impact canada

MODÈLES DE MESURE DES IMPACTS

Guide des méthodes de mesure des impacts

2019

Canada

MODÈLES DE MESURE DES IMPACTS

GUIDE DES MÉTHODES DE MESURE DES IMPACTS

This publication is also available in English:
Measuring impact by design - a guide to methods for impact measurement.

CP22-174/2019F-PDF
ISBN: 978-0-660-29541-1

MESSAGE DU BUREAU DU CONSEIL PRIVÉ

Nous allons l'honneur de dévoiler la première série de lignes directrices pour la mesure des impacts de l'Unité des impacts et de l'innovation (UII), à l'appui de ses travaux dans le cadre de l'Initiative Impact Canada. Ce document se veut à la fois une introduction accessible à ce sujet ainsi qu'une référence pour les personnes intervenant dans la conception, la mise en œuvre, l'approvisionnement et les stratégies de mesure des impacts pour les projets d'Impact Canada. Fondé sur des pratiques exemplaires, le document *Modèles de mesure des impacts* a été rédigé pour amener les lecteurs à repenser la stratégie de mesure des impacts traditionnellement utilisée au sein de la fonction publique fédérale.

Dans son rôle de direction de l'Impact Canada – une initiative pangouvernementale – l'UIII travaille avec un réseau en constante expansion de partenaires à l'élaboration d'un programme innovant et axé sur les résultats. Nous sommes conscients que les dépenses associées aux programmes sont des investissements que nous faisons au nom des Canadiens, et directement dans leur intérêt, et que nous devons faire des efforts pour mieux comprendre en quoi ces investissements améliorent la vie des citoyens. Cela signifie que nous devons mieux comprendre ce qui fonctionne, pour qui et dans quel contexte. Nous devons aussi mieux comprendre quels types d'investissements sont susceptibles de maximiser les impacts social, économique et environnemental que nous cherchons.

« Nous sommes conscients que les dépenses associées aux programmes sont des investissements que nous faisons au nom des Canadiens, et directement dans leur intérêt, et que nous devons faire des efforts pour mieux comprendre en quoi ces investissements améliorent la vie des citoyens.»

De bonnes pratiques de mesure des impacts sont fondamentales pour comprendre ces aspects, et nous avons le devoir de faire preuve de rigueur dans nos efforts. Nous reconnaissons que nous devons continuer d'accroître la capacité du gouvernement à miser efficacement sur ces approches. C'est pour cette raison que nous avons assoupli les autorisations relatives à Impact Canada; ainsi les subventions et les contributions peuvent être utilisées pour financer les organismes de recherche qui détiennent une expertise dans le domaine des techniques énoncées dans ce guide. Nous encourageons nos ministères partenaires à envisager de tirer avantage de cette souplesse accrue.

Le guide *Modèles de la mesure des impacts* est un des nombreux modes de soutien que l'UII fournit pour respecter son engagement d'améliorer les pratiques de mesure des impacts pour Impact Canada. Nous sommes très enthousiastes à l'idée de poursuivre la collaboration avec nos partenaires dans la mise en œuvre de ces importantes approches axées sur les résultats à l'échelle du secteur public.

Matthew Mendelsohn
Sous-sécretaire du Cabinet,
Résultats et livraison

Rodney Ghali
Secrétaire adjoint du Cabinet,
Unité de l'impact et de l'innovation

TABLE DES MATIÈRES

INTRODUCTION

Améliorer les résultats pour les Canadiens, c'est la raison d'être d'Impact Canada, une initiative pangouvernementale lancée dans le cadre du budget de 2017 dirigée par l'Unité des impacts et de l'innovation (UII) du Bureau du Conseil privé. L'Initiative Impact Canada accélère l'utilisation d'approches novatrices et expérimentales à l'échelle du gouvernement du Canada. Afin d'atteindre cet objectif central, l'Initiative Impact Canada adopte une approche axée sur les résultats à tous les niveaux. L'importance accordée aux résultats est ce qui relie chacune des fonctions opérationnelles de base de l'Unité des impacts et de l'innovation (UUI). Ces fonctions comprennent la mise en œuvre de nouveaux outils comme l'introspection comportementale dans la conception des programmes, l'utilisation de mesures financières novatrices pour améliorer la mise en œuvre et l'application de méthodes de pointe pour l'analyse des impacts afin d'évaluer la mesure dans laquelle des résultats ont réellement été obtenus.

