Essais Randomisés Contrôlés

Carlo Barone

SciencesPo
LABORATORY FOR INTERDISCIPLINARY EVALUATION OF PUBLIC POLICIES

LIEPP FICHE MÉTHODOLOGIQUE n°2

2023

Cette fiche fait partie d'un ensemble de fiches méthodologiques publiées par le LIEPP. A ce titre, elle bénéficie du soutien apporté par l'ANR et l'État au titre du programme d'Investissements d'avenir dans le cadre de l'IdEx Université Paris Cité (ANR-18-IDEX-0001).

Essais Randomisés Contrôlés

Carlo BARONE (Sciences Po, CRIS, LIEPP)

carlo.barone@sciencespo.fr

www.sciencespo.fr/liepp

Comment citer cette publication :
BARONE, Carlo, **Essais Randomisés Contrôlés**,
LIEPP Fiche méthodologique n°2, 2023-05

This publication was originally written in English:
BARONE, Carlo, **Randomised Controlled Trials**,
LIEPP Methods Brief n°1, 2023-05

EN BREF :

Les essais randomisés contrôlés (ERC) visent à mesurer l'impact d'une intervention donnée en comparant les résultats d'un groupe expérimental (recevant l'intervention) et d'un groupe de contrôle (ne la recevant pas), auxquels les individus sont assignés de manière aléatoire. Il s'agit d'une méthode quantitative utile d'évaluation *ex ante*, pour tester l'impact d'un programme à un stade où il n'a pas encore atteint la totalité de sa population cible (ce qui rend le groupe de contrôle possible).

Mots-clés : Méthode quantitative, méthode expérimentale, groupes expérimentaux/de traitement et de contrôle, affectation aléatoire, traitement, contamination

I. En quoi consiste cette méthode ?

Les essais randomisés contrôlés (ERC) évaluent l'impact d'une politique en comparant deux groupes : l'un d'eux se voit accorder l'accès à la politique (groupe expérimental), tandis que l'autre est temporairement exclu de la politique (groupe témoin ou contrôle). L'équipe de recherche traduit les objectifs de la politique en mesures quantitatives de résultats et évalue l'efficacité de la politique en comparant ces résultats dans les deux groupes. Si le groupe expérimental affiche des meilleurs résultats en moyenne sur ces dernières mesures, nous concluons que la politique est efficace. Toutefois, cette conclusion est valable si, et seulement si, nous pouvons supposer que les deux groupes étaient parfaitement équivalents. C'est pourquoi l'affectation aux deux groupes doit se faire de manière aléatoire : si l'échantillon est suffisamment grand, l'affectation aléatoire garantit que les deux groupes sont, en moyenne, initialement équivalents sur toutes les caractéristiques, connues ou non par l'équipe de recherche, mesurées ou non dans l'étude d'évaluation. Par conséquent, toute différence dans les résultats observés après la mise en œuvre de la politique peut être interprétée comme un impact causal de la politique.

Pour réaliser un ERC, on recrute un échantillon d'individus qu'on invite à participer à l'étude, en leur expliquant qu'ils et elles peuvent être affecté-e-s soit au groupe expérimental, soit au groupe témoin. Parmi les participant-e-s qui ont accepté de participer, la moitié sera affectée aléatoirement au groupe expérimental et l'autre moitié au groupe témoin. Ce ratio 50%-50% est le plus courant car il permet d'obtenir des estimations plus précises que les ratios non-équilibrés (par exemple 70%-30%). En amont de réaliser l'intervention, il est possible d'effectuer une première mesure des résultats (baseline) qui sont mesurés à l'issue de l'intervention chez les participants après, pour que cela serve de base de référence. Cette mesure n'est pas strictement nécessaire, mais elle est souvent effectuée pour plusieurs raisons, par exemple parce qu'elle permet d'étudier les impacts du traitement de manière plus dynamique en comparant les variations des résultats entre les deux groupes avant et après l'intervention.

