiques publiques au sens où nous l'avons défini dans cette Note. Une bonne évaluation devrait s'appuyer sur un triptyque formé d'un coordonnateur, des administrations concernées et d'experts indépendants :

– *la coordination* de l'évaluation doit être assurée par une institution extérieure au pouvoir exécutif. La logique démocratique voudrait que ce soit le Parlement qui ait la charge de commanditer ces évaluations. Cela imposerait de lui en donner la capacité technique, c'est-à-dire du personnel pour coordonner réellement la mise en place de l'évaluation, son appréciation et sa diffusion auprès des parlementaires et du grand public. La Cour des comptes est un autre candidat, avec suffisamment de poids institutionnel pour commanditer des évaluations indépendantes[15]. Quel que soit le choix du commanditaire institutionnel, celui-ci devra coordonner la préparation de l'évaluation, s'assurer que la pluralité des approches est bien respectée, et contrôler que toutes les mesures aptes à faciliter les évaluations ont bien été prises (en particulier l'accès aux données). Il devra également organiser la confrontation et la diffusion des résultats. Le choix des organismes évaluateurs devrait se faire par le biais d'appels d'offres publics et être effectué de façon transparente ;

Une bonne évaluation devrait s'appuyer sur un triptyque formé d'un coordonnateur, des administrations concernées et d'experts indépendants.

– *les administrations concernées* doivent apporter leur expertise institutionnelle. La collaboration avec les évaluateurs doit se faire sans pression, sous la surveillance des organismes commanditaires. Les services statistiques des ministères ont des compétences techniques en matière d'évaluation, en plus de leurs compétences institutionnelles. Ils peuvent alors organiser des évaluations paral-

4. Le triptyque de l'évaluation



[15] Le coordonnateur peut aussi être issu de l'administration du moment que ce n'est pas l'administration en charge de la politique évaluée. On pense ici à l'Inspection générale des finances et à l'Inspection générale des affaires sociales.

lèles aux évaluations indépendantes et participer utilement aux débats sur les résultats. Mais ils doivent faciliter la réalisation de l'évaluation indépendante, notamment en ouvrant un accès complet et éclairé aux données ;

– *les experts* doivent apporter leur compétence scientifique en tant qu'évaluateurs. Leur indépendance doit être assurée, entre autres, par leur rotation, de manière à éviter toute captation en fonction des résultats passés d'évaluations. Les experts doivent se soumettre aux contraintes liées au secret statistique et être transparents quant à leurs activités annexes pouvant engendrer des conflits d'intérêts. Il est primordial qu'ils collaborent avec les autres disciplines, notamment dans les phases préparatoires et celles de diffusion des résultats.

Conclusion

Si elle nécessite de combiner une expertise technique, une expertise administrative et une organisation rigoureuse garantissant indépendance et pluralisme, l'évaluation des politiques publiques n'en est pas moins à la portée d'un gouvernement désireux de faire le tri dans ses politiques. Il faut cependant souligner trois conditions *sine qua non* de réussite et de crédibilité de l'évaluation : l'accès aux données, le temps de l'expertise, la publication des résultats. Ces conditions ne doivent pas être considérées comme des contraintes, mais plutôt comme les ingrédients clés d'une évaluation crédible, sur laquelle le processus de décision pourra réellement s'appuyer en toute transparence.

113 rue de Grenelle 75007 PARIS **Téléphone : 01 42 75 53 00** **Télécopie : 01 42 75 51 27** **www.cae-eco.fr**