Dans le cadre de son mandat de co-conception d'approches novatrices et expérimentales avec les ministères du gouvernement du Canada, l'UII travaille avec ces derniers à la mise en place d'une stratégie d'évaluation rigoureuse pour toutes les initiatives lancées dans le cadre de d'Impact Canada par l'intermédiaire d'un processus de planification des projets. À cette fin, ce guide se concentre sur l'utilisation de techniques pour la mesure rigoureuse des impacts qui pourraient s'appliquer à de nombreuses circonstances dans le contexte de d'Impact Canada. Le guide *Modèles de mesure des impacts* est susceptible d'être particulièrement pertinent pour le programme et les efforts de prestation de services qui permettront d'adopter une approche « contrefactuelle » et de comprendre ce qui a été accompli, au-delà de ce qui se serait produit d'une façon ou d'une autre. Dans cette optique, les méthodes décrites dans ce guide peuvent servir à mieux déterminer les impacts causales pouvant être directement attribuées à l'initiative en question. De telles approches, lorsqu'elles sont adéquatement mises en œuvre, peuvent apporter des preuves plus solides et une plus grande assurance que les résultats observés sont réellement attribuables à l'initiative elle-même plutôt qu'à d'autres facteurs.

Les expériences ou les essais contrôlés randomisés (ECR) ne datent pas d'hier. Les ECR ont commencé à être utilisés au milieu des années 1920 dans le secteur des sciences agricoles, et sont depuis mis à profit par une grande variété d'évaluateurs et de professionnels des sciences sociales dans de nombreux domaines (Jamison, May 2017, p. 15). Ils ont connu une hausse de popularité dans la dernière décennie à mesure que leur utilisation, plus particulièrement dans le domaine du développement international, s'est répandue.

« Un des principaux thèmes de ce document suggère que, moyennant une planification adéquate, la majorité des programmes peuvent miser sur une méthode expérimentale ou quasi expérimentale pour la mesure des impacts qui exige peu ou pas d'ajustements au mode normal de fonctionnement (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2016, pp. 188-193). »

Les quasi-expériences ont en règle générale des origines plus récentes. Néanmoins, nous pouvons constater qu'elles sont maintenant de plus en plus couramment utilisées en tant que modèles de mesure des impacts. Grâce à ce guide, l'UII espère accélérer leur adoption dans les contextes canadiens. En tant que catégorie de méthodes, les quasi-expériences offrent un équilibre en-

tre rigueur et pragmatisme et, par conséquent, conviennent bien à l'exercice de mesure des impacts. Elles sont assez rigoureuses pour nous fournir des estimations plausibles des impacts, et assez pragmatiques pour être déployées dans des contextes où les méthodes randomisées ne s'appliquent pas. Un des principaux thèmes de ce document suggère que, moyennant une planification adéquate, la majorité des programmes peuvent miser sur une méthode expérimentale ou quasi expérimentale pour la mesure des impacts qui exige peu ou pas d'ajustements au mode normal de fonctionnement (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2016, pp. 188-193).

PRINCIPAUX OBJECTIFS

Ce document n'est pas un guide « étape par étape » sur l'application des évaluations expérimentales ou quasi expérimentales. Il se veut plutôt une ressource clé pour aider le personnel chargé des politiques et des programmes à mieux comprendre les principales approches, leurs avantages et leurs désavantages, ainsi que les conditions dans lesquelles elles peuvent être utilisées. Il n'est pas nécessaire que les utilisateurs de ce document possèdent une formation approfondie en méthodes de recherche ou en statistique.

Ce document s'adresse principalement au personnel de première ligne qui supervise l'élaboration et la mise en œuvre des initiatives d'Impact Canada. Ces employés sont bien placés pour planifier proactivement la mesure des impacts lors de la conception d'un programme. Une planification adéquate à ce stade précoce est un facteur essentiel pour assurer une mesure efficace des impacts. Sans prévisions de ce type, on peut facilement rater d'importantes occasions de saisir des données de base ou de créer des groupes comparables dès les premières étapes d'un programme. Ce genre d'erreur involontaire est très courant et n'est pas facile à gérer. L'UII a établi un processus pour collaborer avec les ministères et les soutenir dans l'élaboration d'une stratégie d'évaluation efficace en tant qu'élément central de la conception du programme. Chaque approche est adaptée aux besoins du ministère ainsi qu'au contexte et aux résultats visés pour chaque initiative.

Le guide est structuré de manière à fournir un aperçu de certains concepts clés liés à la mesure des impacts, de même qu'un aperçu non technique de la principale série de méthodes expérimentales et quasi expérimentales utilisées pour mesurer les impacts. Il vise à en décrire la logique de manière accessible, et il illustre les concepts clés à l'aide d'exemples, lorsque c'est possible.

