# L'assignation aléatoire comme méthode d'évaluation des politiques publiques

Adrien Bouguen, Juliette Seban

Dans Économie & prévision 2014/1-2 n° 204-205 , pages 119 à 143
Éditions Direction générale du Trésor

**ISSN 0249-4744**
**DOI 10.3917/ecop.204.0119**
**Date de mise en ligne : 14/01/2016**

**Article disponible en ligne à l'adresse**
https://shs.cairn.info/revue-economie-et-prevision-2014-1-page-119?lang=fr

CAIRN.INFO

Découvrir le sommaire de ce numéro, suivre la revue par email, s'abonner...
Scannez ce QR Code pour accéder à la page de ce numéro sur Cairn.info.

**Distribution électronique Cairn.info pour Direction générale du Trésor.**
Vous avez l'autorisation de reproduire cet article dans les limites des conditions d'utilisation de Cairn.info ou, le cas échéant, des conditions générales de la licence souscrite par votre établissement. Détails et conditions sur cairn.info/copyright.
Sauf dispositions légales contraires, les usages numériques à des fins pédagogiques des présentes ressources sont soumises à l'autorisation de l'Éditeur ou, le cas échéant, de l'organisme de gestion collective habilité à cet effet. Il en est ainsi notamment en France avec le CFC qui est l'organisme agréé en la matière.

# L'assignation aléatoire comme méthode d'évaluation des politiques publiques

**Adrien Bouguen**(*)

**Juliette Seban**(**)

*Alors que, dans une économie concurrentielle, le marché est censé assurer la sélection des stratégies industrielles ou commerciales les plus efficaces, les politiques publiques sont mises en œuvre dans un contexte qui n'est généralement pas propice à la détermination de leur efficacité. Fonder les politiques publiques sur des preuves scientifiques rigoureuses ouvre pourtant la voie à la justification de l'action publique en facilitant la généralisation de programmes qui ont témoigné de leur efficacité et, à l'inverse, en permettant de motiver des modifications apportées à ceux dont les résultats sont moins convaincants.*

*Évaluer l'impact d'une politique publique pose des problèmes pratiques et théoriques complexes. Ainsi, il est par exemple délicat d'établir un lien entre la participation à un dispositif d'insertion et le retour à l'emploi des bénéficiaires, car ce retour peut être influencé par de nombreux facteurs, tels que la conjoncture économique globale, les autres dispositifs disponibles, le retour naturel vers l'emploi, etc. Idéalement, on aimerait connaître la situation des bénéficiaires si ces derniers n'avaient pas bénéficié de la politique. En réalité, une personne est soit bénéficiaire soit non bénéficiaire. Il est donc impossible de connaître sa situation dite* contrefactuelle*, c'est-à-dire la situation qui aurait été observée si la personne n'avait pas bénéficié de la politique ou* vice versa*. Comparer simplement la situation des bénéficiaires avec leur situation avant la mise en place de la politique ou avec la situation d'un groupe de non bénéficiaires donne la plupart du temps des réponses biaisées ne reflétant pas l'impact de la politique étudiée. L'exigence de causalité a donc poussé les chercheurs à développer des stratégies plus sophistiquées.*

*Issue de la science médicale, l'assignation aléatoire constitue une stratégie particulièrement intéressante. Son principe est simple. Il s'agit de comparer,* après intervention*, les résultats de deux groupes d'individus constitués aléatoirement,* avant l'intervention *: un groupe dit "test" bénéficiant de la politique et un groupe "témoin", qui n'en bénéficie pas. L'assignation aléatoire permet de s'assurer qu'initialement, les groupes sont statistiquement identiques. Toute différence entre ces deux groupes expérimentaux observée à la fin de l'évaluation ne peut donc provenir que de l'intervention dont a bénéficié le groupe test.*

*Cet article présente tout d'abord de façon théorique le problème de l'inférence causale – comment établir un lien causal rigoureux entre le programme évalué et les effets mesurés – et la manière dont l'assignation aléatoire le résout.*

(*) Au moment de la rédaction de cet article, Adrien Bouguen était doctorant à l'École d'économie de Paris (Paris School of Economics, PSE).
E-mail : abouguen@povertyactionlab.org

(**) Au moment de la rédaction de cet article, Juliette Seban était doctorante au CES-CNRS, Université Paris 1, Panthéon-Sorbonne.

Cet article n'engage que ses auteurs et non les institutions auxquelles ils appartiennent. Il n'engage *a fortiori* ni la direction générale du Trésor, ni le ministère des Finances et des Comptes publics, ni le ministère de l'Économie, de l'Industrie et du Numérique.



Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

*Il présente ensuite les éléments clés qui déterminent le choix d'un protocole d'évaluation par assignation aléatoire : le niveau d'assignation, les différentes stratégies d'assignation, les problèmes d'acceptabilité et d'éthique, ainsi que les déterminants de la puissance statistique d'une évaluation. Les avantages et les limites des différentes options dans cette étape essentielle de construction du protocole sont détaillés.*

*Enfin, l'article décrit les menaces potentielles qui pèsent sur les résultats d'analyse, en précisant celles qui sont communes à toutes les méthodes d'évaluation et celles qui sont spécifiques à l'évaluation aléatoire. Sont distingués les risques pesant sur la validité interne, c'est-à-dire sur les conditions permettant une estimation non biaisée d'un programme, des menaces sur la validité externe, c'est-à-dire sur les conditions permettant la généralisation des résultats locaux.*

*En conclusion, cet article rassemble la littérature théorique existante sur la portée et les limites de l'assignation aléatoire comme méthode d'évaluation des politiques publiques et lui apporte un éclairage plus opérationnel* via *des exemples concrets.*

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

L'approche empirique a longtemps été considérée comme secondaire en économie, Keynes prétendant même que la valeur d'une théorie économique ne pouvait être invalidée par des faits stylisés. Dans un autre domaine, celui de la médecine, les études empiriques ont pourtant constitué l'une des révolutions les plus importantes du vingtième siècle, contribuant à la fois à de nouvelles découvertes cliniques[1] et à l'établissement d'une validation clinique rigoureuse. Dès 1962, la législation américaine a imposé de fait la méthodologie de l'évaluation aléatoire, qui est restée jusqu'à aujourd'hui la règle d'or dans l'industrie pharmaceutique. Il s'agissait de constituer de manière aléatoire un groupe test à qui une substance active était prodiguée et de le comparer à un groupe témoin à qui était proposé un placebo (*i.e.* un médicament sans substance active). Si l'utilisation de la méthode aléatoire pour tester l'impact des politiques publiques a été introduite à cette même époque (l'expérimentation des *Perry Preschool* date de 1961), elle n'a pas connu le même succès que dans le domaine médical, où cette méthode s'est imposée comme la norme de validation clinique. À la différence des sciences médicales, l'application de la méthode aléatoire aux sciences sociales se heurtait en effet à au moins trois difficultés :

– une difficulté politique : le temps politique n'est pas toujours adapté au temps nécessaire à une évaluation rigoureuse de l'efficacité des politiques publiques ;

– une difficulté technique : les collectes de données nécessaires à ce type d'évaluation (taille d'échantillon, suivi de cohortes, etc.) n'étaient sans doute pas réalisables sans outils informatiques performants ;

– une difficulté théorique : en sciences sociales, le placebo n'existe pas.

L'ampleur des problèmes d'identification posés par le biais de sélection (détaillés dans la première partie de cet article) a cependant donné un nouvel essor à la méthode aléatoire appliquée à l'évaluation des politiques publiques. Au début des années 1990, une exposition rigoureuse du problème dit du contrefactuel[2] a convaincu que l'utilisation du tirage aléatoire pouvait être une manière pertinente d'estimer l'effet d'une politique publique. À l'image de la méthode utilisée en médecine, il s'agit de comparer les résultats de deux groupes constitués aléatoirement : un groupe test bénéficiant de la politique et un groupe témoin. Ainsi, depuis une quinzaine d'années, de nombreux programmes de lutte contre la pauvreté ont été évalués avec la méthode aléatoire dans les pays en développement. Dans un contexte de restriction budgétaire, l'évaluation rigoureuse des politiques publiques intéresse également de plus en plus les pays développés, notamment les institutions publiques françaises[3] et européennes[4]. S'il est vrai qu'il existe de nombreuses autres méthodes d'évaluation *ex post* et non expérimentales (régression multivariée, appariement, etc.), celles-ci reposent sur des hypothèses parfois lourdes, qui rendent leurs résultats dépendants du modèle utilisé. *A contrario*, la méthode aléatoire permet d'identifier les effets d'une politique avec un nombre restreint d'hypothèses. Les contraintes qu'impose la construction d'un protocole expérimental rendent néanmoins l'utilisation de la méthode aléatoire difficile et les résultats obtenus à la suite de ces expérimentations ne sont pas exempts de limites. Cet article se propose de présenter de manière synthétique les avantages théoriques de la méthode aléatoire, les conditions dans lesquelles cette méthode peut être mise en œuvre mais également les difficultés inhérentes à une approche expérimentale. Cet article rassemble la littérature théorique existante sur la portée et les limites de l'approche expérimentale et lui apporte un éclairage plus opérationnel *via* des exemples concrets de problématiques rencontrées lors de la mise en œuvre de ce type d'évaluation.

Dans la première partie de cet article, une version condensée de la formalisation du problème de l'inférence causale est présentée. Il y est détaillé notamment la manière dont l'assignation aléatoire résout le problème du biais de sélection. En deuxième partie, les différents protocoles d'évaluations possibles sont examinés, ainsi que leurs conséquences en termes d'applicabilité, d'éthique et de précision statistique. Enfin, les menaces potentielles qui pèsent sur les résultats d'analyse sont détaillées dans une troisième partie, en distinguant celles qui sont communes à toutes les méthodes d'évaluation et celles qui sont spécifiques à l'évaluation aléatoire.

## Pourquoi l'assignation aléatoire ?

### Inférence causale et contrefactuel

L'évaluation de l'impact d'un programme de soutien scolaire (cours du soir, nouvelle méthodologie d'apprentissage, etc.) ou d'une politique d'aide au retour à l'emploi (dispositif d'accompagnement ou d'incitation tel que le Revenu de solidarité active "activité" ou la prime pour l'emploi) se heurte au problème de l'inférence causale. La causalité découle en effet de la comparaison d'une situation observée avec une situation contrefactuelle, c'est-à-dire la situation qui aurait été observée si la politique n'avait pas été mise en œuvre. Idéalement, pour évaluer l'impact d'une politique de soutien scolaire sur des enfants ayant des difficultés d'apprentissage de la lecture, il faudrait que les résultats scolaires des élèves ayant reçu la politique soient comparés avec les résultats qu'auraient obtenus les mêmes élèves s'ils n'avaient pas bénéficié de ce programme. Une même personne ne

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

pouvant être simultanément observée dans une situation de bénéficiaire et de non bénéficiaire, le problème du "contrefactuel" n'a pas de solution simple.

La plupart du temps, observer le même individu avant et après la mise en œuvre d'une politique (ce qu'on appelle la technique de la "simple différence") ne résout que très imparfaitement le problème. Comme les élèves améliorent spontanément leurs compétences scolaires durant l'année, observer le même individu juste avant le début de l'intervention (en début d'année par exemple) et lorsque l'intervention se termine (en fin d'année scolaire par exemple) ne permet en aucun cas d'isoler l'effet de la politique. En effet, ce dernier se mêlerait à l'effet naturel de l'amélioration des compétences.

Comparer les élèves ayant bénéficié de la politique et ceux qui n'en ont pas bénéficié ne résout généralement pas davantage le problème. En effet, dans la plupart des cas, les bénéficiaires possèdent un certain nombre de caractéristiques observables et non observables qui les différencient intrinsèquement des élèves non bénéficiaires. Ainsi, un élève volontaire pour participer à un groupe de soutien scolaire est sans doute plus motivé, provient d'une famille plus informée ou est au contraire moins compétent que ceux qui ne sont pas volontaires. La simple comparaison des résultats de ces deux groupes ne peut donc suffire à évaluer l'effet de la politique sur la réussite éducative.

D'autres techniques économétriques plus sophistiquées (différence de différences, appariement, régression discontinue, variables instrumentales, régression multivariée, etc.) peuvent estimer l'effet d'une politique de manière beaucoup plus convaincante (Givord, 2010-2014). La validité de ces approches dépend de façon décisive de la crédibilité des hypothèses sur lesquelles elles reposent.

La stratégie adoptée par la méthode aléatoire est toute différente. Il s'agit de mener une expérimentation dans laquelle l'évaluateur sera en mesure d'allouer de façon aléatoire les bénéficiaires et les non bénéficiaires d'une politique. Par cette stratégie, dite "expérimentale" ou "contrôlée" (*Randomized Controlled Trial* en anglais), l'évaluateur pourra produire un estimateur non biaisé de l'effet de la politique, qu'on appelle aussi effet du traitement. Cependant, les méthodes liées à cette stratégie impliquent souvent un coût et un temps d'évaluation importants et ne sont pas exemptes de limites. Il peut arriver en effet que les conditions pour obtenir un estimateur non biaisé de l'effet de la politique ne soient pas réunies.

**Modèle de Rubin et intérêt du tirage aléatoire**

La première formalisation théorique de l'inférence causale et du problème soulevé par le contrefactuel a été établie en 1974 par Donald Rubin (Rubin, 1974). Rubin pose le problème légèrement différemment de la manière dont la recherche l'avait abordé précédemment. Il s'intéresse au résultat potentiel (en termes de revenu, par exemple) auquel l'individu *i* aurait pu aboutir s'il avait bénéficié de la politique ($Y_i^1$) ainsi que celui qu'il aurait présenté s'il n'avait pas bénéficié de la politique ($Y_i^0$). L'estimation de l'effet du "traitement", c'est-à-dire de la politique, pour l'individu *i* s'estime donc par simple différence entre ($Y_i^1$) et ($Y_i^0$). Comme l'évaluateur est souvent intéressé par l'effet moyen d'une politique publique sur un échantillon (représentatif) de la population, l'objectif de l'évaluation réside dans l'estimation de la différence moyenne entre les deux "résultats" potentiels, c'est-à-dire, de $e = E(Y_i^1 - Y_i^0)$ où $E$ représente l'espérance mathématique.

Cependant, un individu ne peut être observé simultanément dans une situation de bénéficiaire et de non bénéficiaire. Dès lors, au moins l'un des deux "résultats" potentiels ($Y_i^0$ ou $Y_i^1$) n'est pas observé. L'estimation de la différence *e* est donc impossible. Une manière naïve de se rapprocher de l'effet du traitement consisterait à estimer la différence *D* entre le résultat potentiel obtenu chez les individus ayant bénéficié de la politique et celui observé chez ceux n'en ayant pas bénéficié. On comparerait alors le résultat obtenu par les bénéficiaires avec ceux obtenus par les non bénéficiaires. Toutefois, la condition d'indépendance ($Y^0, Y^1 \perp T$) entre les résultats et l'affectation au groupe test *T* nécessaire à l'établissement d'une estimation sans biais de l'effet du traitement ne serait en général pas assurée (encadré 1). La simple comparaison des résultats des bénéficiaires et des non bénéficiaires aboutit donc à une estimation biaisée de l'effet du traitement.

Les méthodes dites non aléatoires (régression multivariée, différence de différences ou appariement) résolvent ce problème en conditionnant l'hypothèse d'indépendance à un certain nombre de variables observables ($Y^0, Y^1 \perp T / X$). Cette hypothèse d'indépendance conditionnelle est cependant particulièrement difficile à soutenir dans le cas de l'évaluation de politiques publiques. Le fait même de participer à une politique publique implique un certain nombre de caractéristiques, parfois très difficiles à observer, qui sont souvent corrélées avec le résultat final. Même en contrôlant par un nombre important de caractéristiques, la simple participation au programme rend les bénéficiaires et les non bénéficiaires intrinsèquement différents. Ainsi, si

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

## Encadré 1 : estimation non biaisée de l'effet du traitement

Soit $e$ l'effet sur une population expérimentale d'une politique publique, ou traitement, au regard d'un critère de résultat $Y$ :

(1) $e = E(Y_i^1 - Y_i^0)$

où $Y_i^1$ (respectivement, $Y_i^0$) est le résultat obtenu par l'individu $i$ s'il bénéficie du traitement (respectivement, s'il n'en bénéficie pas). Cependant, pour un individu donné $i$, l'évaluateur observe soit $Y_i^1$, soit $Y_i^0$ : l'estimation de (1) est donc impossible.