Si la randomisation est une condition nécessaire pour pouvoir émettre des inférences causales plausibles concernant l'intervention, elle n'est pas une condition suffisante. En particulier, le groupe de contrôle doit rester exclu pendant toute la période de mise en œuvre de la politique, c'est-à-dire que nous devons éviter toute forme de "contamination" du traitement. Cela implique, par exemple, que les individus des deux groupes ne communiquent pas sur les objectifs et les contenus du traitement. De plus, lorsque les individus sont affecté-e-s au groupe de contrôle, ils et elles peuvent réagir en essayant de remplacer le traitement par un traitement similaire. La contamination et le remplacement du traitement peuvent invalider les inférences causales s'ils se produisent à grande échelle. Par conséquent, la condition essentielle est que le groupe de contrôle agisse "comme d'habitude" et il est important que l'équipe de recherche conçoive et présente l'étude de manière à garantir que ce soit le cas. Ainsi, si la randomisation est importante, il est tout aussi important de garantir le plus haut degré de contrôle de ces conditions

expérimentales. L'expression "essai randomisé contrôlé " décrit donc les deux conditions essentielles à la réalisation d'inférences causales solides : la répartition aléatoire et le contrôle des conditions expérimentales.

II. En quoi cette approche est-elle utile pour l'évaluation des politiques publiques ?

Les ERC visent à estimer les impacts causaux des politiques, c'est-à-dire à évaluer si les politiques produisent des changements correspondant aux objectifs de ces politiques. Le principal défi est que, même si une politique donnée est totalement inefficace, des changements peuvent intervenir en raison d'autres politiques, ou d'autres paramètres économiques ou socioculturels. Par exemple, on peut proposer un programme de formation à des chômeurs et chômeuses pour améliorer leur employabilité et observer ensuite les taux d'emploi des personnes participant à ce programme. Cependant, il n'est pas certain que le changement observé dans ce résultat puisse être attribué à la politique. Par exemple, il pourrait être dû au cycle économique ainsi qu'à tout autre type de politique économique, de travail ou de protection sociale (par exemple, des incitations fiscales à l'embauche, des modifications des règles d'éligibilité aux allocations de chômage, etc.). Par conséquent, une simple comparaison avant-après sans groupe témoin de comparaison ne permet pas d'isoler le véritable impact causal de cette politique.

Les ERC ne sont pas le seul type de méthode d'évaluation de l'impact causal, par exemple les modèles de régression avec discontinuité sont une autre option (voir fiche séparée). Les ERC sont une forme d'évaluation *ex ante*, c'est-à-dire qu'ils doivent être réalisés avant que la politique ne soit appliquée à l'ensemble de la population des bénéficiaires potentiels. En effet, les ERC supposent que la politique ne soit pas appliquée à certains individus, qui constituent le groupe de contrôle. Si la politique a déjà été généralisée, les ERC sont irréalisables. On peut alors recourir à d'autres types de méthodes d'évaluation de l'impact causal pour isoler le véritable impact causal de la politique.

III. Un exemple d'application : quels sont les messages qui favorisent le mieux la conformité fiscale ?

La conformité fiscale (*tax compliance*), c'est-à-dire la déclaration véridique des revenus imposables et le paiement des impôts en temps voulu, est essentielle pour financer les services publics. Une équipe de recherche s'est associée à l'administration fiscale belge pour tester l'impact de différents messages encourageant la conformité fiscale (De Neve et al, 2019). Entre 2014 et 2016, l'équipe a assigné de manière aléatoire environ 2,5 millions de contribuables à recevoir différents messages : des messages simplifiés présentant les informations clés en termes plus simples, des messages de dissuasion visant à rendre explicites les conséquences de la non-conformité, et des messages de morale fiscale visant à motiver les contribuables à apprécier l'importance de la conformité pour la fourniture de biens publics. Les 4 millions de contribuables restant-e-s ont été affecté-e-s à un groupe témoin où la communication avec les contribuables est restée inchangée (cette taille d'échantillon est exceptionnelle, la plupart des ERC se basant sur quelques centaines ou milliers de cas). À l'aide de données administratives, l'équipe de recherche a mesuré l'impact de l'intervention sur la probabilité d'effectuer un paiement ou de déclarer ses impôts, ainsi que sur le montant des revenus déclarés. Une communication plus simple a eu l'effet le plus important sur le respect des obligations fiscales, incitant les gens à déclarer et à payer leurs impôts plus tôt. L'ajout de messages de dissuasion a encore amélioré le respect des obligations fiscales, tandis que les messages moralisateurs se sont révélés inefficaces.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode ?