« Le guide est structuré de manière à fournir un aperçu de certains concepts clés liés à la mesure des impacts, de même qu'un aperçu non technique de la principale série de méthodes expérimentales et quasi expérimentales utilisées pour mesurer les impacts. »

Voici les principaux objectifs de ce document :

- Approfondir les connaissances des spécialistes des prgrammes et des politiques à l'égard des concepts clés liés à la mesure des impacts;
- Mieux faire connaître les différentes méthodes expérimentales et quasi expérimentales applicables à la mesure des impacts;
- Permettre au personnel de première ligne chargé des programmes et des politiques de promouvoir l'amélioration des évaluations des impacts, plus particulièrement pour les initiatives de financement en fonction des résultats dans le contexte de l'Initiative Impact Canada;
- Offrir une ressource clé au personnel des programmes et des politiques pour lui permettre de se montrer critique à l'égard des stratégies et des rapports de mesure des impacts;
- Faire comprendre quels types de méthodes de mesure des impacts conviennent le mieux au contexte particulier de chaque programme.

L'INITIATIVE IMPACT CANADA

Annoncée dans le budget de 2017, Impact Canada est un effort pangouvernemental qui aidera les ministères à accélérer l'adoption de méthodes de financement fondées sur les résultats afin de produire des résultats significatifs pour les Canadiens. Les méthodes fondées sur les résultats constituent une nouvelle manière de gérer le financement des subventions et des contributions : il est question de passer de l'approche traditionnelle axée sur le processus et les produits à une approche selon laquelle les paiements sont fonction de l'obtention de résultats mesurables sur les plans économique, environnemental et/ou social.

Impact Canada préconise l'utilisation de différentes méthodes de financement novatrices, notamment les suivantes :

- Défis – L'attribution de prix à celui qui trouve le premier ou le plus efficacement une solution à un problème défini, ou le recours à des concours ouverts et structurés pour solliciter des propositions afin de financer les meilleures idées susceptibles de résoudre des problèmes thématiques.

- Financement en fonction des résultats – Utilisation d'instruments personnalisés afin de privilégier la méthode selon laquelle on verse des fonds aux bénéficiaires de financement en fonction des résultats sociaux positifs et mesurables obtenus (p. ex. obligations à impact social, mécanismes de paiement à la réussite).

- L'application de l'introspection comportementale et d'autres approches fondées sur des preuves pour améliorer la prestation des programmes et des services.

Impact Canada est appuyé par un centre d'expertise situé au sein de UII. L'équipe compte une vaste expérience de l'exécution de programmes nouveaux et novateurs au sein du gouvernement, dans les domaines suivants :

- **Approches novatrices du financement et du partenariat –** le personnel s'occupe directement de mettre en œuvre des méthodes de financement novatrices, notamment les obligations à impact social et les investissements à impacts sociales, et de lancer des défis d'externalisation ouverte à grande échelle pour trouver des solutions à des problèmes pressants. Ces expériences nécessitent toutes l'adoption d'approches multisectorielles qui ont permis de rassembler des intervenants venant du gouvernement, du secteur privé et des secteurs philanthropiques et à but non lucratif pour réaliser des résultats communs.

- **Mesure des impacts –** l'équipe a travaillé avec des partenaires pour concevoir et élaborer conjointement des méthodes fondées sur des données probantes afin d'améliorer les résultats des programmes. Le personnel tire parti d'un ensemble de méthodes d'évaluation et de mesure des impacts, et travaille avec des partenaires à faire avancer de nouvelles méthodes de mesure des impacts.

- **Introspection comportementale –** l'équipe possède une grande expérience du soutien à la réalisation et à la prestation d'expériences et de projets intégrant les méthodes d'introspection comportementale. Le travail de l'équipe comprend, entre autres, la mise en application de principes et d'approches fondés sur des données probantes, la réalisation d'expériences de petite à grande envergure, et l'application stratégique de la science comportementale à l'élaboration de politiques directement à l'appui du mandat de base et des engagements du gouvernement du Canada.

POURQUOI MESURER LES IMPACTS?

POURQUOI MESURER LES IMPACTS?

Fondamentalement, la mesure des impacts est une façon impartiale d'orienter les décisions au moyen d'une méthodologie scientifique rigoureuse. La principale raison pour laquelle nous mesurons les impacts d'un programme est que cela nous permet de déterminer si le programme en question a livré ou non les résultats attendus. En outre, lorsqu'elles sont bien conçues, les évaluations des impacts peuvent nous aider à répondre à certaines questions :

- Quels types de programmes sont les plus susceptibles d'apporter de grands avantages à la société?
- Quels types de programmes devrions-nous abandonner ou éviter?
- Comment les avantages d'un programme sont-ils répartis entre les différents groupes sociaux?
- Y a-t-il des volets du programme qui apportent plus d'avantages que d'autres?
- Comment un programme génère-t-il des impacts?
- Comment pouvons-nous améliorer un programme pour accroître ses impacts?
- Quels types de programmes devrions-nous élargir ou mettre à l'échelle?