Soit $D$, la différence moyenne de résultat entre les bénéficiaires et les non bénéficiaires du traitement :

(2) $D = E(Y_i^1 | T_i = 1) - E(Y_i^0 | T_i = 0)$

où $T_i$ est une indicatrice valant 1 si l'individu $i$ est assigné au groupe de traitement et 0 sinon. Lorsque $T_i$ est égal à 1 (respectivement, 0), l'individu $i$ est dit test ou traité (respectivement témoin ou non traité). En ajoutant et retranchant la valeur non observée $E(Y_i^0 | T_i = 1)$ (c'est-à-dire le résultat qu'auraient obtenu les individus du groupe test s'ils n'avaient pas été traités), on obtient :

$$D = E(Y_i^1 | T_i = 1) - E(Y_i^0 | T_i = 0)$$
$$+ E(Y_i^0 | T_i = 1) - E(Y_i^0 | T_i = 1)$$
$$= E(Y_i^1 - Y_i^0 | T_i = 1) - E(Y_i^0 | T_i = 0) + E(Y_i^0 | T_i = 1)$$

$E(Y_i^1 - Y_i^0 | T_i = 1)$ représente l'effet du traitement sur les individus du groupe test, c'est-à-dire la valeur d'intérêt à isoler[a], tandis que $E(Y_i^0 | T_i = 1) - E(Y_i^0 | T_i = 0)$ représente le biais de sélection. Si les résultats potentiels hors traitement des individus du groupe test et du groupe témoin sont égaux, alors $D$ mesure sans biais l'effet du traitement. Cette situation n'a cependant aucune raison de se produire naturellement. En effet, il est fréquent que la valeur $E(Y_i^0 | T_i = 1)$ soit supérieure à $E(Y_i^0 | T_i = 0)$. C'est le cas par exemple si les élèves les plus motivés ont tendance à participer davantage à une politique de soutien scolaire que les autres élèves : on parle alors d'un phénomène d'auto-sélection. Dans ce cas, $D$ est supérieur à l'effet du traitement sur la population test. Symétriquement, si les enseignants tendent à sélectionner dans un dispositif de soutien scolaire les élèves les plus en difficulté, alors $E(Y_i^0 | T_i = 0)$ est supérieure à $E(Y_i^0 | T_i = 1)$. Dans ce cas, $D$ est inférieur à l'effet du traitement sur la population test. $E(Y_i^0 | T_i = 1)$ étant inobservée, il n'existe pas de test permettant de s'assurer qu'une évaluation de l'effet de la politique fondée sur l'estimation de $D$ ne souffre pas d'un biais de sélection.

Si $E(Y_i^0 | T_i = 1) \neq E(Y_i^0 | T_i = 0)$, alors l'hypothèse d'indépendance des résultats au traitement $Y^0, Y^1 \perp T$ ne tient pas et l'évaluation de l'effet du traitement par estimation de $D$ est biaisée. Inversement, en cas de tirage aléatoire des individus bénéficiaires et non bénéficiaires réalisé de manière à assurer que $E(Y_i^0 | T_i = 1) = E(Y_i^0 | T_i = 0)$ (c'est-à-dire dans le cas où l'hypothèse dite d'assignation aléatoire, ou hypothèse 1, est satisfaite) le biais de sélection identifié en (2) disparaît.

On peut alors écrire :

(3) $Y_i = c + \delta T_i + \varepsilon_i$

où $Y_i$ est le résultat observé, $c$ et $\delta$ sont des coefficients inconnus à estimer et $\varepsilon_i$ est un terme d'erreur d'espérance nulle. $c$ et $\delta$ sont égaux à :

$c = E(Y_i^0 | T_i = 0)$

et $\delta = E(Y_i^1 | T_i = 1) - E(Y_i^0 | T_i = 0) = D$

$T_i$ étant une variable indicatrice, on montre aisément que l'estimateur $\hat{\delta}$ des moindres carrés ordinaires (MCO) de $\delta$ dans (3) s'écrit :

$$\hat{\delta} = \hat{E}(Y_i^1 | T_i = 1) - \hat{E}(Y_i^0 | T_i = 0)$$
$$= \hat{E}(Y_i | T_i = 1) - \hat{E}(Y_i | T_i = 0)$$

où $\hat{E}(.)$ désigne la moyenne empirique de la distribution indiquée entre parenthèses.

$\hat{\delta}$ est un estimateur de $D$, l'effet du traitement sur la population test, qui a de bonnes propriétés (dont celle d'être sans biais) si un certain nombre d'hypothèses, outre l'hypothèse 1, sont vérifiées par l'assignation, permettant d'assurer notamment la non-corrélation entre la variable de traitement $T_i$ et le terme d'erreur $\varepsilon_i$, ainsi que la non-corrélation des termes d'erreur $(\varepsilon_i, \varepsilon_j)_{i \neq j}$ deux à deux. En particulier, on suppose que le résultat potentiel $Y_i^1$ ou $Y_i^0$ n'est pas affecté par le statut de traitement des autres individus $j \neq i$. Cette hypothèse est connue dans la littérature sous la dénomination d'hypothèse SUTVA, pour *Stable Unit Treatment Value Assumption* (Angrist et Imbens, 1996). Nous reviendrons plus en détails sur ces hypothèses *infra*.

(a) $E(Y_i^1 - Y_i^0 | T_i = 1)$ n'est pas équivalent à $E(Y_i^1 - Y_i^0)$. En effet, $E(Y_i^1 - Y_i^0 | T_i = 1)$ représente l'effet du traitement sur les individus qui sont traités alors que $E(Y_i^1 - Y_i^0)$ est l'effet du traitement sur l'ensemble de la population. Dans la mesure où très peu de politiques sont susceptibles de concerner tout le monde, $E(Y_i^1 - Y_i^0)$ n'est presque jamais identifiable et n'est au demeurant pas très utile. La recherche et le politique s'intéressent davantage à l'effet d'une politique sur les personnes qui sont susceptibles d'en bénéficier, c'est-à-dire à $E(Y_i^1 - Y_i^0 | T_i = 1)$.

ces deux groupes étaient réellement comparables, on ne pourrait pas expliquer pourquoi certains ont participé et d'autres ont décidé de (ou ont été forcés à) ne pas participer. Si, par exemple, le rendement individuel d'une politique n'est connu que de l'individu lui-même, alors contrôler de manière satisfaisante le biais de sélection sera très difficile, voire impossible. De même, il existe en général un nombre important de variables non observables et corrélées avec la variable d'intérêt, pour lesquelles il n'existe pas de conditionnement possible. Si tel est le cas, la seule possibilité crédible d'évaluer l'effet du traitement de manière non biaisée est de s'assurer que seul le hasard guide la formation des groupes de bénéficiaires et de non bénéficiaires du traitement. Certes, il existe des situations quasi-aléatoires qui

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

peuvent permettre sans expérimentation d'identifier l'effet du traitement. Ces situations sont appelées des expériences "naturelles". Il s'agit de situations où se produisent naturellement des assignations aléatoires (loteries, événement climatique ou politique exogène, etc.). Toutefois, ces situations sont en nombre restreint et ne permettent pas d'évaluer de manière systématique les politiques publiques. En outre, la stratégie expérimentale semble être la mieux à même de reproduire de manière rigoureuse le scénario idéal pour évaluer l'effet du traitement. En effet, lorsqu'un tel tirage aléatoire a lieu avant le début de la politique à évaluer, les groupes test et témoin ont statistiquement les mêmes caractéristiques observables et non observables[5]. En conséquence, la comparaison des moyennes des résultats du groupe témoin et du groupe test permet l'identification de l'effet du traitement.

Cette identification repose néanmoins sur plusieurs hypothèses (définies à la fin de l'encadré 1). Si la première hypothèse du modèle (assignation aléatoire) est acquise dans le cadre d'une expérience aléatoire, l'hypothèse dite SUTVA soulève plus de difficultés. Il existe au moins deux situations fréquentes dans lesquelles l'hypothèse SUTVA n'est pas assurée. La première correspond aux programmes pour lesquels les effets directs de l'intervention peuvent avoir des conséquences indirectes importantes. Dans l'article de Michael Kremer et d'Eduard Miguel sur l'effet des vers intestinaux sur les performances à l'école (Miguel et Kremer, 2004), le traitement (un médicament contre les vers) produit des effets indirects sur les élèves non traités, du fait de leur plus faible exposition aux parasites. La seconde situation renvoie aux dispositifs qui génèrent des effets d'équilibre général susceptibles d'affecter également le groupe des non traités. Si, par exemple, un programme d'aide au retour à l'emploi produit un effet positif direct sur les participants, il peut réduire les chances de retour à l'emploi des non-participants par un effet de concurrence sur les emplois offerts. Les problématiques relatives aux effets indirects ou aux effets d'équilibre général et les solutions à ces problèmes seront évoquées plus en détail dans la troisième partie de cet article.

**Conséquence d'une "conformité" imparfaite**

Jusqu'à présent, seuls les cas de conformité parfaite ont été considérés : tous les individus alloués au traitement y participent, alors que les individus témoins n'y participent pas. Cette situation n'a cependant que peu de raisons de se produire. Si, dans 50 % des écoles tirées aléatoirement, un programme de soutien scolaire est mis en œuvre, il est peu probable que l'ensemble des élèves de ces écoles bénéficie du dispositif. Ce ne sera évidemment pas le cas si le programme n'est pas obligatoire. En outre, même si le programme est en mesure de recevoir tous les volontaires, certains parents n'inscriront sans doute pas leurs enfants. On dit alors que la conformité (*compliance*, en anglais) est imparfaite. L'interprétation causale du paramètre $\delta$ défini dans l'équation (3) de l'encadré 1 s'en trouve modifiée. En effet, $\delta$ (qui s'estime en pratique en soustrayant au résultat moyen du groupe test le résultat moyen du groupe témoin ; soit $\hat{\delta}$ cet estimateur) identifie désormais l'effet moyen du traitement sur une population composée de participants et de non participants. On parle alors de l'effet moyen de l'intention de traiter (*Intention To Treat* ou *Intent to Treat*, ITT). Cette valeur reste pertinente dans de nombreux cas. Lorsque le décideur politique veut estimer l'effet d'une politique éducative de soutien scolaire, il s'intéresse à l'effet global de cette politique sur l'ensemble de l'école qui a bénéficié de ce dispositif, que les élèves aient participé ou non au programme. Si le programme était généralisé, certains élèves y participeraient, d'autres s'en abstiendraient dans des conditions similaires à celles observées lors de l'évaluation[6].

Le décideur peut cependant s'intéresser également à l'effet du programme de soutien scolaire sur les élèves qui ont effectivement participé au programme. Il s'agit de l'effet moyen du programme sur les traités (*Treatment On The Treated*, TOT). Comme indiqué dans l'encadré 2, cet effet est estimé en divisant l'estimateur de l'intention de traiter par la proportion des individus du groupe test qui participent effectivement au traitement (ce qu'on appelle le taux d'exposition dans le groupe test) corrigée de la proportion des individus affectés au groupe témoin mais qui participent néanmoins au traitement (*i.e.* le taux d'exposition dans le groupe témoin). Si, par exemple, la comparaison entre groupe test et groupe témoin donne un effet en ITT significatif de + 20 points de pourcentage sur une variable de résultat binaire et que la participation à la politique est de 80 % dans le groupe test et de 0 % dans le groupe témoin, alors l'effet de la politique sur les participants est de 0,2/0,8 = + 25 points de pourcentage. En plus des trois premières hypothèses de l'ITT, deux hypothèses supplémentaires doivent être posées pour identifier le TOT de manière non biaisée (encadré 2).

Comme n'importe quelle stratégie d'identification par variable instrumentale, le TOT capte l'effet sur les individus qui sont sensibles à la variable instrumentale, c'est-à-dire ceux qui participent au traitement lorsqu'ils sont assignés au groupe test et ceux qui n'y participent pas lorsqu'ils sont assignés au groupe témoin (encadré 2). Ces individus sont parfois appelés les *compliers* (ceux qui respectent l'assignation aléatoire). Grâce à l'assignation aléatoire, les *compliers* du groupe test sont comparables aux *compliers* du groupe témoin. L'importance de l'hypothèse de monotonicité (hypothèse 5 de l'encadré 2) réside dans le fait que la présence d'individus rebelles (*defiers*, participant au traitement en cas d'affectation au groupe témoin et n'y participant pas en cas d'affectation au groupe



Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

## Encadré 2 : les cinq hypothèses de l'estimation aléatoire dans le cas d'une conformité imparfaite et les liens entre ITT et TOT

Les notations de cet encadré sont les mêmes que celles de l'encadré 1. L'indicatrice $T_i$ indique toujours dans quel groupe expérimental l'individu *i* est aléatoirement affecté : $T_i = 1$ si *i* est affecté au groupe test et $T_i = 0$ s'il est affecté au groupe témoin. Cependant, si la conformité est imparfaite, l'assignation à l'un ou l'autre des groupes expérimentaux n'équivaut pas tout à fait au fait de bénéficier ou pas du traitement. Une seconde indicatrice $P_i$ est donc introduite, qui vaut 1 si l'individu *i* participe effectivement au traitement et 0 sinon.

### Hypothèses de l'intention de traiter (*Intention To Treat*, ITT) et interprétation de l'estimateur ITT

Reprenons le modèle (3) introduit dans l'encadré 1 :

(3) $Y_i = c + \delta T_i + \varepsilon_i$

Si les hypothèses 1 à 3 *infra* sont vérifiées, l'assignation aléatoire permet d'obtenir une évaluation ayant de bonnes propriétés (dont l'absence de biais) de l'effet moyen de l'intention de traiter, c'est-à-dire de l'effet global du traitement sur la population test, à travers l'estimation par les MCO $\hat{\delta}$ de $\delta$ dans (3) telle que précisée à la fin de l'encadré 1 :

$$\hat{\delta} = \hat{E}(Y_i|T_i = 1) - \hat{E}(Y_i|T_i = 0) = \hat{\delta}_{ITT}$$

*Hypothèse* 1 : l'assignation au traitement est aléatoire, au sens où chaque individu de l'échantillon a la même probabilité d'être affecté au groupe test ou au groupe témoin.

*Hypothèse* 2 (*SUTVA*) : le résultat potentiel $Y_i$ d'un individu *i* n'est fonction que de l'affectation ou non au groupe de traitement $T_i$ et de la participation au traitement $P_i$ de l'individu *i*. Il n'est pas fonction de l'affectation et de la participation des autres individus de l'échantillon (que l'on pourrait représenter par les vecteurs *P* et *T*). Ainsi, $Y_i(T,P) = Y_i(T_i,P_i)$. Cela signifie qu'il n'existe pas d'externalité (ou d'effets d'équilibre général – *cf. infra*) de l'assignation aléatoire d'un individu sur le résultat final ou sur le comportement de participation d'un autre individu.

*Hypothèse* 3 (*hypothèse d'exclusion*) : $E(T_i \varepsilon_i) = 0$. Cette hypothèse ne découle pas nécessairement du tirage aléatoire. En effet, si la simple assignation dans un groupe modifie le résultat final, alors l'hypothèse d'exclusion n'est pas assurée. Par exemple, les effets dits de Hawthorne ou de John Henry (modification spontanée des comportements des individus dès lors qu'ils sont affectés dans l'un ou l'autre des groupes expérimentaux) sont susceptibles de biaiser les résultats de l'analyse[(a)].

### Hypothèses du traitement sur les traités (*Treatment On the Treated*, TOT)

Considérons maintenant le modèle (4) constitué des deux équations suivantes :

(4.1) $Y_i = c + \delta P_i + \varepsilon_i$

(4.2) $P_i = \alpha_0 + \alpha_1 T_i + v_i$

où $\alpha_0$, $\alpha_1$, $c$ et $\delta$ sont des coefficients inconnus à estimer et $\varepsilon_i$ et $v_i$ des termes d'erreur d'espérance nulle.