Dans certains contextes, les ERC sont irréalisables car les risques de contamination ou de remplacement du traitement sont trop élevés. Par exemple, lorsque les individus peuvent facilement communiquer sur le contenu informatif d'une intervention et sont très motivé-e-s pour le faire. Certaines politiques ne peuvent pas être testées avec un ERC car, par construction, elles impliquent l'ensemble de la population, nous ne pouvons donc pas exclure temporairement le groupe de contrôle. C'est par exemple le cas de plusieurs politiques macroéconomiques ou de défense (par exemple, un changement dans les dépenses militaires).

De plus, alors qu'on assigne le plus souvent des individus au groupe de traitement ou de contrôle, on peut aussi assigner des familles entières, des rues ou des villages au traitement ou au contrôle. C'est le cas, par exemple, lorsqu'une intervention donnée est plus efficace, ou ne peut être mise en œuvre, qu'à ces niveaux supra-individuels. Ces types de randomisations de niveau supérieur (randomisation en grappes) peuvent être nécessaires ou extrêmement pratiques, mais ils exigent des échantillons de grande taille et donc des budgets importants.

Enfin, il faut garder à l'esprit que la validité interne (c'est-à-dire la force des inférences causales dans le cas étudié) n'est qu'un des critères de qualité de la recherche en évaluation. Un autre critère important est la validité externe, c'est-à-dire la possibilité de généraliser des conclusions au-delà de l'échantillon étudié. Ce deuxième critère, lorsqu'il est appliqué aux ERC, exige des échantillons importants et aléatoires de la population étudiée, et que le nombre de participant-e-s abandonnant l'étude reste limité. Un troisième critère important concerne la validité et la fiabilité des mesures des résultats, y compris la capacité d'observer les résultats à long terme d'une politique, et la couverture de tous les effets potentiels (positifs et négatifs) de la politique.

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres ?

Comme expliqué ci-dessus, la principale force des ERC est qu'ils permettent d'évaluer l'impact causal réel d'une politique avant de l'appliquer à l'ensemble de la population des bénéficiaires. Dans la recherche clinique, les ERC sont la méthode standard pour évaluer l'efficacité de tout type de thérapie ou de médicament et ils sont de plus en plus utilisés pour l'évaluation des politiques publiques, notamment dans les domaines de l'éducation, du marché du travail, de la santé et du logement.

Les applications les plus courantes de cette méthode impliquent la randomisation entre deux groupes d'individus. Cependant, il arrive que l'on puisse organiser trois groupes d'individus ou plus afin de comparer des variantes qualitativement différentes d'une intervention ou des dosages différents de l'intervention. Par exemple, dans une étude visant à promouvoir l'utilisation des services de vélo en libre-service, on peut comparer le groupe de contrôle à un premier groupe de traitement disposant d'informations sur le vélo en libre-service, à un deuxième groupe de traitement recevant une incitation financière à son utilisation et à un troisième groupe recevant une incitation financière plus importante.

Les ERC ne sont pas toujours facilement réalisables. En particulier, les responsables politiques ou les participant-e-s potentiel-le-s peuvent refuser le principe de la randomisation, par exemple parce qu'ils pensent que cela pose un problème éthique car elles excluent les individus du groupe de contrôle des avantages de la politique. Cette critique oublie toutefois que l'exclusion est temporaire, c'est-à-dire qu'elle ne dure que le temps nécessaire pour démontrer que la politique est efficace. Cette exclusion temporaire permet d'évaluer si la politique est efficace avant de la généraliser à l'ensemble de la population. En outre, les ressources disponibles dans les études d'évaluation *ex ante* ne permettent de traiter qu'une petite

partie de la population totale, de sorte qu'il serait de toute façon impossible de traiter tout le monde : l'assignation aléatoire donne à chacun les mêmes chances d'être traité.