Dans le contexte plus large de l'innovation dans le secteur public, la pertinence de la mesure des impacts ne fait qu'augmenter :

- La mesure des impacts s'inscrit dans la reddition de comptes, un principe de base pour un bon gouvernement. Nous devons faire preuve de rigueur dans nos efforts pour montrer comment les dépenses publiques sont liées à l'atteinte de meilleurs résultats pour les Canadiens. L'accent accru que nous mettons sur les résultats et livraison au sein du secteur public en est une preuve évidente.

- Les administrateurs généraux sont maintenant tenus d'affecter un pourcentage des dépenses associées aux programmes à l'expérimentation. La mesure des impacts est une excellente façon de nous aider à respecter cet engagement.

- Les modèles de financement en fonction des résultats, comme les obligations à impact social, sont de nouveaux moyens qu'envisagent les ministères pour améliorer les résultats pour les Canadiens. Ces approches sont des éléments fondamentaux d'Impact Canada. Si nous voulons accorder du financement en fonction des résultats, nous devons pouvoir déterminer clairement quand nous avons obtenu les résultats visés et quand nous n'y sommes pas parvenus. C'est l'objectif central de la mesure des impacts. Il est essentiel d'utiliser une méthode rigoureuse pour évaluer les impacts d'une initiative pour donner aux investisseurs initiaux comme aux investisseurs qui accordent du financement en fonction des résultats confiance que les résultats du programme sont mesurés de façon précise et équitable.

- On observe une demande croissante envers les formes économiques d'évaluation, notamment le rendement social des investissements et l'analyse coûts-avantages. La mesure des impacts est une étape centrale et essentielle de ces approches. Pour évaluer le rendement de l'investissement, nous devons d'abord acquérir une compréhension approfondie des impacts de nos programmes.

- Les décideurs cherchent de plus en plus à innover parce qu'ils veulent découvrir ce qui « fonctionne » lorsqu'ils abordent des problèmes sociaux de longue date ou complexes. Au fil du temps, l'accumulation de plusieurs évaluations des impacts dans des domaines précis pourrait appuyer cet objectif.

Dans l'ensemble, nous travaillons dans un contexte de plus en plus axé sur l'innovation en tant que moyen d'améliorer les résultats pour les Canadiens. Pour cette raison, la mesure des impacts doit être intégrée dans notre façon de travailler.

DÉFINITION DE LA MESURE DES IMPACTS

Étant donné que l'atteinte de résultats pour les Canadiens est un élément central d'Impact Canada, comprendre dans quelle mesure les programmes changent la vie des Canadiens revêt une importance majeure. Générer des impacts a une signification précise. Une ***retombée*** s'entend de tout changement dans les résultats ***généré*** par des investissements dans un programme ou des politiques.

En pratique, les termes « résultats » et « impacts » sont souvent employés de manière interchangeable. Bien qu'ils soient interreliés, ces deux concepts ont chacun une signification qui leur est propre. Un résultat se définit comme tout avantage sur le plan social, environnemental ou économique qu'une politique ou un programme vise à maintenir ou à améliorer d'une quelconque manière. La participation au marché du travail est un bon exemple. La mesure des résultats répond à des questions principalement descriptives, comme : Quel est le taux de participation actuel au marché du travail? En quoi la participation au marché du travail varie-t-elle d'une région à l'autre? A-t-elle changé au fil du temps?

Par contraste, une retombée est la mesure dans laquelle un programme produit un changement dans le cadre d'un résultat. Les impacts sont définies comme des changements (positifs ou négatifs) dans les résultats sur le plan social, environnemental ou économique qui sont ***directement attribuables à un investissement***. Pour faire fond sur l'exemple de la participation au marché du travail précédemment cité, les impacts pourraient être évaluées à l'aide de la question suivante : dans quelle mesure un programme de formation professionnelle a-t-il ***changé*** le taux de participation au marché du travail pour les citoyens formés? Dans le cas présent, une règle de base serait de voir les ***résultats*** comme un nom, et les ***impacts*** comme un verbe.

«Les impacts sont définies comme des changements (positifs ou négatifs) dans les résultats sur le plan social, environnemental ou économique qui sont directement attribuables à un investissement.»