L'assignation aléatoire permet d'obtenir une évaluation ayant de bonnes propriétés de l'effet moyen du traitement sur les individus traités (TOT) à partir de l'estimation du paramètre $\delta$ dans le modèle (4) si les hypothèses 1 à 5 sont vérifiées.

*Hypothèse* 4 (*non-zero effect assumption*) : *il existe une corrélation non nulle entre l'assignation au traitement et la participation* : $corr(T_i,P_i) \neq 0$. Cette hypothèse est facilement vérifiable et ne pose en général pas de difficulté.

*Hypothèse* 5 (*hypothèse de monotonicité*) : soit $P_i^1$ la participation potentielle de l'individu *i* s'il est affecté au groupe test et $P_i^0$ la participation potentielle de l'individu *i* s'il est affecté au groupe témoin. La monotonicité implique que $P_i^1 \geq P_i^0$ quel que soit *i*. Ceci suppose que le traitement influence la participation des individus toujours dans le même sens. Sans l'hypothèse de monotonicité, il serait possible d'identifier un groupe d'individus "rebelles" (les *defiers*) tels que $P_i^1 = 0$ et $P_i^0 = 1$. De tels individus participeraient au traitement seulement en cas d'affectation au groupe témoin et n'y participeraient pas en cas d'affectation au groupe test.

Notons que l'hypothèse 3 appliquée au modèle (4) implique que l'effet de $T_i$ sur $Y_i$ n'est produit que par l'effet de $T_i$ sur $P_i$.

Les variables $P_i$ et $T_i$ étant des indicatrices, on montre aisément que l'estimateur des doubles moindres carrés de $\delta$ dans (4.1), la participation $P_i$ étant instrumentée par l'affectation au traitement $T_i$ (équation (4.2)) s'exprime comme l'estimateur de Wald de $\delta$ dans le modèle (4) :

$$\hat{\delta}_{TOT} = \frac{\hat{E}(Y_i|T_i=1) - \hat{E}(Y_i|T_i=0)}{\hat{E}(P_i|T_i=1) - \hat{E}(P_i|T_i=0)}$$

### Lien entre ITT et TOT et interprétation de l'estimateur TOT

L'estimateur TOT s'exprime immédiatement en fonction de l'estimateur ITT comme suit :

$$\hat{\delta}_{TOT} = \frac{\hat{\delta}_{ITT}}{\hat{E}(P_i|T_i=1) - \hat{E}(P_i|T_i=0)}$$

En d'autres termes, l'estimateur TOT s'exprime comme le rapport entre l'estimateur ITT et l'écart entre les taux de participation au traitement des individus du groupe test et du groupe témoin (cet écart est appelé le taux de conformité). L'estimateur TOT s'interprète comme l'effet moyen du traitement sur les individus qui sont sensibles à l'assignation aléatoire, *i.e.* ceux pour qui $(P_i^1 - P_i^0) = 1$, c'est-à-dire les *compliers* (*cf.* démonstration en annexe 1) :

$$\hat{\delta}_{TOT} = \hat{E}[(Y_i^1 - Y_i^0)|\hat{E}(P_i^1 - P_i^0) = 1]$$

Cet estimateur est aussi appelé *Local Average Treatment Effect* ou LATE (Angrist et Imbens, 1994).

(a) *cf.* troisième partie, première sous-partie, deuxième alinéa (intitulé "Impact du tirage sur le comportement des agents") pour plus de détails.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

test) rendrait l'identification de la population des *compliers* impossible (*cf.* encadré 2 et annexe 1). Notons qu'en vertu du tirage aléatoire, il existe également deux autres groupes, toujours équivalents dans les groupes test et témoin :

– les *never-takers*, qui ne participent jamais au traitement, qu'ils soient affectés au groupe test ou au groupe témoin (c'est-à-dire tels que $P_i^1 = P_i^0 = 0$, en reprenant les notations de l'encadré 2). Dans notre exemple de soutien scolaire, les *never-takers* correspondent aux élèves ou à leurs parents refusant systématiquement toute forme de soutien ;

– les *always-takers*, tels que $P_i^1 = P_i^0 = 1$, c'est-à-dire les individus qui participent toujours au traitement, quelle que soit leur affectation. Dans notre exemple, il s'agirait d'un groupe d'élèves capables de bénéficier du soutien scolaire (par pression ou influence des parents), même si celui-ci ne leur est pas proposé.

Sur ces deux groupes, l'effet du traitement ne pourra jamais être identifié car leurs individus constitutifs ne sont pas *sensibles* au traitement : le fait de leur proposer ou non le traitement ne modifie en rien leur statut de traité ou non. Dans de nombreux cas, le décideur politique peut être spécifiquement intéressé par l'effet sur les *compliers*, car c'est sur cette population que le programme aura naturellement un effet. Dans notre exemple du programme de soutien scolaire, l'effet détecté captera l'impact sur les individus ayant bénéficié de la politique et qui n'en auraient pas bénéficié sans l'intervention. Cette restriction aux *compliers* pose cependant un problème quant à la validité externe de l'expérimentation. Une fois la mesure généralisée, le groupe de volontaires peut avoir des caractéristiques différentes de celles observées sur les volontaires lors de l'expérimentation[7].

## Définir le protocole

Cette partie aborde les éléments clés qui déterminent le choix d'un protocole d'évaluation. En tout premier lieu, il convient de s'intéresser aux déterminants qui président au choix du niveau d'assignation (individuel ou par grappe). Ce choix est guidé autant par des questions de précision statistique que par des questions éthiques ou de faisabilité politique. En pratique, il existe plusieurs manières de procéder à ce tirage aléatoire. Ces différentes stratégies de tirage, leurs avantages et limites respectifs, seront développés. Enfin, les déterminants de la puissance statistique d'une évaluation seront présentés. La construction du protocole constitue une étape essentielle de la mise en place d'une évaluation aléatoire. Elle peut prendre plusieurs mois et se fait en concertation avec les différents acteurs.

### Niveau d'assignation aléatoire

L'assignation aléatoire peut s'effectuer à différents niveaux : celui de l'individu, du ménage, de l'école ou, encore, de la ville. Au moins cinq aspects importants doivent être pris en compte pour déterminer le niveau de tirage optimal.

#### *Les contraintes liées à l'intervention elle-même*

Le niveau de tirage est d'abord déterminé par les contraintes du programme lui-même et le niveau auquel le programme est mis en œuvre. Si un programme concerne une modification de l'organisation d'un établissement scolaire, le tirage devra par définition s'effectuer au niveau de l'établissement.

#### *Les effets de diffusion (externalités)*

En relation avec le point précédent, l'évaluateur devra prendre en compte les effets de diffusion potentiels. Plus les effets de diffusion sont importants, moins le tirage individuel devient pertinent. Prenons comme exemple un programme de soutien scolaire offert seulement à certains élèves d'une même classe. L'impact du programme sur les élèves concernés (élèves test) risque de se diffuser aux élèves témoin scolarisés dans la même classe, par l'interaction entre les élèves ou par l'intermédiaire de l'enseignant, qui disposerait alors d'une classe plus homogène. Si un tel phénomène se produit, le groupe témoin sera "contaminé" puisqu'il aura été affecté indirectement par le programme. Pour éviter ces externalités, une assignation aléatoire au niveau des classes ou des écoles sera préférée[8].

#### *Les contraintes organisationnelles ou politiques*

Il est parfois inconcevable politiquement de n'offrir un programme qu'à certaines écoles au sein d'une même ville. Il faut alors prendre la ville comme unité d'assignation. De même, il est possible que, pour des raisons pratiques, le partenaire responsable de la mise en œuvre du programme ou de l'intervention (une association, une agence d'État ou une entreprise) ne puisse pas traiter des unités trop éloignées les unes des autres. Une discussion avec les parties prenantes (les autorités administratives et politiques, les partenaires responsables de l'intervention, etc.) avant le début de l'évaluation est donc absolument nécessaire pour tenir compte des contraintes de chacun.

#### *Les questions éthiques*

Des questions éthiques peuvent jouer dans le choix du niveau de l'assignation. Il peut s'avérer en effet difficilement concevable d'offrir un programme à certaines personnes d'un même quartier ou à des élèves d'une classe et pas aux autres. Un tirage au niveau des quartiers ou des écoles permet souvent de remédier à ce problème : tous les élèves éligibles des quartiers/écoles test reçoivent le programme et



Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

aucun des quartiers/écoles témoin. Différents types de protocole ont été imaginés pour rendre plus acceptable le tirage aléatoire (voir *infra*, deuxième partie, deuxième sous-partie, intitulée "Différents protocoles d'évaluation").

### *La puissance statistique*

Enfin, la puissance statistique joue un rôle central dans le choix d'un protocole d'évaluation. Comme détaillé *infra* dans la troisième sous-partie ("Puissance d'une évaluation"), les tirages aléatoires réalisés à des niveaux supérieurs (classe, ville, etc.) présentent le désavantage de réduire la précision statistique, ce qui a de fortes chances de se traduire par une nécessaire augmentation de la taille de l'échantillon. Or, une telle augmentation a bien évidemment des conséquences importantes en termes de coût, à la fois pour le partenaire et pour l'évaluateur.

### *Un exemple*

Prenons l'exemple du recrutement local sur profil des enseignants dans le secondaire[9], programme dit "Éclair", qui vise à augmenter la compétition au sein des établissements ainsi qu'entre les enseignants afin de générer des gains d'efficacité et de motivation. Imaginons que l'évaluateur ait la charge d'analyser l'effet de ce programme de l'éducation nationale, qui consiste à donner plus de liberté à certains établissements dans le recrutement de leurs enseignants : certains enseignants seraient recrutés "sur dossier" alors que d'autres suivraient la voie d'allocation classique ("système par point"). Comme ce programme affecte de manière prioritaire les enseignants, on pourrait imaginer d'octroyer aléatoirement à certains d'entre eux le statut d'"Éclair". Les enseignants poseraient leur candidature directement dans des établissements scolaires, ces derniers pouvant accepter ou refuser de les embaucher. On effectuerait alors des mesures sur la motivation des enseignants et, éventuellement, les résultats scolaires de leurs élèves. Cette stratégie de tirage au niveau individuel présenterait l'avantage de disposer d'une très bonne puissance statistique. Elle pose néanmoins plusieurs problèmes au regard des critères évoqués ci-dessus. D'abord, le niveau d'assignation individuel ne correspond pas au niveau auquel le décideur politique avait initialement imaginé mettre en œuvre la politique : le programme devait affecter des établissements "éclairs" et non pas des enseignants "éclairs". Même si l'analyse que l'évaluateur produirait serait en lien avec la politique proposée, elle ne la reproduirait pas fidèlement : les effets au niveau de l'établissement (effet de motivation collective, de modification du mode de management, etc.) ne seraient pas pris en compte. En outre, même si un tirage au niveau individuel ne pose pas à proprement parler de problème éthique – on ne retire rien à l'enseignant témoin, seuls les critères de nomination sont modifiés – ce tirage peut poser un problème d'acceptabilité politique ou syndicale. Dans ce cas d'espèce, cependant, il n'est pas sûr qu'un tirage au niveau des établissements soit plus acceptable. Peut-être un tirage au niveau des circonscriptions pourrait-il être mieux accepté par les différents acteurs. Comme évoqué en deuxième partie, troisième sous-partie ("Puissance d'une évaluation"), un tirage au sort à un niveau si élevé nécessiterait cependant une taille d'échantillon importante. Il faudrait tirer au moins une centaine de circonscriptions, ce qui reviendrait à faire passer des dizaines de milliers de tests de compétences scolaires. Il revient à l'évaluateur d'estimer de manière adéquate le coût et la faisabilité d'un tel protocole.

## Différents protocoles d'évaluation

### *Demande excédentaire*

Le cas le plus naturel pour introduire l'assignation aléatoire est celui où la demande pour un programme excède l'offre. Pour des raisons pratiques ou financières, le partenaire est parfois incapable de fournir le programme à toute la population cible : il peut ainsi décider dans un premier temps de n'octroyer le programme qu'à une partie de la population totale. Prenons l'exemple des internats d'excellence[10]. Avec un nombre limité de lits et des candidats souvent nombreux, les internats constituent un cadre très favorable à un tirage aléatoire. L'assignation aléatoire est une règle aisément justifiable par les partenaires et facilement comprise et acceptée par les bénéficiaires potentiels. Elle assure, une fois la population cible définie, que les candidats aient une probabilité égale de bénéficier de la politique. Cette procédure peut même être considérée comme plus neutre qu'une sélection fondée sur des critères subjectifs et, parfois, arbitraires.

### *Mise en place progressive ou rotation*

Le protocole dit de mise en place progressive (*phase-in* en anglais) constitue une autre approche intéressante. Lorsqu'un programme doit être progressivement étendu pour couvrir l'ensemble d'une population éligible, il est possible d'utiliser cette graduation pour évaluer l'impact du programme. Il est alors nécessaire de sélectionner aléatoirement l'ordre dans lequel les unités éligibles (personnes, écoles) bénéficieront du programme. Ce protocole donne à chaque unité éligible une chance égale de recevoir le traitement, lors de la première phase ou d'une phase ultérieure du programme. Tant que certaines unités n'auront pas intégré le programme, elles constitueront le groupe de comparaison servant à estimer le contrefactuel pour les groupes déjà soumis au traitement. Cette méthode a notamment permis d'évaluer le programme de transferts monétaires conditionnels PROGRESA, dont bénéficient aujourd'hui les ménages pauvres du Mexique rural (Gertler, 2004).

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

La limite principale de cette méthode réside dans l'impossibilité de mesurer des effets de long terme. En effet, une fois toutes les unités traitées, il n'existe plus de groupe témoin permettant la comparaison. Par ailleurs, l'évaluateur doit rester attentif aux potentiels effets d'anticipation. Il est en effet possible que le groupe témoin modifie son comportement lors de la première phase sachant qu'il va recevoir le traitement lors d'une phase ultérieure.

La mise en place d'un système de rotation constitue un autre protocole possible : les individus bénéficiaires du dispositif la première année deviennent le groupe témoin l'année suivante et *vice versa*. Par conséquent, l'ensemble de la population cible bénéficie du programme pendant une année. Cette stratégie ne permet pas non plus de mesurer des effets de long terme et peut également être concernée par des effets d'anticipation. Elle s'avère cependant moins coûteuse dans sa mise en œuvre.

***Encouragement***

Lorsqu'exclure des personnes d'un programme est totalement inacceptable pour des raisons éthiques ou pratiques, une option consiste à mettre en place un protocole d'encouragement. Au lieu d'assigner aléatoirement le traitement, on assigne une incitation (un "encouragement") à participer au programme. Le groupe test aura reçu un encouragement pour participer au programme et le groupe témoin n'en aura pas reçu. Ainsi, personne n'est exclu du programme, mais certains sont incités à y participer. L'encouragement peut prendre différentes formes : un appel, une lettre, un courriel. Cette méthode fonctionne si l'encouragement est suffisamment efficace pour qu'une proportion accrue de personnes dans le groupe test participe effectivement au programme (hypothèse 4 de l'encadré 2). En pratique, il est important de tester l'efficacité de l'encouragement en amont. En effet, dans de nombreux cas, l'incitation ne suffit pas à renforcer suffisamment la participation. Se pose alors un problème de puissance statistique (voir deuxième partie, troisième sous-partie "Puissance d'une évaluation"). Il faut par ailleurs être attentif à ce que l'encouragement ne constitue pas un traitement en soi car, si c'était le cas, l'hypothèse d'exclusion (hypothèse 3 de l'encadré 2) ne serait plus assurée.

Ce protocole implique que l'effet d'intention de traiter (ITT) mesure l'impact non pas d'une politique mais d'un encouragement. Seul l'effet de traitement sur les traités (TOT) mesure de manière pertinente l'effet de la politique elle-même. En pratique, malheureusement, l'effet de l'encouragement est souvent trop faible pour constituer un bon instrument. Ce protocole, pourtant prometteur, n'est donc que rarement employé.