Il est primordial que les équipes de recherche expliquent en termes simples pourquoi la randomisation est éthique et pourquoi elle est nécessaire pour garantir la fiabilité des comparaisons entre les deux groupes. Lorsque cela est possible, l'acceptabilité sociale de la randomisation peut être accrue en créant une liste d'attente, c'est-à-dire que le groupe témoin reçoit la politique à la fin de l'étude, ou un traitement compensatoire (un traitement différent de celui étudié et qui n'affecte pas le résultat de l'étude). Par exemple, dans une étude fournissant des informations sur les services de garde d'enfants aux mères enceintes afin d'améliorer le recours à ces services, le groupe de contrôle peut recevoir cette information à la fin de l'étude ou recevoir un autre type d'information, par exemple sur les pratiques saines pendant la grossesse. Toutefois, si une liste d'attente est créée, il n'est pas possible d'observer les résultats à long terme car le groupe de contrôle n'est plus exclu de l'intervention. Les listes d'attente et les traitements compensatoires peuvent également être utilisés pour réduire le risque que les personnes assignées au groupe de contrôle abandonnent le traitement. Il est en effet important que les taux d'abandon des deux groupes soient similaires afin de préserver leur équivalence tout au long de l'étude.

Par rapport aux expériences en laboratoire, les ERC ont une validité écologique plus élevée, dès lors que l'on étudie des personnes dans des situations de vie réelles et dans des contextes naturels. Par conséquent, le risque que leur comportement soit influencé par la conscience de faire partie d'une étude est moins important. En même temps, par rapport aux expériences en laboratoire, les ERC permettent un degré de contrôle plus faible sur le comportement des participant-e-s. Dans les expériences cliniques et psychologiques, la conscience d'être traité-e est souvent neutralisée par l'administration de placebos au groupe de contrôle, c'est-à-dire des traitements spécifiquement conçus pour n'avoir aucun effet. Dans les politiques sociales, cette pratique est moins courante car nous avons tendance à considérer les avantages découlant de la conscience d'être traité comme faisant partie intégrante de la politique.

Plus fondamentalement, si les ERC sont un outil fiable pour évaluer les impacts causaux des politiques, ils ne sont pas en position de force pour étudier les processus sous-jacents. Par exemple, si un ERC conclut qu'une politique est inefficace ou moins efficace que prévu, cette méthode est incapable d'expliquer ce qui n'a pas fonctionné et comment nous pouvons améliorer cette politique. C'est pourquoi il est extrêmement utile d'intégrer les ERC aux techniques qualitatives d'évaluation des processus. De plus, les croyances et les perceptions de la politique qu'ont les bénéficiaires et les agent-e-s de la mise en œuvre peuvent être étudiées en utilisant des entretiens qualitatifs ou des enquêtes.

Quelques références bibliographiques pour aller plus loin

De Neve, Jan-Emmanuel. et Imbert, Clement. et Spinnewijn, Johannes. et Tsankova, Teodora. et Luts, Maarten. 2019. "How to Improve Tax Compliance? Evidence from Population-wide Experiments in Belgium." *Working paper*.

White, Howard. et Sabarwal, Shagun. et De Hoop, Thomas. 2014. *Essais Contrôlés Randomisés*, Notes méthodologiques, Évaluation d'impact no 7, Unicef.
https://www.unicef-irc.org/publications/pdf/MB7FR.pdf

Gibson, Michael. et Sautmann, Anja. Dernière mise à jour : avril 2021. *Introduction to randomized evaluations*, Abdul Latif Jameel Poverty Action Lab.
https://www.povertyactionlab.org/resource/introduction-randomized-evaluations

Gertler, Paul. et Martinez, Sebastian. et Premand, Patrick. et Rawlings, Laura. et Vermeersch, Christel. 2016. *Impact Evaluation in Practice*, deuxième édition. World Bank Group
https://openknowledge.worldbank.org/bitstream/handle/10986/25030/9781464807794.pdf?sequence=2&isAllowed=y (Chapitres 3 et 4 de ce manuel)