Il est important de faire une distinction entre résultats et impacts puisque, bien souvent, nous devons nous intéresser à des éléments qui vont au-delà des résultats qui nous intéressent. Souvent, nous voulons déterminer dans quelle mesure nos actions (politiques, programmes ou interventions) mènent à des impacts sur les résultats (des changements aux résultats). Voilà le rôle de la mesure des impacts. Par conséquent, la mesure des impacts va plus loin que la mesure des résultats, et permet d'établir à quel point une politique ou un programme influe sur les résultats d'intérêt. Autrement dit, c'est une façon d'isoler la mesure dans laquelle un programme change les résultats des effets de tout autre élément susceptible d'avoir également changé les résultats.

À partir de cette définition, les sections suivantes fourniront un aperçu de certains concepts clés liés aux impacts; nous passerons ensuite à une description des principales méthodes de mesure des impacts que nous devrions connaître.

PROBLÈMES ASSOCIÉS À LA MESURE DES IMPACTS

Lorsqu'il s'agit de mesurer les impacts d'un programme, il devient rapidement clair que nous devons surmonter un problème : il n'est pas toujours évident que les résultats que nous observons sont bel et bien le produit des interventions qui nous intéressent.

Imaginons qu'un groupe d'individus participent à un programme visant à accroître leur niveau d'activité physique. Après trois mois, nous observons qu'ils sont (en moyenne) plus actifs qu'ils l'étaient avant de participer au programme. Le programme a-t-il été efficace? Dans ce scénario, il est impossible de dire si c'est le programme lui-même qui a amené un changement dans les résultats ou si ce changement était en tout ou en partie attribuable à d'autres facteurs.

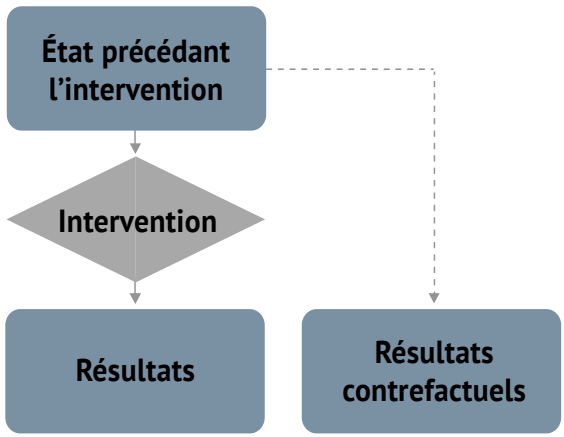
- Supposons que le programme a commencé en avril et s'est terminé en juin. Il est possible que les participants soient simplement devenus plus actifs avec le retour du beau temps.

- Supposons qu'une autre organisation a offert un programme différent à ce même groupe d'individus. Il est également possible que ce soit ce programme qui ait amené les impacts voulues, et que notre programme n'en ait amené aucune – ou ait même eu des impacts négatives.

- Supposons qu'une étude majeure sur les bienfaits d'un style de vie sain ait été publiée au même moment. Nous ne pouvons pas nécessairement isoler les effets du programme des effets de cette nouvelle information que nos participants pourraient avoir reçue.

«il n'est pas toujours évident que les résultats que nous observons sont bel et bien le produit des interventions qui nous intéressent.»

Figure 1



Cet exemple illustre très bien pourquoi il est essentiel de faire une distinction entre les résultats et les impacts. Les résultats changent souvent, et souvent pour des raisons que nous ne connaissons pas nécessairement. Si nous voulons comprendre les impacts d'un programme en particulier, mesurer des résultats changeants ne suffit pas. Cette stratégie ne nous permettra pas d'attribuer ce changement à une cause unique, comme un programme. Nous devons trouver une façon de comparer ce qui s'est produit avec ce qui se serait produit. C'est ce que nous appelons des scénarios ***factuels*** et ***contrefactuels***. Le scénario factuel est ce qui s'est réellement produit (participation au programme). Le scénario contrefactuel est ce qui se serait produit en l'absence du programme. Par définition, le scénario contrefactuel ne peut jamais être observé parce qu'il représente ce qui *ne s'est pas* produit. Donc, la difficulté qui réside dans la mesure des impacts est de trouver une façon de reconstru-

ire ce qui *se serait produit* en l'absence d'un programme, ce qui nous permet de comparer ces deux scénarios et d'évaluer les véritables impacts.

«Le scénario factuel est ce qui s'est réellement produit (participation au programme). Le scénario contrefactuel est ce qui se serait produit en l'absence du programme.»