***Traitements multiples***

Il est également possible de comparer différents types de traitement. En pratique, les décideurs préfèrent généralement pouvoir comparer les avantages relatifs de différentes interventions plutôt que de connaître l'impact d'une seule. Prenons l'exemple de l'évaluation *Opérateurs privés de placement/Cap vers l'entreprise* (OPP/CVE) (Behaghel *et alii*, 2009), qui évalue l'impact sur le retour à l'emploi de plusieurs types d'accompagnement de chômeurs. L'accompagnement par Pôle Emploi fait partie du droit commun et il n'est pas possible d'en exclure des individus pour évaluer les bénéfices de cet accompagnement. Il est en revanche possible de comparer ce dernier avec d'autres types d'accompagnement. Dans cette expérimentation, les demandeurs d'emploi étaient assignés à trois traitements différents : un accompagnement par les opérateurs privés de placement, un accompagnement classique par Pôle Emploi et un accompagnement renforcé par Pôle Emploi. Ainsi, l'évaluation a permis de comparer l'efficacité relative de chaque traitement. En toute rigueur, cependant, seule une comparaison des estimations en ITT (encadré 2) peut être réalisée. Rien n'assure en effet que les *compliers* d'une branche de traitement soient comparables aux *compliers* d'une autre.

**Puissance d'une évaluation**

La plupart des évaluations d'impact cherchent à tester une hypothèse simple qui se résume à la question suivante : le dispositif que l'on cherche à évaluer a-t-il un impact ou bien est-il sans effet ?

La puissance d'une évaluation est la probabilité que, pour une taille d'échantillon, une taille d'effet et un niveau de significativité donnés, l'hypothèse nulle d'absence d'effet soit rejetée si la politique a effectivement un impact. On exprime généralement la puissance en *Effet Minimum Détectable* (EMD). L'EMD correspond à l'impact positif minimal que doit avoir le dispositif étudié pour pouvoir être détecté comme tel avec une probabilité conventionnelle $\pi$ fixée à un niveau suffisamment élevé. Cette probabilité conventionnelle, qui correspond à la puissance statistique, est le plus souvent fixée à 80 %. L'EMD est en général présenté en unités d'écart-type de la variable d'intérêt à des fins de comparaison. Pour une probabilité conventionnelle $\pi$ donnée, un EMD faible est un indicateur de forte puissance d'une évaluation. En effet, celle-ci permet dans ce cas de détecter des effets d'ampleur même limitée. Au contraire, un EMD élevé est le signe d'une puissance faible de l'évaluation, puisqu'elle n'est susceptible de détecter que des effets de grande ampleur.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

Afin de s'assurer que le dispositif expérimental permettra d'identifier les effets du traitement avec une probabilité suffisamment élevée, il est impératif que des calculs de puissance soient réalisés *ex ante* – c'est-à-dire avant le début de l'évaluation. Comme ils mettent en jeu des notions statistiques parfois complexes, ces calculs sont réalisés par l'équipe en charge de l'évaluation.

Même si le calcul de puissance repose sur un certain nombre d'hypothèses difficiles à vérifier avant le début de l'évaluation, il permet d'avoir une idée des objectifs en termes de taille de l'échantillon, de taux de réponse et de participation qui seront nécessaires pour la détection des effets. Ce calcul conditionne donc souvent la faisabilité d'une évaluation.

***Assignation aléatoire au niveau individuel***

La formule (6) (en annexe 2) indique qu'il existe trois éléments essentiels à prendre en compte dans le calcul de puissance.

Premièrement, l'évaluateur doit porter une attention particulière au taux de conformité. En fonction du protocole d'évaluation, il est possible que la conformité soit imparfaite : certains individus du groupe test ne sont pas traités alors que certains individus du groupe témoin le sont. Ceci a des implications sur la précision de l'évaluation, qu'il faut prendre en compte lors du calcul de puissance. On définit donc le taux de conformité, qui correspond à la différence entre le taux de participation dans le groupe test ($p^1$) et le taux d'individus témoins traités ($p^0$). Le taux de conformité ($p^1 - p^0$) est une estimation du ratio de *compliers* contenus dans les groupes expérimentaux. Une diminution du taux de conformité se traduit par une détérioration de la précision de l'estimateur de l'effet du traitement. Il en résulte une augmentation de l'EMD. Notons que, contrairement aux autres facteurs sous une racine (comme la taille d'échantillon par exemple), le taux de conformité joue proportionnellement sur l'EMD : il affecte donc plus fortement la puissance que tous les autres facteurs. Le taux de conformité est généralement estimé en fonction d'études antérieures ou tiré de l'expérience des partenaires responsables de la mise en œuvre du programme.

Deuxièmement, lorsque la taille de l'échantillon diminue, l'EMD augmente : là encore, le pouvoir de détection de l'évaluation se réduit. L'évaluateur devra maximiser la taille de l'échantillon sous contrainte budgétaire : le coût marginal d'une observation supplémentaire étant faible pour l'évaluation, cette contrainte portera essentiellement sur le coût de la mise en œuvre du programme évalué. Cependant, la taille de l'échantillon dépend également du nombre de personnes volontaires pour participer au programme évalué. Il est souvent difficile d'estimer l'ampleur de ce volontariat en amont et celui-ci est parfois surestimé par le partenaire qui met en place le programme. C'est notamment le cas pour les dispositifs d'insertion professionnelle destinés aux jeunes qui, lorsqu'ils sont mis en place à grande échelle, attirent un public bien inférieur à leur public cible. Ceci est dû notamment à la présence de nombreux autres programmes disponibles sur le territoire. Il est important que l'évaluateur intègre cette contrainte opérationnelle dans son calcul de puissance et opte pour des hypothèses plutôt conservatrices.

Enfin, le taux de tirage (au sens de la proportion $\tau$ d'individus assignés au groupe test dans l'échantillon) joue un rôle dans le calcul de puissance. La minimisation de la variance de l'estimateur de l'effet du programme sur $\tau$ dans l'équation (5) de l'annexe 2 donne une valeur optimale de $\tau = 0{,}5$. La puissance sera donc maximale pour une probabilité de tirage égale à 50 %.

Une manière peu coûteuse d'augmenter la puissance réside dans la réduction de la variance du modèle d'évaluation ($\sigma$). Plus la population étudiée est homogène, plus il est facile de détecter une différence entre le groupe test et le groupe témoin. Après le tirage au sort assignant les individus dans les deux groupes expérimentaux, il est possible de diminuer la variance de l'estimateur de l'effet du traitement en contrôlant par des caractéristiques non affectées par le traitement mais qui sont susceptibles d'influencer la variable de résultat. Le plus pertinent est de considérer les variables collectées à l'enquête initiale(11) et, dans l'idéal, la valeur initiale de la variable de résultat (Bloom *et alii*, 2007). De même, lors du tirage au sort, la variance de l'estimateur peut être réduite par stratification de l'échantillon. Au lieu d'opérer un tirage sur l'échantillon total, l'évaluateur affecte aléatoirement les individus à l'intérieur de cellules définies par des strates. La stratification assure un parfait équilibre entre les groupes test et témoin sur ces variables de strates. Plus les strates sont corrélées à la variable d'intérêt, plus l'impact sur la puissance est fort. Par exemple, en éducation, une stratification par sexe et par compétences initiales permet de capter une partie non négligeable de la variance initiale.

En pratique, il est fréquent qu'on ne parvienne pas à collecter des données sur l'ensemble des unités de l'échantillon, soit en raison de la non-réponse aux enquêtes de suivi, soit parce que les données administratives utilisées sont incomplètes. Ceci entraîne une baisse de la puissance statistique et doit donc être pris en compte dans le calcul de puissance. Selon l'égalité (6) de l'annexe 2, si la non-réponse est équilibrée entre groupe test et groupe témoin et que les non-répondants affichent des comportements de participation dans le traitement similaires à ceux des répondants, alors la précision est affectée simplement par le facteur $1/\sqrt{N}$. La taille d'échantillon étant placée sous une racine, l'effet de

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

la non-réponse est donc marginal pour des taux d'attrition faibles mais de plus en plus problématique à mesure que la non-réponse augmente. Ainsi, sur un échantillon de 1 000 unités, les dix premières unités non-répondantes ont un effet de 0,5 % sur l'EMD final, alors qu'une non-réponse affectant 100 unités a un impact d'un peu plus de 5 %. Il est généralement admis qu'une attrition anticipée supérieure à 10 % doit être prise sérieusement en considération dans le calcul de puissance. Si les non-répondants sont en moyenne différents des répondants, on s'attend à un effet plus fort de la non-réponse sur la précision statistique. En effet, si le taux de non-réponse diffère selon les groupes, l'effet du terme $1/\tau(1-\tau)$ de (6) s'ajoute à l'impact de la taille de l'échantillon. Plus grave encore, si l'attrition affecte principalement les participants du groupe test ou les non-participants du groupe témoin, alors le terme $1/(p^1 - p^0)$ de (6) a un effet négatif proportionnel à la non-réponse.

***Assignation aléatoire au niveau du groupe***

Lorsque l'assignation aléatoire est effectuée non au niveau de l'individu mais au niveau du groupe (école, quartier), on dit qu'il y a tirage par grappe. Il est alors essentiel de prendre en compte dans l'analyse le fait que les termes d'erreurs des individus au sein d'un groupe peuvent ne pas être indépendants. En effet, les enfants d'une même école ou d'un même quartier peuvent être soumis à des chocs communs, ont souvent des caractéristiques communes et entretiennent ensemble des relations qui les conduisent à agir conjointement. Il existe donc souvent une corrélation à l'intérieur des grappes qui est supérieure à celle observée entre individus de groupes différents. La corrélation intra-grappe $\rho$, égale à la proportion de la variance totale (cumul des variances intra et inter-grappes) du terme d'erreur du modèle d'évaluation expliquée par la variance intra-grappe (supposée pour simplifier identique pour toutes les grappes) doit être prise en compte dans le calcul de la puissance (voir modèle correspondant en fin d'annexe 2).

D'après l'équation (7) de l'annexe 2, si $\rho = 0$, c'est-à-dire s'il n'existe pas de corrélation à l'intérieur des grappes, alors (7) revient à (6) et un tirage au niveau des grappes n'influence pas la puissance. Inversement, lorsque $\rho = 1$, c'est-à-dire que les individus à l'intérieur de chaque grappe sont identiques en termes de résultat *Y,* alors $\rho$ pèse sur la puissance de manière maximale. De même, le niveau de $\rho$ détermine l'influence que le nombre d'individus tirés par grappe aura sur la puissance : lorsque $\rho$ est élevé, le nombre d'individus à l'intérieur de la grappe influence relativement peu la puissance, beaucoup moins en tout cas que le nombre total de grappes. À taille d'échantillon donnée, il est donc beaucoup plus avantageux d'augmenter le nombre de grappes ($G$) plutôt que le nombre d'unités par grappes ($n$).

Le niveau de la corrélation intra-grappe dépend des interactions existant entre les membres d'une même grappe. Il peut être estimé à partir d'études antérieures. Empiriquement, d'après notre expérience opérationnelle, l'ordre de grandeur de $\rho$ est le plus souvent compris entre 5 % et 20 %. Une valeur raisonnable au vu des expérimentations antérieures est de 10 %. Néanmoins $\rho$ dépend de la manière dont les grappes sont constituées. Plus les grappes diffèrent, plus $\rho$ augmente.

Le niveau d'assignation aléatoire choisi influence donc fortement le niveau de précision statistique de l'évaluation. Un tirage au niveau des grappes, parfois plus simple politiquement et éthiquement, présente une puissance statistique réduite, d'autant plus que la corrélation intra-grappe est élevée.

***Exemple de calcul de puissance***

Prenons l'exemple d'un programme d'accompagnement des jeunes dans la réalisation de leurs projets, les Groupements de Créateurs, dont l'évaluation est en cours[12]. Un indicateur du succès de ce programme est le taux de "situations positives" des jeunes. La situation d'un jeune $i$ est considérée comme "positive" ($Y_i = 1$) si ce dernier est en contrat salarié pour plus de six mois, en formation qualifiante ou en situation de création d'activité, par opposition au chômage ou à des contrats salariés de courte durée ($Y_i = 0$). Le taux de "situations positives" est donc un indicateur d'insertion à moyen terme sur le marché du travail.

Dans le cadre de l'évaluation de ce dispositif, des calculs de puissance ont été réalisés en amont par l'équipe d'évaluation pour déterminer la taille d'échantillon acceptable, sous les hypothèses suivantes (en reprenant les notations de l'annexe 2) :

– un test bilatéral de $H_0 : \delta = 0$ contre $H_a : \delta \neq 0$ au seuil $\alpha = 5$ % (risque d'erreur de type I), avec une puissance de 80 % (soit un risque d'erreur de type II de 20 %) ;

– une assignation aléatoire au niveau des individus, avec une probabilité de 0,5 d'être affecté dans chacun des groupes ;

– un taux de réponse estimé à 80 %. Le taux de réponse est un enjeu crucial pour ce type d'évaluation, qui se fonde sur des enquêtes téléphoniques auprès d'un public difficile à contacter ;

– la variable de résultat étant dichotomique, sa variance théorique est égale à $p(1-p)$, où $p = \Pr(Y_i = 1)$. À partir du suivi effectué par le partenaire, on estime que 60 % des jeunes à la sortie de Groupements de Créateurs sont dans une "situation positive". On peut donc estimer approximativement la variance de la variable de résultat par : $0{,}6 \times 0{,}4 = 0{,}24$ (ce que l'on retient en première approximation comme ordre de grandeur pour le $\sigma$ de la formule de calcul (6) de l'EMD, en annexe 2).



Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

Le tableau 1 présente l'effet minimum détectable en points de pourcentage[13] obtenu en appliquant la formule (6) sous ces hypothèses pour différentes valeurs de taille d'échantillon et de taux de conformité.

**Tableau 1 : calculs effectués en amont de l'évaluation des Groupements de Créateurs de l'effet minimum détectable, selon la taille de l'échantillon et le taux de conformité**

| Taux de conformité (%) | Taille d'échantillon | | | | |
|---|---|---|---|---|---|
| | 1 200 | 1 000 | 800 | 600 | 400 |
| 100 | 8,85 | 9,70 | 10,84 | 12,52 | 15,34 |
| 85 | 10,42 | 11,41 | 12,76 | 14,73 | 18,04 |
| 70 | 12,65 | 13,86 | 15,49 | 17,89 | 21,91 |

Lecture : les EMD sont exprimés en points de pourcentage.
Source : auteurs.

Avec un effectif de 1 200 jeunes et un taux de conformité de 85 %, l'impact du dispositif devrait accroître le taux de situations positives de plus de 10,4 points pour que l'évaluation ait au moins 80 % de chances de conclure à un impact significatif, ce qui signifie que le taux de situations positives devrait passer de 60 % à environ 70 %[14]. Remarquons que la progression des effets minima détectables n'est pas linéaire avec la baisse de l'effectif. En effet, une réduction de la taille de l'échantillon de 1 200 à 800 jeunes serait moins coûteuse qu'une diminution de 800 à 400 jeunes. Notons aussi le très fort effet du taux de conformité sur l'EMD. Ce taux est un élément central de la détection des effets. En particulier, le taux de conformité est inférieur à 100 % lorsque l'assignation d'un individu dans le groupe test ne garantit pas sa participation au programme. Or, même si les individus sont volontaires et éligibles au moment du tirage, il y a toujours une déperdition pour ce type de programme. Par expérience, les partenaires responsables de la mise en œuvre ont une tendance naturelle à surestimer la future participation des individus à leur programme. L'évaluateur doit donc effectuer un travail d'estimation de la participation avant le début de l'évaluation.

## Menaces sur l'analyse

La première partie de cet article a montré comment le tirage aléatoire permet, sous certaines conditions, d'estimer de manière non biaisée l'effet d'un dispositif. De ce fait, la méthode aléatoire est souvent considérée comme disposant d'une validité interne forte[15]. Il existe cependant des situations dans lesquelles les hypothèses centrales du modèle d'Angrist et Imbens ne sont plus respectées. Il s'agit de situations de violation de la validité interne. D'autres menaces pèsent sur l'évaluation aléatoire lorsque le dispositif évalué localement est généralisé. On parle alors de menaces sur la validité externe[15]. Les conditions qui permettent la généralisation de résultats locaux sont parfois difficiles à réunir. S'il est vrai que la plupart des menaces sur la validité de l'analyse décrites ci-après concernent toutes les méthodes d'évaluation (expérimentales et non expérimentales), certaines caractéristiques de la méthode aléatoire la rendent plus sensible à ces problématiques. Afin que le lecteur puisse se faire une idée précise des avantages et inconvénients de la méthode aléatoire par rapport aux autres méthodes, les menaces communes à toutes les méthodes seront distinguées de celles qui pèsent plus spécifiquement sur l'évaluation aléatoire.

### Menaces sur la validité interne

#### *Attrition et collecte de données*

Le problème de la qualité de la collecte des données n'affecte pas uniquement les évaluations aléatoires : toutes les techniques d'évaluation y sont sensibles. La méthode aléatoire revendiquant la qualité de sa validité interne, les chercheurs spécialisés dans ce domaine se sont cependant penchés très scrupuleusement sur ce problème. Une mauvaise collecte de données, quantitativement et qualitativement, a potentiellement deux conséquences : elle réduit la précision statistique (*cf.* deuxième partie, troisième sous-partie, intitulée "Puissance d'une évaluation") et peut engendrer l'apparition d'un biais d'estimation.

*La collecte de données : une tâche complexe*

Il existe une vaste littérature sur les facteurs importants à prendre en compte pour minimiser la non-réponse et améliorer la qualité des données. Tout dépend de la méthode de collecte employée (en face à face, par téléphone, par courrier, par courriel), des agents auprès desquels les données sont collectées (les individus concernés, les administrations, qui disposent déjà de données individuelles, les entreprises, les instituts de statistique, etc.) et de l'acceptation du protocole expérimental.

Si ces problèmes se retrouvent dans nombre d'autres techniques d'évaluation, le processus de collecte de données est sans doute plus délicat pour la méthode d'assignation aléatoire. En effet, les données utilisées dans une évaluation aléatoire sont, la plupart du temps, soit générées (passation de tests ou de questionnaires) soit collectées pour l'évaluation elle-même et dans son seul but. Il est rare (même si cela n'est pas exclu) de disposer par avance des données sur les personnes faisant partie de l'évaluation. Il est donc nécessaire de mettre en place une structure capable d'assurer la collecte de données et/ou de convaincre les partenaires et les personnes qui ont accès aux données de transmettre ces informations à l'équipe des évaluateurs. Dans le cadre des évaluations dans le domaine de

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

l'éducation, par exemple, les administrations sont souvent peu enclines à transmettre les informations nécessaires à l'évaluation de manière nominative. Un long travail de persuasion est alors nécessaire à la bonne conduite de l'évaluation. Concernant les évaluations dans le domaine de l'emploi, les évaluateurs ont souvent recours à des prestataires externes pour la passation de questionnaire par téléphone. Ces prestataires sont peu habitués à ce type de demande, notamment à l'exigence en termes de qualité des données collectées et de taux de réponse. Par expérience, l'attrition s'avère assez forte pour ce type d'enquête notamment auprès d'un public jeune.

*Les conséquences sur la validité interne*

La validité interne de l'estimateur de l'effet du traitement est affectée s'il existe une corrélation entre le tirage aléatoire $T$ et 1) la probabilité de répondre ou 2) la qualité de cette réponse. Cette situation peut se produire lorsque le traitement est directement lié à la collecte de données. Par exemple, pour évaluer l'efficacité des internats d'excellence, un tirage aléatoire a été réalisé sur un ensemble de candidats présélectionnés. Les candidats du groupe test ont été autorisés à s'inscrire à l'internat alors que les candidats du groupe témoin ont dû trouver un autre établissement scolaire. Dans ce cas précis, la collecte de données auprès des internes est facilitée par le traitement lui-même : les internes sont généralement moins absents, changent moins souvent d'établissement et sont potentiellement moins décrocheurs. De plus, lors de la passation de tests de compétences scolaires, l'ensemble des internes peut être regroupé et surveillé facilement au sein de l'internat. *A contrario*, les élèves témoins sont scolarisés dans des établissements répartis sur tout le territoire, ce qui peut potentiellement affecter à la fois le taux de réponse et la qualité des données collectées les concernant. Si les conditions de passation des tests sont meilleures à l'internat que dans les autres établissements scolaires, l'évaluation risque une distorsion vers le haut de l'estimation. De même, si les élèves les plus en difficulté du groupe témoin quittent le système scolaire (et qu'aucune donnée ne soit alors collectée sur eux) pendant que ceux du groupe test parviennent à se maintenir à l'internat d'excellence, alors le résultat de l'évaluation risque une distorsion vers le bas.

Tester si la qualité des données collectées est comparable dans les groupes test et témoin n'est pas une tâche aisée. Isoler ce qui est causé par les conditions de collecte de l'information de ce qui correspond à l'effet du traitement est presque impossible. Si les élèves du groupe témoin passent moins de temps sur les épreuves de compétences scolaires, est-ce dû aux conditions de passation ou à un degré de motivation inférieur ?

*A contrario*, Il existe plusieurs tests permettant de détecter la présence d'un biais de non-réponse. On peut simplement poser :

$$A_{i,q} = \alpha_0 + \alpha_1 T_i + u_i$$

où $A_{i,q}$ désigne l'attrition observée pour l'individu $i$ au questionnaire/test $q$, $T_i$ l'assignation au traitement, $\alpha_0$ et $\alpha_1$ deux paramètres à estimer et $u_i$ le terme d'erreur. $A_{i,q}$ vaut 1 en cas d'attrition totale (correspondant à la non-réponse à l'ensemble du questionnaire) et 0 sinon. Si $\alpha_1 \neq 0$, alors il existe une attrition différentielle entre individus tests et témoins qui a toutes les chances de biaiser le résultat de l'évaluation. L'absence de rejet de l'hypothèse nulle $\alpha_1 = 0$ ne garantit cependant pas l'absence de biais. Il est possible par exemple qu'une attrition non différentielle concerne des sous-ensembles différents dans le groupe test et dans le groupe témoin. Pour tester ce biais, on peut effectuer le test suivant, dans le cadre du modèle :

$$Y_{i,b} = \beta_0 + \beta_1 T_i + \varepsilon_i \quad \forall i \text{ tel que } A_{i,q} = 0$$

où $Y_{i,b}$ est le résultat en enquête initiale $(b)$[16], $T_i$ l'assignation au traitement, $\beta_0$ et $\beta_1$ deux paramètres à estimer et $\varepsilon_i$ le terme d'erreur, l'échantillon étant restreint aux seuls répondants à l'enquête finale[17]. Comme, en vertu du tirage aléatoire, les groupes sont initialement équilibrés, un $\beta_1$ significativement différent de 0 indique un biais d'attrition : l'effet du traitement sera biaisé vers le bas si $\beta_1 < 0$ et biaisé vers le haut si $\beta_1 > 0$.

La méthode aléatoire présente donc l'avantage de disposer d'un test robuste et parfaitement identifié de la présence d'un biais d'attrition, un test qui ne présente pas les mêmes qualités lorsque d'autres méthodes sont utilisées. Cela étant dit, les techniques utilisées pour résoudre le biais sont similaires en évaluation aléatoire et dans d'autres méthodes. Tout au plus, l'existence d'un test robuste et non biaisé de la présence d'un biais d'attrition a sans doute davantage sensibilisé à cette menace les chercheurs pratiquant les évaluations aléatoires.

*Techniques de redressement de l'attrition*

Il existe de nombreuses techniques permettant de redresser le biais de sélection. Ces techniques nécessitent cependant des hypothèses parfois fortes. Un effort considérable doit donc être entrepris pour minimiser en amont la probabilité de non-réponse. Cela peut passer par une diversification des voies par lesquelles les informations sont collectées. On peut par exemple prévoir différentes versions d'un même questionnaire afin qu'il soit administré à la fois au téléphone, par papier, sur internet, par l'intermédiaire des réseaux sociaux, etc. Suivant les cas, une incitation financière peut être appropriée, notamment lorsque les conditions de collecte de l'information impliquent un coût (psychologique ou



Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

d'opportunité) important. Bien entendu, ces différents modes de collecte de l'information ainsi que les incitations financières peuvent potentiellement générer des biais de collecte qu'il faudra prendre en compte dans l'analyse. Par exemple, les tests passés à la maison ne captent sans doute pas la même information que ceux passés à l'école. Un juste équilibre doit donc parfois être trouvé entre le biais qualitatif et le biais quantitatif.

Lorsque tous les efforts ont été entrepris pour collecter les informations sur l'ensemble de l'échantillon et que le taux de réponse reste inférieur à 100 %, une procédure de *suivi intensif* peut être mise en place. Il s'agit de sélectionner aléatoirement un sous-ensemble d'individus pour lesquels tous les efforts possibles ainsi que tous les moyens disponibles seront mobilisés, l'objectif étant de parvenir à un taux de réponse sur ce sous-ensemble de 100 %. Lors de l'analyse, il s'agira alors de pondérer ce sous-ensemble de répondants proportionnellement à son taux de tirage (c'est-à-dire à la proportion de l'échantillon de non-répondants qu'il représente). D'autres techniques paramétriques et non paramétriques de redressement de l'attrition existent. On notera l'utilisation fréquente des bornes dites de Horowitz-Manski (Horowitz et Manski, 2000) ou Lee (Lee, 2009), qui permettent d'établir des bornes supérieures et inférieures de l'effet du traitement. Le modèle de Heckman (1979) est plus rarement utilisé à cause des problèmes d'identification qu'il pose. La méthodologie du suivi intensif mentionnée plus haut s'inspire cependant de ce travail.

Comme évoqué précédemment, lorsque le groupe témoin est construit *ex post* – comme c'est le cas lorsque les participants au traitement sont appariés avec des non-participants (régressions multivariées ou appariement ; *cf.* Givord, 2010-2014) – la question de l'attrition est sans doute moins cruciale. En effet, les participants restent appariés avec leurs plus proches homologues non-participants.

***Impact du tirage sur le comportement des agents***

Comme évoqué en première partie de l'article, la validité de l'inférence causale repose sur l'hypothèse d'indépendance $Y_i^0, Y_i^1 \perp T$. Si l'assignation aléatoire permet dans la plupart des cas d'assurer cette condition, il existe au moins deux situations bien connues des évaluateurs dans lesquelles la simple assignation aléatoire modifie les comportements des individus indépendamment de l'effet du dispositif. Les élèves du groupe test peuvent se sentir valorisés d'avoir été sélectionnés dans le dispositif alors que la déception éprouvée par le groupe témoin peut modifier son comportement. Ces effets affectent particulièrement l'évaluation de politiques publiques car celles-ci ne peuvent être comparées à des mesures placebo (comme en médecine) et le comportement des individus (les croyances, la motivation, etc.) constitue une part non négligeable de l'efficacité du traitement. Dans la littérature, la modification du comportement du groupe test sous l'effet du traitement est connue sous la dénomination de l'effet Hawthorne, celle du groupe témoin comme l'effet John Henry.

Ces phénomènes comportementaux peuvent modifier l'interprétation des résultats et/ou être des vecteurs de biais d'analyse. Si le seul fait d'être acceptés à l'internat d'excellence modifie le comportement des individus, alors l'effet de Hawthorne n'est pas à proprement parler un biais d'analyse. Il peut être considéré comme un effet du dispositif. Seule est alors modifiée l'interprétation des résultats. Dans ce cas, l'estimateur de l'impact du traitement capte l'effet direct du dispositif (meilleur environnement scolaire) auquel s'ajoute un effet Hawthorne (meilleure motivation). Dans d'autres cas, la modification du comportement peut poser des problèmes plus délicats. Si les individus ont un intérêt direct à orienter les conclusions de l'évaluation, alors il est possible que la simple assignation à un groupe expérimental modifie leur comportement. Des enseignants pourraient par exemple augmenter artificiellement les notes qu'ils donnent aux élèves. De même, des sociétés de microcrédit pourraient dissimuler l'arrêt des remboursements.

L'ampleur du biais généré par ces problèmes de comportements dépend donc de façon centrale du type de dispositif évalué et du contexte dans lequel il est mis en œuvre. Notons que ce type de problème ne concerne pas uniquement la méthode aléatoire et qu'il n'existe pas d'outil de mesure de l'ampleur des biais engendrés par la modification du comportement des agents. L'ampleur des biais générés par les effets Hawthorne et John Henry doit être évaluée au cas par cas, suivant le contexte, la capacité et l'intérêt des acteurs à modifier leurs comportements.

***Non-respect du protocole d'assignation aléatoire***

Comme évoqué en première partie de l'article, la conformité imparfaite ne se traduit pas à proprement parler par un biais statistique. Il est en effet toujours possible de retrouver l'effet du traitement sur les traités en utilisant l'assignation aléatoire comme instrument de la participation (encadré 2). De même, la forme réduite (ITT) reste interprétable comme l'effet global du traitement. Tout ceci reste juste à condition que les comportements de participation (et de non-participation) correspondent à ceux qui auraient été observés en l'absence d'expérimentation.

Prenons par exemple le cas d'une politique éducative visant à proposer à des écoles tirées aléatoirement une ligne budgétaire pour assurer des cours du soir pour les élèves en difficulté. Imaginons qu'une erreur lors de la mise en œuvre du programme ait conduit à proposer le budget à deux écoles témoins.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

Lors de l'analyse, suivant la stratégie empirique décrite en première partie, la forme réduite (ITT) comparera donc les groupes test et témoin tels que définis lors du tirage, alors même que deux écoles témoin sont en réalité traitées. Bien évidemment, l'effet du traitement estimé en ITT sera plus faible que si les deux écoles n'avaient pas été traitées. Dans ce cas-là, il existe bien une distorsion vers le bas causée par le non-respect du protocole expérimental.

Il revient donc à l'évaluateur d'accorder une attention toute particulière au respect du protocole. Ainsi, l'évaluation aléatoire nécessite pour l'évaluateur d'être présent dès le début de la mise en œuvre du programme et implique qu'il entretienne des relations très étroites avec les partenaires mettant en place le projet évalué. La mise en place et le respect d'un protocole expérimental présentant nécessairement des contraintes pour les partenaires, il est absolument capital que ceux-ci comprennent parfaitement l'intérêt de l'évaluation et s'y prêtent volontairement. L'évaluation aléatoire ne peut pas être conçue comme une manière de sanctionner ou de noter une association ou une politique particulière. Elle doit au contraire orienter les choix politiques sur des bases scientifiques. Le non-respect du protocole expérimental constitue bien évidemment la menace la plus sérieuse qui pèse sur les évaluations aléatoires. Cette menace leur est en grande partie spécifique.

***Externalités***

Cette question a déjà été évoquée dans la première partie de l'article lors de la discussion sur l'hypothèse SUTVA. Cette hypothèse suppose que le résultat final et le comportement de participation de l'individu *i* ne soient induits que par l'assignation individuelle dans un groupe de tirage et non par le résultat final ou le comportement des autres individus de l'échantillon. En présence d'externalités[18], c'est-à-dire d'effets indirects du traitement d'un individu sur un individu tiers, l'hypothèse SUTVA est potentiellement invalidée. Si l'externalité n'affecte que les individus à l'intérieur du groupe test, alors elle modifie l'interprétation du coefficient d'intérêt mais ne crée pas de biais à proprement parler. En effet, l'effet du traitement estimé inclura alors à la fois l'effet direct mais aussi l'effet de contexte généré par l'évaluation. Dans bien des cas, le décideur politique considérera que c'est cet effet global (direct plus indirect) qui l'intéresse. Notons cependant que, pour éviter un biais d'estimation, aucun individu témoin ne doit être traité, sinon les effets externes se diffuseraient au sein des autres membres du groupe témoin. Notons également qu'en présence d'externalité au sein du groupe test, il n'est pas possible d'estimer l'effet moyen sur les traités (TOT) car une telle estimation conduirait à attribuer aux seuls participants l'ensemble de l'effet généré par la politique (externalités incluses).

Il existe des situations moins favorables où des effets externes des individus traités se diffusent aux individus non traités. Dans de tels cas, la validité interne de l'évaluation est menacée. Dans leur article d'évaluation de l'effet d'un traitement contre les vers intestinaux au Kenya, Miguel et Kremer (2004) notent qu'un tirage aléatoire au niveau individuel affecterait à la fois les élèves tests et, par diffusion, les élèves témoins. On observerait alors un groupe témoin plus faiblement exposé aux vers intestinaux par le simple fait que les élèves tests en seraient protégés. L'une des manières de résoudre cette difficulté est d'effectuer un tirage à un niveau supérieur (au niveau régional, au niveau de l'école) et de s'assurer que les individus de ces groupes n'interagissent pas entre eux (en éloignant par exemple les groupes les uns des autres). En faisant varier aléatoirement le nombre d'écoles tests dans une même région, le protocole adopté par Miguel et Kremer (2004) permet de mesurer l'ampleur de l'externalité.