Mesurer les *impacts* est ce qui nous permet d'isoler les effets d'un programme sur un résultat (ou des résultats) d'intérêt. C'est comme si on se demandait : dans quelle le mesure le changement observé dans les résultats est-il attribuable au programme lui-même? Comme nous le verrons plus loin dans ce document, la meilleure approche à adopter consiste à comparer deux groupes similaires en tous points, sauf qu'un groupe participe au programme, l'autre non. Nous appellerons le premier ***groupe de traitement*** ou ***groupe test***, et le second, ***groupe témoin*** ou ***groupe de comparaison***. Étant donné que la seule chose qui distingue ces deux groupes est leur participation à un programme, toute différence observée dans les résultats obtenus par chacun d'eux peut être attribuée à cette intervention, et seulement à celle-ci. C'est ce qui nous démontre que le programme a des impacts. Les sections subséquentes de ce document fourniront une introduction aux principales méthodes utilisées pour mesurer les effets d'un programme ou d'une politique afin d'en déterminer les véritables impacts.

LE PRINCIPE *CETERIS PARIBUS*

En latin, ceterus paribus signigie «toutes choses demeurant égales».

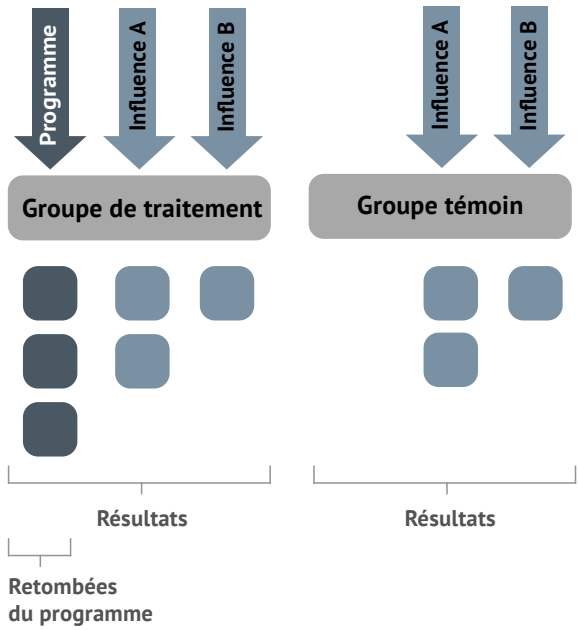
Pour pouvoir comprendre les résultats contrefactuels, il faut comparer les résultats d'un groupe de traitement avec ceux d'un groupe témoin. Pour ce faire, nous devons nous assurer que le groupe témoin respecte ce qu'on appelle le principe ***ceteris paribus***. En latin, *ceteris paribus* signifie « toutes choses demeurant égales ». En répartissant des individus de façon aléatoire dans deux groupes, ou en créant des groupes correspondants où nous nous assurons que les participants présentent les mêmes variables clés, nous minimisons la probabilité que des différences systématiques existent entre les deux groupes, ce qui nous permet de tenir pour acquis que tout élément pouvant affecter les résultats a la même influence sur les deux groupes. Ainsi, lorsqu'un de ces deux groupes ne participe pas à l'intervention, les résultats obtenus par ce groupe représentent les résultats qu'aurait obtenus le groupe de traitement en l'absence de l'intervention. En outre, le principe ***ceteris paribus*** nous permet de nous assurer que les groupes réagiraient de la même façon à une intervention.

Gertler et coll. (2016, p.52) mentionnent trois critères qu'il convient de considérer au moment d'évaluer si le groupe de traitement et le groupe témoin respectent le principe *ceteris paribus* :

1. Il devrait y avoir un équilibre entre les deux groupes en ce qui concerne les caractéristiques moyennes, autant les caractéristiques observables (âge, sexe, etc.) que les caractéristiques non observables (motivation, capacités, etc.);

2. Le traitement (programme ou politique) ne peut en aucun cas affecter directement ou indirectement le groupe de comparaison (par exemple, les membres du groupe de comparaison ne peuvent pas recevoir le traitement);

Figure 2



3. Les résultats du groupe de comparaison devraient changer de la même façon que ceux du groupe de traitement si le groupe de comparaison était exposé au programme.

Lorsque ces critères sont satisfaits, nous pouvons établir que les deux groupes respectent le principe *ceteris paribus*, c'est-à-dire les deux groupes sont similaires en tous points avant le traitement, sauf qu'un groupe participe au programme, l'autre non. Il s'agit donc d'un groupe témoin ***valide***. Cela nous permet d'attribuer la différence entre les résultats des deux groupes au programme lui-même puisqu'aucun autre élément ne les distingue.