Il est également possible de distinguer l'effet direct de l'effet de diffusion en adoptant par exemple un *partial population model* dans lequel, avant tirage, les évaluateurs définissent un groupe d'individus dits présélectionnés. Cette présélection peut être effectuée par l'évaluateur lui-même (par exemple en préférant réserver la politique à des individus d'un sexe particulier ou en excluant de la politique des personnes d'un certain âge) ou elle peut s'opérer naturellement à travers un système de candidature ou de volontariat (l'évaluateur laisse alors les individus s'inscrire à la politique librement). Le tirage est ensuite effectué à un niveau supérieur (école, entreprise). Dans ce cadre, l'effet direct du traitement se calcule en comparant les individus présélectionnés des deux groupes. La comparaison des élèves non présélectionnés dans les groupes test et témoin permet de capter l'effet des présélectionnés tests sur leurs pairs (les non présélectionnés). Ce protocole, employé dans l'évaluation du dispositif dit de la Mallette des parents (Avvisati *et alii*, 2010), avec un tirage au niveau des écoles, a permis d'identifier des effets de diffusion forts au sein des classes.

Cette menace ne se limite évidemment pas aux seules évaluations aléatoires : toute diffusion des effets entre les groupes affectés par la politique et ceux qui ne sont pas directement affectés a pour conséquence de modifier l'ampleur de l'effet du traitement quelle que soit la méthode d'estimation utilisée.

**Menaces sur la validité externe**

Les évaluations aléatoires assurent – lorsqu'elles sont menées de manière rigoureuse – qu'au sein de l'échantillon expérimental, l'estimation est très proche du réel effet de la politique évaluée : on dit alors que l'évaluation aléatoire dispose d'une bonne validité interne. Cependant, rien ne permet *a priori* d'assurer que ces estimations locales ont une



Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

pertinence générale – c'est-à-dire une bonne validité externe. Deux sortes de menaces pèsent sur la validité externe : le problème lié à l'équilibre général et celui lié à la généralisation d'effets locaux.

### *Effet d'équilibre général*

En relation avec la question de l'externalité se pose le problème de l'équilibre général. Alors que les externalités affectent la validité interne de l'évaluation, les effets d'équilibre général affectent la validité externe. Lorsqu'un dispositif est généralisé, il peut produire des effets de bouclage macroéconomiques ou macrosociaux susceptibles de perturber le résultat évalué localement. Certains programmes pertinents d'un point de vue microéconomique (baisse du coût du travail) ont des conséquences macroéconomiques potentiellement perverses (baisse de la demande globale).

Des programmes d'éducation ayant localement des effets positifs peuvent s'avérer inefficaces lorsqu'ils sont généralisés. Revenons à l'exemple du programme "Éclair" évoqué plus haut. Imaginons que, pour évaluer cette politique, une équipe de recherche autorise certains établissements à recruter les enseignants nouvellement diplômés sur profil. Les premiers établissements tests seront peut-être capables de recruter des candidats de meilleure qualité que les établissements témoins, ce qui serait susceptible d'améliorer leur fonctionnement. Cet effet local, qui pourrait être évalué par la méthode aléatoire et qui disposerait donc d'une validité interne forte, n'aurait cependant aucune validité externe. En effet, au fur et à mesure que le nombre d'établissements disposant de cette liberté augmenterait, la concurrence entre établissements pour recruter les meilleurs candidats s'intensifierait. À dotation égale, l'effet de cette politique serait sans doute marginal. Si les salaires étaient laissés libres et que les compétences pédagogiques des enseignants étaient facilement mesurables (ce qui n'est typiquement pas le cas), une augmentation du salaire des meilleurs enseignants et une baisse de la rémunération des moins bons seraient probables. Cette politique pourrait avoir des effets bénéfiques (en augmentant par exemple la motivation et l'effort des enseignants et des élèves) mais risquerait aussi d'avoir des effets pervers, en générant une plus grande inégalité sur le territoire entre les établissements capables d'attirer les meilleurs enseignants et les autres. En tout état de cause, l'effet local capté par l'expérimentation ne correspond absolument pas à l'effet général du dispositif.

Une des manières d'identifier l'effet d'équilibre général serait d'appliquer la stratégie employée par Miguel et Kremer (2004) sur des zones éducatives. C'est ce à quoi s'est employée une équipe de recherche qui évaluait l'impact d'une politique de retour à l'emploi sur les jeunes diplômés (Crépon *et alii*, 2013). Il s'agissait d'évaluer une amélioration du dispositif de suivi des demandeurs d'emploi mis en place par Pôle Emploi. Cette politique est particulièrement concernée par le problème d'équilibre général dans la mesure où un résultat positif localement mesuré sur un petit échantillon d'individus pourrait ne traduire qu'une accélération du retour à l'emploi par rapport aux jeunes non traités et non pas une création nette d'emplois. Si chaque individu de ce même bassin d'emplois disposait de la même technologie de recherche, l'effet du traitement pourrait bien être nul. Certes, en favorisant un meilleur appariement entre demandeur et entreprise, il est possible que cette politique ait des effets économiques importants. Toutefois, le protocole expérimental de tirage individuel ne permet en aucun cas de le démontrer. Afin d'identifier l'ampleur de ce phénomène, les chercheurs ont effectué un tirage en deux étapes. Ils ont tout d'abord assigné à chacun des bassins d'emplois un pourcentage de jeunes concernés par le dispositif (0 %, 25 %, 50 %, 75 %, 100 %). Ensuite, à l'intérieur de chacun des bassins d'emplois, ils ont affecté aléatoirement le dispositif à un nombre de jeunes correspondant au taux de traitement du bassin d'emplois considéré. La comparaison des groupes test et témoin dans chaque bassin d'emplois identifie l'effet local du dispositif alors que les comparaisons des groupes témoins entre bassins permet de détecter l'existence d'un effet d'équilibre général.

### *Généralisation d'estimations locales*

Outre le problème de l'équilibre général, il n'est pas certain qu'un effet estimé dans des conditions locales puisse avoir une pertinence globale. En effet, les conditions de participation (qui sont les *compliers* ?), la manière dont la politique a été mise en œuvre, les conditions économiques et sociales locales peuvent être différentes lorsque la politique est généralisée. Pour que les résultats d'une évaluation puissent avoir une pertinence plus générale, l'évaluateur doit porter attention à différents éléments :

– il est tout d'abord essentiel de vérifier dans quelle mesure l'échantillon de l'évaluation est représentatif de la population cible du programme. La validité externe des résultats dépend fortement de cette condition. Idéalement, il faudrait recourir aux méthodes classiques d'échantillonnage. En pratique, néanmoins, il est plus simple pour le partenaire et l'évaluateur de se concentrer sur une zone géographique donnée. Il faut alors vérifier que les conditions locales observées dans cette région ont une pertinence globale ;

– ensuite, l'évaluateur doit s'intéresser aux individus qui ont été sensibles à l'assignation aléatoire (les *compliers*). Les *compliers* dans la population cible totale sont sans doute différents de ceux identifiés lors d'une évaluation locale. Ce problème est d'autant plus ennuyeux que les programmes évalués par la méthode aléatoire sont innovants. L'équipe responsable de l'évaluation est susceptible de modifier la population des volontaires en améliorant la communication ou les incitations lors de la

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

généralisation du dispositif. Les *compliers* risquent alors d'être plus nombreux et donc différents de ceux observés au cours de l'évaluation locale. Il convient donc de préciser sur quelle sous-population les résultats sont estimés et dans quelle mesure ces résultats sont plus ou moins susceptibles de s'appliquer à la population cible générale ;

– l'équipe responsable de la mise en œuvre du dispositif au cours de l'évaluation n'est généralement pas comparable à celle qui mettra en œuvre le dispositif si ce dernier est généralisé. Les équipes responsables de la mise en œuvre du programme ont parfois à cœur de prouver son efficacité. Un tel niveau de motivation ne se retrouvera probablement pas lors de la généralisation. Inversement, les programmes évalués étant fréquemment innovants, le partenaire a souvent peu d'expérience lors des premières années de mise en œuvre. Pour assurer une validité externe, l'évaluation d'impact doit s'accompagner d'une évaluation de processus qui documentera la façon dont le programme a été mis en œuvre et qui permettra sa reproductibilité ;

– le contexte institutionnel et culturel peut également influencer le résultat estimé localement. Rien n'assure que l'effet estimé à une période *t* dans une région *r* est généralisable à d'autres contextes. Pour tenter de répondre à cette difficulté, il est essentiel de documenter au maximum le contexte, notamment institutionnel, dans lequel le programme a été mis en place.

Ces problèmes se posent également avec d'autres techniques d'évaluation, sur données préexistantes. Ces techniques s'appuient sur des variations exogènes de participation au traitement causées par une circonstance particulière et en estiment les effets sur un sous-ensemble de *compliers*. Néanmoins, parce qu'elles nécessitent un gros effort de coordination et de mise en œuvre, les évaluations aléatoires se restreignent parfois à des échantillons plus petits, sur des durées plus courtes ou sur des régions spécifiques. Inversement, les techniques d'évaluation non expérimentales peuvent disposer de bases de données plus fournies et donc plus représentatives de la population. Il est cependant difficile d'établir des généralités à ce sujet. Certaines évaluations aléatoires ont été menées sur de grandes bases de données. Par exemple, les évaluations aléatoires menées en partenariat avec Pôle Emploi impliquent généralement des bases de données de dizaines de milliers d'observations. D'autres évaluations aléatoires ont été menées sur de très longues périodes. L'évaluation aléatoire du programme STAR de réduction des tailles de classes s'est ainsi étendue sur une vingtaine d'années[19]. Un jugement au cas par cas de la validité externe des estimations obtenues doit là encore prévaloir.

L'une des solutions directes à ce problème de la validité externe est la réplication. Idéalement, pour s'assurer de la validité d'un résultat, il faudrait reproduire à l'identique la même évaluation dans plusieurs régions et à diverses époques avec des équipes différentes. En pratique, malheureusement, la complexité et le coût de la mise en œuvre de ces évaluations rendent leur reproduction à l'identique difficile. De plus, les évaluateurs chercheurs sont souvent enclins à modifier le protocole d'évaluation afin d'en tirer des résultats novateurs. Dans l'idéal, il serait donc nécessaire que des organismes d'évaluation non liés à la recherche mais formés rigoureusement aux techniques de l'évaluation soient créés pour poursuivre le travail mené par les laboratoires de recherche en répliquant à l'identique les protocoles d'évaluations. De telles structures destinées à favoriser ces réplications ne sont malheureusement pas encore disponibles dans les pays développés[20].

En complément de la nécessaire réplication des évaluations, il est important que les évaluations aléatoires apportent un éclairage sur les raisons pour lesquelles les interventions évaluées fonctionnent ou non. Ceci doit passer par la documentation précise des conditions dans lesquelles l'intervention s'est déroulée (qualité des acteurs qui l'ont mise en œuvre, qualité de l'intervention elle-même, type de public qu'elle a concerné, etc.). Ces informations descriptives serviront notamment à établir une mesure d' "intensité du traitement", qui permettra de lier les conditions de l'intervention avec la taille de l'effet.

### Évaluation aléatoire et théorie économique[21]

L'une des manières de contourner les problèmes pesant sur la validité externe est d'intégrer les résultats d'évaluation dans des modèles de comportements individuels ayant, eux, plus de chance d'être généralisables. Il s'agit alors d'utiliser l'évaluation aléatoire non plus comme une méthode de validation de politiques publiques mais plutôt comme une entreprise de recherche sur les comportements humains. Au lieu de tester différentes versions d'une politique semblable, on pourrait par exemple théoriser un certain nombre de comportements humains et les tester en utilisant un protocole expérimental adapté. Les articles d'évaluation récents s'orientent presque systématiquement dans cette direction. Par exemple, dans leur article sur l'introduction des classes de niveau au Kenya, Duflo, Dupas et Kremer (2011) décomposent les différents mécanismes comportementaux qui pourraient rendre compte de l'impact d'une telle politique. Si l'effet évalué du dispositif (effet positif des classes de niveau ne se faisant pas au détriment des élèves les plus faibles) ne peut pas être directement généralisé à d'autres contextes, les mécanismes de comportements des enseignants mis en lumière par l'évaluation octroient néanmoins aux résultats une portée

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

beaucoup plus large qu'au seul périmètre de l'expérience réalisée.

Les évaluations aléatoires permettent également de tester la théorie économique existante ou d'estimer des paramètres pouvant être ensuite intégrés à des modèles structurels. Tout modèle structurel comportemental nécessite, par exemple, de faire des hypothèses sur des élasticités clefs. En économie de la taxation (élasticité du revenu imposable à la taxation), en économie du travail (élasticité de l'offre de travail à une modification des salaires) ou en économie du développement (élasticité prix de biens ou de services sociaux ou de santé), l'estimation des élasticités est rendu complexe par l'endogénéité de l'évolution de la demande à celle des prix. Les évaluations aléatoires peuvent constituer une stratégie très intéressante pour estimer ces élasticités. Au Kenya, par exemple, des chercheurs (Cohen et Dupas, 2010) se sont intéressés à la question de l'élasticité de la demande de moustiquaires ainsi que de leur utilisation par rapport au prix. S'il existe un consensus sur le fait qu'un service ou un bien améliorant la santé doit être subventionné, il n'existe pas d'estimation précise de l'impact d'une augmentation des prix sur la demande ou sur l'utilisation de ce service ou de ce bien. Les chercheurs sont parvenus à estimer l'élasticité prix en faisant varier aléatoirement le prix de vente des moustiquaires. Leurs résultats aboutissent aux deux conclusions suivantes. D'une part, la demande de moustiquaires est très sensible au prix. En effet, elle diminue fortement lorsque les moustiquaires ne sont plus gratuites mais sont vendues, même à un prix relativement faible. D'autre part, l'intensité de l'utilisation des moustiquaires ne varie pas avec le prix. Une telle stratégie pourrait tout à fait être répliquée dans des pays développés où nos connaissances restent faibles sur l'impact des prix des services sociaux ou médicaux sur leur demande ou leur utilisation.

## Conclusion

La méthode d'assignation aléatoire présente des avantages évidents pour l'établissement de l'inférence causale dans le cadre de l'évaluation des politiques publiques. Dans de nombreux cas, cette méthode est même la seule à pouvoir établir un lien causal rigoureux entre le programme évalué et les effets mesurés. Si des menaces pèsent sur l'estimation de l'impact du programme, la plupart de ces menaces concernent également les autres méthodes d'évaluation. Les menaces sur la validité interne peuvent souvent être anticipées et prises en compte dans la conception du protocole d'évaluation. Lorsque les menaces sur la validité interne ont été prises en compte, l'enjeu principal de l'évaluation reste la validité externe, à laquelle une attention particulière doit être prêtée.

Par ailleurs, la méthode d'assignation aléatoire nécessite d'être en capacité, d'une part, d'effectuer un tirage au sort sur un nombre suffisant d'unités et, d'autre part, de faire respecter ce tirage. Il existe une grande variété de protocoles adaptés à des situations diverses, qui permettent d'élargir le champ d'application de la méthode d'assignation aléatoire à un grand nombre de politiques publiques. Cependant, la méthode reste inopérante dans certaines situations. Il est par exemple extrêmement difficile d'évaluer par cette méthode des politiques macroéconomiques : politique de taux de change, de taux d'intérêt, politique industrielle, etc. Il est également difficile d'évaluer par cette méthode des programmes considérés comme de première nécessité, car l'exclusion, même temporaire, de certains individus de ces programmes sous des conditions aléatoires ne serait pas acceptée. Enfin, pour s'assurer que l'assignation aléatoire soit réalisée dans de bonnes conditions et que le protocole soit respecté, il est recommandé que l'évaluateur soit présent avant et pendant le déploiement du programme, ce qui réduit la portée de cette méthode pour évaluer des dispositifs déjà en place. Dans la perspective de fonder les politiques publiques sur des preuves scientifiques, la coexistence de techniques d'évaluation expérimentales et non-expérimentales se justifie donc pleinement.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

## Notes

(1) Des découvertes aussi importantes que celles de la pénicilline ou d'autres antibiotiques ont été impulsées par des découvertes empiriques pures.

(2) *Cf*. Angrist et Imbens (1994 ; 1996) ou Givord (2010-2014), dans ce numéro.

(3) Voir, par exemple, les travaux financés par le Fond d'Expérimentation pour la Jeunesse.

(4) Voir, notamment, l'article publié par le réseau *European Expert Network on Economics of Education* (Bouguen et Gurgand, 2012).

(5) En d'autres termes, l'espérance de la différence de résultat entre le groupe test et le groupe témoin en l'absence de traitement est égale à zéro.

(6) Nous verrons que cela dépend de la nature de la conformité imparfaite (*cf*. troisième partie, première sous-partie, troisième alinéa, intitulé "Non-respect du protocole d'assignation aléatoire"). Nous verrons également que la généralisation du comportement de participation dépend de la validité externe de l'évaluation (*cf*. troisième partie, deuxième sous-partie, intitulée "Menaces sur la validité externe").

(7) Nous reviendrons sur cette question dans la troisième partie de l'article.

(8) Pour plus de détails sur les conséquences des externalités et les manières de les prendre en compte dans l'analyse se reporter à la troisième partie, première sous-partie, quatrième alinéa, intitulé "Externalités". Le lecteur pourra aussi consulter l'article de Bléhaut et Rathelot (2014), dans ce numéro, pour voir comment les auteurs traitent ce problème.

(9) Cet exemple est tiré de Bouguen et Gurgand (2012).

(10) Le lecteur intéressé peut se reporter à cet égard à l'évaluation menée par le laboratoire J-PAL Europe (https://www.povertyactionlab.org/fr/evaluation/sourdun-internats-excellence).

(11) L'enquête initiale consiste à collecter, avant le tirage, des informations auprès de l'ensemble de l'échantillon (groupes test et de contrôle) : tests scolaires, cognitifs, non cognitifs, chômage, revenu, salaire, etc. Cette enquête est facultative dans une évaluation aléatoire mais elle s'avère utile pour détecter l'attrition différentielle, gagner en puissance ou identifier des effets sur des sous-populations (effets hétérogènes) – sur l'attrition, voir *infra* pour plus de détails.

(12) Pour plus d'informations sur cette évaluation, se reporter à : http://www.povertyactionlab.org/evaluation/groupements-de-créateurs-encouraging-youth-entrepreneurship-france.

(13) Dans cet exemple, l'EMD est exprimé en points de pourcentage. Cela a été rendu possible par le fait que la variable de résultat est une variable binaire. Nous avons pu estimer la variance $\sigma^2$ sur la base de la formulation théorique de la variance d'une variable binaire et d'une estimation de la probabilité que cette dernière soit égale à 1. Dans le cas de variables de résultat continues, il est souvent plus difficile de trouver une approximation de $\sigma^2$. Il est alors préférable d'exprimer l'EMD en pourcentage d'écart-type (c'est-à-dire de calculer EMD/$\sigma$), sans se soucier à ce stade de la valeur que prendra l'écart-type $\sigma$ (*cf*. annexe 2). Comme l'impact final du programme sera souvent exprimé en pourcentage d'écart-type, cette pratique est tout à fait acceptable.

(14) Le même type de calcul peut être effectué pour des variables de résultat continues en utilisant comme valeur d'écart-type estimé des chiffrages provenant d'autres études. Par exemple, si l'on utilisait un test non-cognitif de confiance en soi pour mesurer l'effet du Groupement de Créateur, il serait possible de reprendre l'écart-type estimé lors de précédentes études utilisant ce même test. Il s'agirait d'une solution concurrente à celle qui est indiquée dans la note précédente (13), à savoir le calcul de l'EMD en pourcentage d'écart-type.

(15) La validité interne d'une évaluation réalisée à partit d'une expérimentation renvoie à la fiabilité des conclusions que l'évaluateur tire de l'analyse statistique des données fondée sur cette expérimentation. La validité interne se distingue de la validité externe, qui renvoie à la possibilité de généraliser les résultats d'une évaluation à l'ensemble de la population, au-delà du seul échantillon des données expérimentales.

(16) Pour rappel, l'enquête initiale est l'enquête conduite avant le tirage au sort auprès de l'ensemble de l'échantillon expérimental (test et témoin) - *cf*. note (11).

(17) L'enquête finale est l'enquête menée auprès de l'ensemble de l'échantillon expérimental (test et témoin) après la mise en œuvre de l'intervention. Le test d'attrition présenté est effectué après la réalisation de l'enquête finale. Par conséquent, on dispose alors de l'information sur les individus qui ont ou non répondu.

(18) Le terme externalité prend parfois des appellations différentes selon le contexte de l'évaluation. En économie de l'éducation, on parle souvent d'effet de pair alors que, sur le marché du travail, on parlera plutôt d'effet de diffusion. L'analyse de la ségrégation a vu l'apparition de l'expression "effet de contexte". Tous ces termes recoupent plus ou moins la même idée : un effet indirect d'une politique sur un individu non concerné par la politique.

(19) Voir Chetty *et alii* (2011).

(20) Dans les pays en développement, l'organisation *Innovations for Poverty Action* mène de nombreuses réplications : http://www.poverty-action.org/.

(21) Pour plus de détails sur la manière d'intégrer les évaluations aléatoires à des modèles théoriques, se reporter à White (2009).



Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

## Bibliographie


**Angrist J. et Imbens G. (1994).** "Estimation and Identification of the Local Average Treatment Effect", *Econometrica*, vol. 62, n° 2, pp. 467-475.

**Angrist J. et Imbens G. (1996).** "Identifying Causal Effect Using Instrumental Variables", *Journal of Statistical Association*, vol. 91, n° 434, juin, pp. 444-455.

**Avvisati F., Guyon N., Gurgand M. et Maurin E. (2010).** "Getting Parents Involved: A Field Experiment in Deprived Schools", CEPR Discussion Paper, n° 8020, 45 pages.

**Behaghel L., Crépon B., Gurgand M. et Le Barbanchon T. (2009).** "Sample Attrition Bias in Randomized Experiments: A Tale of Two Surveys", IZA Discussion Paper, n° 4162, 29 pages.

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

**Bloom H. (1995).** "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs", *Evaluation Review*, vol. 19, n° 5, October, pp. 547-556.

**Bloom H., Richburg-Hayes L. et Rebeck Black A. (2007).** "Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions", *Education Evaluation and Policy Analysis*, vol. 29, n° 1, pp. 30-59.

**Bouguen A. et Gurgand M. (2012).** "Randomized Controlled Experiments in Education", European Expert Network on Economics of Education (EENEE) Analytical Report, n° 11, février, 31 pages.

**Chetty R., Friedman J., Saez E., Whitmore Schanzenbach D. et Yagan D. (2011).** "How Does your Kindergarten Classroom Affect your Earnings: Evidence from Project STAR", *Quarterly Journal of Economics*, vol. 126, n° 4, pp. 1593-1660.

**Cohen J. et Dupas P. (2010).** "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment", *Quarterly Journal of Economics*, vol. 125, n° 1, pp. 1-45.

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

**Duflo E., Dupas P. et Kremer M. (2011).** "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya", *American Economic Review*, vol. 101, n° 5, pp. 1739-1774.

**Duflo E., Glennerster R. et Kremer M. (2006).** "Using Randomization in Development Economics Research: A Toolkit", NBER Technical Working Paper, n° 333, 87 pages.

**Erkel-Rousse H. (2014).** "Le calcul de puissance en assignation aléatoire : un point de vue de statisticien", Commentaire de l'article d'A. Bouguen et J. Seban, *Économie et Prévision*, n° 204-205, pp. 145-161, ce numéro.

**Gertler P. (2004).** "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment", *American Economic Review*, vol. 94, n° 2, pp. 336-341.

**Givord P. (2010, 2014).** "Méthodes économétriques pour l'évaluation de politiques publiques", Document de travail de l'Insee, Direction des Études et Synthèses Économiques, n°G2010/008, 74 pages, et *Économie et Prévision*, n° 204-205, pp. 1-28, ce numéro.

**Heckman J. (1979).** "Sample Selection Bias as a Specification Error", *Econometrica*, vol. 47, n° 1, pp. 153-161.

**Horowitz J. et Manski C. (2000).** "Non parametric Analysis of Randomized Experiments With Missing Covariates and Outcome Data", *Journal of the American Statistical Association*, vol. 95, n° 449, pp. 77-84.

**Lee D. (2009).** "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects", *Review of Economic Studies*, vol. 76, n° 3, pp. 1071-1102.

**Miguel E. et Kremer M. (2004).** "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities", *Econometrica*, vol. 72, n° 1, pp. 159-217.

**Rubin D. (1974).** "Estimating Causal Effect of Treatment in Randomized and Non Randomized Studies", *Journal of Education Psychology*, vol. 66, n° 5, pp. 688-701.

**White H. (2009).** "Theory-Based Impact Evaluation: Principles and Practice", *International Initiative for Impact Evaluation* (3ie) Working Paper, n° 3, New Delhi, 20 pages.


Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

## Annexe 1 : interprétation de l'estimateur TOT[(a)]

Dans toute cette annexe, nous reprenons les notations de l'encadré 2. Comme cela a été montré dans cet encadré, l'estimateur TOT s'exprime comme l'estimateur de Wald du paramètre $\delta$ dans le modèle (4), soit :

$$\hat{\delta}_{TOT} = \frac{\hat{E}(Y_i|T_i=1) - \hat{E}(Y_i|T_i=0)}{\hat{E}(P_i|T_i=1) - \hat{E}(P_i|T_i=0)}$$

Le numérateur de cet estimateur se décompose de la façon suivante :

$$\begin{aligned}\hat{E}(Y_i|T_i=1) - \hat{E}(Y_i|T_i=0) &= \hat{E}(P_i^1 Y_i^1 + (1-P_i^1)Y_i^0|T_i=1] \\ &\quad - \hat{E}(P_i^0 Y_i^1 + (1-P_i^0)Y_i^0|T_i=0] \\ &= \hat{E}[P_i^1(Y_i^1 - Y_i^0)|T_i=1] - \hat{E}[P_i^0(Y_i^1 - Y_i^0)|T_i=0] \\ &\quad + \hat{E}[Y_i^0|T_i=1] - \hat{E}[Y_i^0|T_i=0]\end{aligned}$$

En vertu de l'hypothèse (1) d'assignation aléatoire et de l'hypothèse (3) d'exclusion (*cf.* encadré 2), qui assurent l'indépendance des résultats et des participations au traitement ($(P_i^1, P_i^0, Y_i^1, Y_i^0) \perp T_i$), on a :

$$\begin{aligned}&\hat{E}[P_i^1(Y_i^1 - Y_i^0)|T_i=1] - \hat{E}[P_i^0(Y_i^1 - Y_i^0)|T_i=0] \\ &\quad = \hat{E}[(P_i^1 - P_i^0)(Y_i^1 - Y_i^0)]\end{aligned}$$

En outre, les mêmes hypothèses impliquent que :

$$\hat{E}[Y_i^0|T_i=1] - \hat{E}[Y_i^0|T_i=0] = 0$$

Par conséquent, le numérateur de $\hat{\delta}_{TOT}$ se résume à :

$$\begin{aligned}\hat{E}(Y_i|T_i=1) - \hat{E}(Y_i|T_i=0) &= \hat{E}[(P_i^1 - P_i^0)(Y_i^1 - Y_i^0)] \\ &= \hat{E}_{P_i^1 - P_i^0}\{\hat{E}[(P_i^1 - P_i^0)(Y_i^1 - Y_i^0)|(P_i^1 - P_i^0)]\}\end{aligned}$$

$(P_i^1 - P_i^0)$ prend les valeurs 1 pour le groupe des *compliers*, 0 pour les *never-takers* ou *always-takers* et -1 pour les *defiers*. On peut donc écrire :

$$\begin{aligned}&\hat{E}[(P_i^1 - P_i^0)(Y_i^1 - Y_i^0)] \\ &\quad = \hat{E}[-1 \times (Y_i^1 - Y_i^0)|(P_i^1 - P_i^0) = -1].\hat{\Pr}[(P_i^1 - P_i^0) = -1] \\ &\quad + \hat{E}[0 \times (Y_i^1 - Y_i^0)|(P_i^1 - P_i^0) = 0].\hat{\Pr}[(P_i^1 - P_i^0) = 0] \\ &\quad + \hat{E}[1 \times (Y_i^1 - Y_i^0)|(P_i^1 - P_i^0) = 1].\hat{\Pr}[(P_i^1 - P_i^0) = 1]\end{aligned}$$

L'hypothèse de monotonicité implique :

$$\hat{\Pr}[(P_i^1 - P_i^0) = -1] = 0$$

De même, de façon évidente :

$$\hat{E}[0 \times (Y_i^1 - Y_i^0)|(P_i^1 - P_i^0) = 0].\hat{\Pr}[(P_i^1 - P_i^0) = 0] = 0$$

Ceci correspond aux effets des *never-takers* et des *always-takers* qui s'annulent en vertu du tirage aléatoire.

Le numérateur de l'estimateur TOT se réduit donc à :

$$\begin{aligned}&\hat{E}(Y_i|T_i=1) - \hat{E}(Y_i|T_i=0) \\ &\quad = \hat{E}[(Y_i^1 - Y_i^0)|(P_i^1 - P_i^0) = 1].\hat{\Pr}[(P_i^1 - P_i^0) = 1]\end{aligned}$$

Quant au dénominateur de $\hat{\delta}_{TOT}$, il s'écrit de même :

$$\begin{aligned}\hat{E}(P_i|T_i=1) - \hat{E}(P_i|T_i=0) &= \hat{E}(P_i^1) - \hat{E}(P_i^0) = \hat{E}(P_i^1 - P_i^0) \\ &= \hat{E}[-1|(P_i^1 - P_i^0) = -1].\hat{\Pr}[(P_i^1 - P_i^0) = -1] \\ &\quad + \hat{E}[0|(P_i^1 - P_i^0) = 0].\hat{\Pr}[(P_i^1 - P_i^0) = 0] \\ &\quad + \hat{E}[1|(P_i^1 - P_i^0) = 1].\hat{\Pr}[(P_i^1 - P_i^0) = 1] \\ &= \hat{\Pr}[(P_i^1 - P_i^0) = 1]\end{aligned}$$

Au total, l'estimateur TOT est donc égal à :

$$\begin{aligned}\hat{\delta}_{TOT} &= \frac{\hat{E}(Y_i|T_i=1) - \hat{E}(Y_i|T_i=0)}{\hat{E}(P_i|T_i=1) - \hat{E}(P_i|T_i=0)} \\ &= \hat{E}[(Y_i^1 - Y_i^0)|(P_i^1 - P_i^0) = 1]\end{aligned}$$

Ainsi, l'estimateur TOT évalue l'effet moyen du traitement sur les individus qui sont sensibles à l'assignation aléatoire, *i.e.* ceux pour qui $(P_i^1 - P_i^0) = 1$, soit les *compliers*. Cet estimateur est aussi appelé *Local Average Treatment Effect* ou LATE (Angrist et Imbens, 1994).

a. Les démonstrations de cette annexe s'inspirent en grande partie de Duflo *et alii* (2006).



Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

# Annexe 2 : le calcul de puissance

Le calcul de puissance et de l'effet minimum détectable (EMD) est détaillé dans le cas d'un tirage au niveau individuel avec conformité parfaite. En fin d'annexe, on fournit à titre indicatif les formulations de l'EMD dans deux cadres un peu plus complexes : celui d'un tirage individuel avec conformité imparfaite, d'une part, et celui d'un tirage au niveau de grappes, d'autre part.