VALIDITÉ INTERNE ET EXTERNE

Deux aspects clés de la mesure des impacts doivent être pris en compte dans toute évaluation :

i. La ***validité interne***, c'est-à-dire la mesure dans laquelle une évaluation permet d'établir efficacement le lien de cause à effet d'une intervention sur les résultats d'intérêt.

ii. La ***validité externe***, c'est-à-dire la mesure dans laquelle les impacts estimées d'une étude peuvent être généralisées pour inclure, par exemple, d'autres régions, groupes de population, etc.

Il est essentiel de comprendre que la validité interne et la validité externe sont des concepts indépendants, et qu'une étude peut avoir une validité interne élevée et une validité externe faible, et vice-versa. Ce guide se concentre principalement sur l'établissement de la validité interne. Une section subséquente est consacrée aux questions liées à la validité externe.

«La validité interne, c'est-à-dire la mesure dans laquelle une évaluation permet d'établir efficacement le lien de cause à effet d'une intervention sur les résultats d'intérêt.»

PRINCIPALES MENACES POUR LA VALIDITÉ INTERNE D'UNE ÉVALUATION DES IMPACTS

Lorsque le principe *ceteris paribus* est respecté, nous pouvons comparer les résultats de notre groupe de traitement et de notre groupe de comparaison, et ***établir une estimation non biaisée des effets[1] du traitement***. Cela revient exactement à dire que les résultats ont une validité interne élevée. Dans le cas des évaluations des impacts qui ne se fondent pas sur des groupes de comparaison valides, certains biais ou certaines menaces pourraient nuire à la validité interne (Campbell, 1957). Les causes les plus fréquentes et les plus importantes de biais sont :

- Biais de sélection;
- Causalité inverse;
- Erreur de mesure systématique;
- Effets du temps;
- Non-conformité;
- Effet placebo;
- Effets comportementaux;
- Biais de l'observateur.

Le but de la mesure des impacts est de choisir un modèle qui élimine, ou du moins réduit au minimum, les sources potentielles de biais. Lorsque les sources de biais ne sont pas adéquatement gérées, la validité interne d'une évaluation des impacts peut être compromise, et les résultats doivent être interprétés avec précaution.

BIAIS DE SÉLECTION

Les *biais de sélection* surviennent lorsque des différences systématiques existent entre les caractéristiques (observables ou non observables) du groupe de traitement et du groupe témoin à cause du processus de répartition des individus dans les deux groupes. Par exemple, cela peut se produire dans un programme de formation professionnelle lorsqu'un plus grand nombre d'individus motivés s'inscrivent au programme ou que les administrateurs du programme sélectionnent les participants qui ont de meilleures capacités. Dans ce cas, le groupe de traitement serait fondamentalement différent du groupe

[1]Les termes « effets du traitement » et « impacts » sont utilisés de façon interchangeable dans ce contexte.

témoin avant même le début du programme de formation professionnelle – plus particulièrement, le groupe de traitement serait plus motivé ou aurait de meilleures capacités que le groupe témoin en raison du processus de sélection. Cela signifie souvent qu'au fil du temps, le groupe de traitement obtiendra naturellement de meilleurs résultats que le groupe témoin, peu importe l'efficacité de la formation professionnelle. Il se pourrait fort bien que dans ce scénario, la comparaison des résultats du groupe de traitement et du groupe témoin mène à une surestimation des impacts du programme de formation professionnelle. Lorsque le biais de sélection est problématique, le groupe témoin n'est pas valide et on tient pour acquis que l'estimation des effets du traitement (toute différence dans les résultats) est *biaisée*.

La raison pour laquelle nous devons nous attarder aux caractéristiques du groupe de traitement et du groupe témoin est que certaines de ces caractéristiques sont très susceptibles d'avoir une incidence sur des résultats d'intérêt qui sont indépendants du programme évalué. Les biais de sélection sont malheureusement très courants dans les études d'évaluation. Pour les programmes auxquels les participants peuvent s'inscrivent de leur propre gré, il pourrait y avoir présence d'un biais de sélection puisque les individus qui choisissent d'y participer sont nécessairement différents de ceux qui choisissent de ne pas y participer. Même si cela peut être attribué au hasard dans certains cas, en pratique, nous ne pouvons jamais tenir pour acquis que ces raisons n'influencent pas les résultats.

Les biais de sélection peuvent entraîner une surestimation, ou une sous-estimation des impacts :

- Il arrive couramment que les individus les plus susceptibles de réussir sont ceux qui s'inscriront de leur propre gré ou qui seront sélectionnés par les administrateurs du programme. Dans ces cas, on tient généralement pour acquis que la simple comparaison des résultats observés entre le groupe de traitement et le groupe témoin mènera à une ***surestimation*** des impacts.
- Dans d'autres cas, les administrateurs du programme pourraient intentionnellement choisir des individus qui ont moins de chances de réussir (p. ex., parce qu'ils ont moins de capacités ou de motivation); une simple comparaison des résultats observés entre le groupe de traitement et le groupe témoin mènera alors à une ***sous-estimation*** des impacts.