Soit le modèle (3) introduit dans l'encadré 1 :

$$Y_i = c + \delta T_i + \varepsilon_i$$

où $Y_i$ est l'indicateur de résultat pour l'individu $i$, $T_i$ son groupe d'assignation, $c$ une constante, $\delta$ le paramètre d'intérêt (l'effet de la mesure) à estimer et $\varepsilon_i$ le terme d'erreur. Nous supposons ici que les observations sont tirées indépendamment d'une population d'individus identiques avant traitement, de sorte que les $(\varepsilon_i)$ peuvent être considérés comme indépendants identiquement distribués (*i.i.d.*), de moyenne nulle et de variance commune, $\sigma^2$, inconnue.

L'estimateur des MCO du paramètre d'intérêt $\delta$ dans le modèle (3) s'exprime comme l'écart des moyennes observées dans chacun des deux groupes (c'est l'estimateur $\hat{\delta}$ de l'encadré 1) :

$$\hat{\delta} = \sum_{i/T_i=1} \frac{Y_i}{N_1} - \sum_{i/T_i=0} \frac{Y_i}{N_0}$$

où $N_1$ et $N_0$ désignent le nombre total d'observations, respectivement dans le groupe test (nombre d'individus $i$ tels que $T_i = 1$) et dans le groupe témoin (nombre d'individus tels que $T_i = 0$). Soient $N$ la taille de l'échantillon total ($N = N_1 + N_0$) et $\tau$ la proportion de la population éligible assignée au groupe test ($\tau = N_1 / N$). L'écart-type de $\hat{\delta}$ dans le modèle (3), conditionnel au traitement, est égal à :

$$\sigma_{\hat{\delta}} = \sqrt{\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_0}} = \sqrt{\frac{\sigma^2}{\tau N} + \frac{\sigma^2}{(1-\tau)N}} = \frac{\sigma}{\sqrt{\tau(1-\tau)N}}$$

En vertu de la loi des grands nombres, $\dfrac{\hat{\delta} - \delta}{\sigma_{\hat{\delta}}}$ converge en loi vers une loi normale centrée réduite. Soit $\Phi$ la fonction de répartition de cette loi. Dans toute la suite, on se place dans le cas d'un échantillon de taille suffisante pour que l'approximation normale soit valide.

Le calcul de puissance s'effectue le plus souvent à partir d'un test bilatéral asymptotique, de l'hypothèse nulle $H_0 : \delta = 0$ (la mesure n'a pas d'effet) contre l'hypothèse alternative $H_a : \delta \neq 0$ (la mesure a un effet).

Notons $\alpha$ le niveau du test, c'est-à-dire le risque d'erreur de type I (probabilité de rejet à tort de l'hypothèse nulle) maximal. Alors la région critique du test de $H_0$ contre $H_a$, c'est-à-dire la zone de rejet de l'hypothèse nulle, se définit comme :

$$W = \left\{ \left| \frac{\hat{\delta}}{\sigma_{\hat{\delta}}} \right| > t_{1-\alpha/2} \right\}$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite.

La puissance $\Pi$ du test bilatéral, qui se définit comme la probabilité de rejeter l'hypothèse nulle avec raison, est égale à :

$$\Pi = \Pr_{|\delta|>0}\left( \left| \frac{\hat{\delta}}{\sigma_{\hat{\delta}}} \right| > t_{1-\frac{\alpha}{2}} \right) = \Pr_{|\delta|>0}\left( \frac{\hat{\delta}-\delta}{\sigma_{\hat{\delta}}} > t_{1-\frac{\alpha}{2}} - \frac{\delta}{\sigma_{\hat{\delta}}} \right) + \Pr_{|\delta|>0}\left( \frac{\hat{\delta}-\delta}{\sigma_{\hat{\delta}}} < -t_{1-\frac{\alpha}{2}} - \frac{\delta}{\sigma_{\hat{\delta}}} \right)$$

$$= \Pr_{|\delta|>0}\left( \frac{\hat{\delta}-\delta}{\sigma_{\hat{\delta}}} < -t_{1-\frac{\alpha}{2}} + \frac{\delta}{\sigma_{\hat{\delta}}} \right) + \Pr_{|\delta|>0}\left( \frac{\hat{\delta}-\delta}{\sigma_{\hat{\delta}}} < -t_{1-\frac{\alpha}{2}} - \frac{\delta}{\sigma_{\hat{\delta}}} \right)$$

$$= \Phi\left( -t_{1-\frac{\alpha}{2}} + \frac{\delta}{\sigma_{\hat{\delta}}} \right) + \Phi\left( -t_{1-\frac{\alpha}{2}} - \frac{\delta}{\sigma_{\hat{\delta}}} \right)$$

Pour que la puissance du test soit suffisante (au moins égale à une valeur fixée $\pi$ jugée assez élevée, en pratique égale à 80 %), il faut donc que :

$$\Phi\left( -t_{1-\frac{\alpha}{2}} + \frac{\delta}{\sigma_{\hat{\delta}}} \right) + \Phi\left( -t_{1-\frac{\alpha}{2}} - \frac{\delta}{\sigma_{\hat{\delta}}} \right) \geq \pi$$

L'effet minimum détectable (EMD) se définit comme l'ampleur minimale de l'effet du traitement $|\delta|$ nécessaire à sa détection par un test de niveau de significativité $\alpha$ et de puissance $\pi$. Or, pour des plages de valeurs acceptables de $\alpha$ et $\pi$ et pour $\delta > 0$ (respectivement $\delta < 0$) vérifiant l'inégalité précédente, on montre[a] que le terme de droite (respectivement de gauche) de cette inégalité est négligeable devant l'autre terme. Il est alors facile de calculer l'EMD :

$$\Phi\left( -t_{1-\frac{\alpha}{2}} + \frac{\delta}{\sigma_{\hat{\delta}}} \right) \geq \pi \Leftrightarrow \delta \geq (t_\pi + t_{1-\alpha/2})\sigma_{\hat{\delta}} \quad (\text{pour } \delta > 0)$$

et :

$$\Phi\left( -t_{1-\frac{\alpha}{2}} - \frac{\delta}{\sigma_{\hat{\delta}}} \right) \geq \pi \Leftrightarrow \delta \leq -(t_\pi + t_{1-\alpha/2})\sigma_{\hat{\delta}} \quad (\text{pour } \delta < 0)$$

D'où :

$$(5)\ EMD = (t_\pi + t_{1-\alpha/2})\sigma_{\hat{\delta}} = (t_\pi + t_{1-\alpha/2})\frac{\sigma}{\sqrt{\tau(1-\tau)N}}$$

(a) NDLR : ce point est détaillé dans le commentaire d'Erkel-Rousse (2014), dans ce numéro.



Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

Par définition, dans le cas d'un test bilatéral, l'EMD vérifie :

$\Pi \geq \pi \Leftrightarrow |\delta| \geq EMD$

Plus l'EMD est faible à $\pi$ et $\alpha$ fixés, plus la puissance de l'évaluation est élevée, dans la mesure où l'expérimentation permet de détecter un effet du traitement même de petite ampleur. Au contraire, plus l'EMD est élevé à $\pi$ et $\alpha$ fixés, plus la puissance de l'évaluation est limitée (car il faut que le traitement ait un effet notable pour pouvoir être détecté).

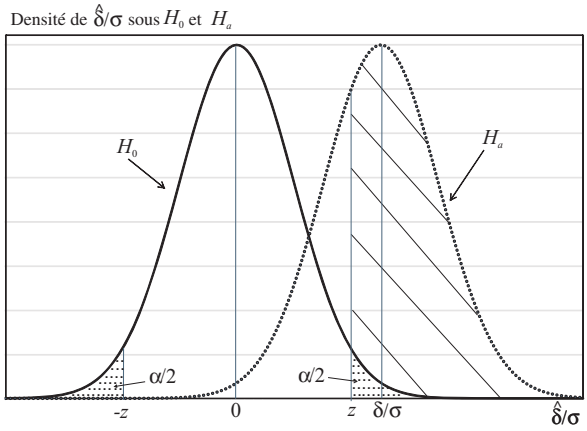
Pour $\alpha = 5$ % et $\pi = 80$ %, et dans le cas d'un échantillon suffisamment grand, on a : $t_{1-\alpha/2} \approx 1{,}96$ et $t_{\pi} \approx 0{,}84$. Au moment du calcul de puissance, qui s'effectue avant le tirage de l'échantillon et, *a fortiori*, avant l'assignation des individus échantillon dans le groupe test et le groupe témoin, $\sigma^2$ ne peut être estimé par les MCO dans le modèle (3), dont la variable dépendante et le régresseur ne sont pas encore observables. Deux options sont alors envisageables. Soit on dispose d'informations *a priori* sur la variance inconditionnelle de $Y$ et on estime $\sigma^2$ sur la base de ces informations[(b)]. Soit on ne dispose pas d'informations sur la variance de $Y$ : dans ce cas, on exprime le résultat du calcul de puissance en unités d'écart-type, ce qui évite d'avoir à estimer ce paramètre. Ceci revient à calculer :

$$EMD / \sigma = \frac{t_{\pi} + t_{1-\alpha/2}}{\sqrt{\tau(1-\tau)N}}$$

On dispose de tous les éléments permettant de calculer l'EMD. On peut alors déterminer les caractéristiques de l'échantillon permettant de réduire l'EMD (en points de pourcentage si on dispose d'une estimation de $\sigma^2$ ou en unités d'écart-type sinon) suffisamment pour que la puissance atteigne un niveau acceptable (au moins 80 %) – *cf.* corps du texte.

Le graphique A2.1 illustre le calcul de puissance pour un test bilatéral de niveau $\alpha$, dans le cas où l'effet $\delta$ est strictement positif et où l'on se situe dans la zone de rejet droite de l'hypothèse nulle : $\hat{\delta} / \sigma_{\hat{\delta}} > t_{1-\alpha/2}$ ou, de manière équivalente, dans ce cas où $\delta / \sigma > z$, avec $z = t_{1-\alpha/2} / \sqrt{\tau(1-\tau)N}$. Sont représentées les densités de la loi approximative (pour $N$ grand) de $\hat{\delta} / \sigma$ sous l'hypothèse nulle (courbe de gauche, centrée sur 0) et sous l'hypothèse alternative (courbe de droite, centrée sur $\delta / \sigma$). L'écart-type de ces deux densités est le même : $1 / \sqrt{\tau(1-\tau)N}$. La puissance est représentée par la surface hachurée délimitée par le niveau du seuil $z$ et la courbe de densité correspondant à l'hypothèse alternative $H_a$. Pour un quantile donné ($t_{1-\alpha/2} = 1{,}96$) = 1,96), le calcul de puissance consiste à modifier l'EMD afin d'atteindre un niveau de puissance acceptable. Ici, l'effet minimum détectable est à modifier jusqu'à ce que la partie hachurée représente 80 % de la surface sous la courbe de la densité sous $H_a$. Notons que, sous l'hypothèse alternative, $\Phi(-\delta/\sigma_{\hat{\delta}} - t_{1-\alpha/2})$ correspond à la surface sous la densité sous $H_a$ située à gauche du seuil $-z$ : elle peut être considérée comme infinitésimale au regard de la partie hachurée.

**Graphique A2.1 : illustration du calcul de puissance**




Source : auteurs.

Par convention, la recherche en sciences sociales privilégie les tests bilatéraux. Aussi les calculs de puissance sont-ils généralement établis sur la base d'un test bilatéral. Cependant, on pourrait envisager d'effectuer le calcul de puissance à partir d'un test unilatéral, de l'hypothèse nulle $H_0 : \delta = 0$ (la mesure n'a pas d'effet) contre l'hypothèse alternative $H_a : \delta > 0$ ou $\delta < 0$ (la mesure a l'impact attendu, correspondant à un effet positif ou négatif selon la manière dont est définie la variable de résultat). Comme souligné par Bloom (1995), l'utilisation d'un tel test unilatéral se justifierait dans de nombreux cas où le sens de l'effet de la politique est clairement établi (rôle du prix sur la demande d'un bien par exemple) ou s'il existe une raison physiologique ou naturelle qui détermine le sens de la politique (donnée anthropométrique ne pouvant pas prendre de valeur négative par exemple). Un argument usuellement avancé pour préférer le test bilatéral lors du calcul de puissance est qu'il est plus conservateur, au sens où, comme, à $\pi$ et $\alpha$ fixés, il aboutit à un EMD plus élevé, il conduit l'évaluateur à opter pour une taille d'échantillon supérieure afin d'assurer une puissance au moins égale à $\pi$. Il est donc souvent considéré comme préférable d'effectuer *ex ante* le calcul de puissance sur la base d'un test bilatéral, quitte à utiliser lors de l'analyse ultérieure un test unilatéral, statistiquement plus puissant mais moins conservateur. Cette question est discutée en détails dans le commentaire d'Erkel-Rousse (2014), dans ce numéro.

Le calcul de puissance précédent a été effectué dans le cas d'un tirage au niveau individuel avec conformité parfaite. Ci-dessous, on fournit les formulations de l'EMD dans deux cadres un peu plus complexes : celui d'un tirage individuel avec conformité imparfaite d'une part, et celui d'un tirage au niveau de grappes (toujours avec conformité imparfaite), d'autre part. Ceci illustre la dépendance de la formulation de l'EMD au niveau de conformité et au type de tirage réalisé.

(b) Sachant qu'en général, les politiques ont un impact très réduit sur les variables de résultat au regard des déterminants fondamentaux de ces dernières, de sorte qu'en pratique l'estimateur des MCO de $\sigma^2$ dans le modèle (3), s'il pouvait être calculé, serait assez proche de la variance inconditionnelle de $Y$.



Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

### *Tirage individuel avec conformité imparfaite*

En cas de conformité imparfaite, on montre que l'écart-type de $\hat{\delta}$ est amplifié d'une manière inversement proportionnelle au taux de conformité $p^1 - p^0$, de sorte que :

$$(6)\ EMD = (t_{\pi} + t_{1-\alpha/2}) \frac{\sigma}{\sqrt{\tau(1-\tau)N}} \frac{1}{(p^1 - p^0)}$$

où $p^1$ est le taux d'exposition (ou de participation) au traitement dans le groupe test et $p^0$ le taux d'exposition (ou de participation) au traitement dans le groupe témoin.

## *Tirage au niveau de grappes*

Dans le cas d'un tirage au niveau de grappes et non plus des individus, le modèle (3) doit être légèrement modifié pour tenir compte d'une structure un peu plus complexe des termes d'erreur, ceux-ci pouvant ne plus être indépendants les uns des autres pour des individus appartenant à une même grappe.

Le modèle (3) devient (Duflo *et alii*, 2006) :

$$Y_{ij} = c + \delta T_{ij} + \varepsilon_{ij}$$

avec $\varepsilon_{ij} = v_j + w_{ij}$

où $E(v_j) = E(w_{ij}) = 0$

$V(v_j) = \nu^2, V(w_{ij}) = \sigma^2$

$(v_j)$ et $(w_{ij})$ *i.i.d.* et non corrélés.

$\sigma^2$ est la variance inter-grappes du modèle. La corrélation intra-grappe $\rho$ est définie comme :

$$\rho = \frac{\mathrm{cov}(Y_{ij}, Y_{i'j})}{\sqrt{V(Y_{ij})\,V(Y_{i'j})}} \quad \forall\, i \neq i'$$

Elle est égale à :

$$\rho = \frac{V(v_j)}{V(Y_{ij})} = \frac{\nu^2}{\nu^2 + \sigma^2}$$

c'est-à-dire à la proportion de la variance totale du modèle due à la variance intra-grappe.

L'estimateur $\hat{\delta}$ des MCO de $\delta$ dans le modèle (3) demeure un estimateur sans biais de $\delta$ dans le modèle modifié mais l'expression de sa variance est légèrement différente. Elle dépend désormais de la variance inter-grappes $\sigma^2$ du modèle et de la corrélation intra-grappe $\rho$. Avec un nombre $G$ de grappes et $n$ d'individus par grappe, on montre que l'EMD s'exprime alors comme suit, dans le cas général d'une conformité imparfaite :

$$(7)\ EMD = (t_{\pi} + t_{1-\alpha/2})\sigma_{\hat{\delta}}$$

$$= (t_{\pi} + t_{1-\alpha/2}) \frac{\sigma}{\sqrt{G\tau(1-\tau)}} \sqrt{\rho + \frac{1-\rho}{n}} \frac{1}{(p^1 - p^0)}$$

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)