Le problème lorsqu'on tente de gérer un biais de sélection, c'est que même si de nombreux facteurs/caractéristiques sont observables (p. ex., sexe, âge, situation d'emploi, revenu, éducation, état de santé), de nombreux facteurs/caractéristiques ne le sont pas. En pratique, les caractéristiques observables sont faciles à gérer dans les évaluations des impacts. Nous pouvons comparer ces types de caractéristiques entre nos deux groupes et, si certaines caractéristiques sont différentes, nous saurons alors que ces deux groupes ne sont pas comparables.

«Les biasis de sélection surviennent lorsque des différences systématiques existent entre les caractéristiques (observables ou non observables) du groupe de traitement et du group témoin à cause du processus de répartition des individus dans les deux groupes.»

Les caractéristiques non observables sont les plus difficiles à gérer, et elles peuvent prendre deux formes. La première forme comprend les éléments difficiles ou carrément impossibles à observer. Des éléments comme les traits de personnalité, le degré de motivation, les habiletés/capacités intellectuelles et les préférences sont généralement non observables, ou du moins très difficiles à mesurer avec précision. La deuxième forme de caractéristique non observable comprend les caractéristiques qui pourraient en fait être observables, mais qui (pour une quelconque raison) ne figurent pas dans les données disponibles. Par exemple, certains éléments considérés comme délicats pourraient ne pas avoir été pris en compte, comme le revenu ou l'orientation sexuelle.

EFFETS DU TEMPS

Les effets du temps (aussi appelés « effets historiques ») sont un enjeu majeur qui peut compromettre la validité des comparaisons simples avant-après des résultats. Ils représentent en partie ce qui se serait produit en l'absence d'une intervention, ce qui signifie qu'ils ne constituent *pas* une retombée des interventions dont nous voulons mesurer les impacts. Chaque fois que nous mesurons un résultat avant une intervention et ce même résultat après une intervention, nous devons considérer que de nombreux éléments peuvent changer au fil du temps dans le contexte global, et qu'un grand nombre de ces éléments peuvent influencer les résultats de façon indépendante du programme. Notre exemple précédent dans lequel un programme d'activité physique se déroule pendant l'été illustre bien le problème que peuvent représenter les effets du temps. Il se peut fort bien que les niveaux d'activité physique fluctuent naturellement avec les changements de saison. Nous devons faire très attention de ne pas attribuer ces changements naturels dans les résultats aux programmes que nous évaluons puisque nous pouvons ainsi facilement surestimer ou sous-estimer les véritables effets du programme.

Dans ces cas, disposer d'un groupe témoin permet d'utiliser des modèles capables d'isoler les effets du temps des effets de l'intervention, à l'aide des techniques décrites plus loin dans ce document. Comme précédemment mentionné, le groupe témoin doit être similaire au groupe de traitement, auquel cas les impacts des effets du temps seront les mêmes au sein du groupe de traitement que du groupe témoin et seront ainsi contrôlées dans le cadre de l'étude.

Toute autre source de biais connexe est ce qu'on appelle la ***régression vers la moyenne***. La régression vers la moyenne est un phénomène où les résultats « extrêmes » (anormalement élevés ou anormalement faibles) se rapprochent naturellement de la valeur moyenne de ces résultats au fil du temps. Prenons pour exemple un programme de tutorat pour les étudiants qui ont des difficultés d'apprentissage. Leurs résultats (dans ce cas-ci, leurs notes) auront tendance à s'améliorer naturellement, simplement parce qu'ils ne peuvent pas vraiment empirer (ils ne peuvent que s'améliorer d'une façon ou d'une autre). Dans ces cas, nous pourrions avoir tendance à surestimer les impacts du programme de tutorat puisqu'il est fort probable que les notes augmenteraient naturellement (dans une certaine mesure) en l'absence de ce programme.

CAUSALITÉ INVERSE

En ce qui concerne le problème de sélection précédemment mentionné, les estimations peuvent être extrêmement biaisées et le seront si la population ciblée par l'intervention a été choisie (ou si les personnes s'inscrivent de leur propre gré) en fonction des résultats qu'on souhaite obtenir à la suite de l'intervention. Par exemple, si un service de police déploie une unité de forces spéciales dans les régions où le taux de criminalité est le plus élevé, une simple comparaison du taux de criminalité entre les régions de traitement et les régions témoins révélera que l'intervention est associée avec des taux de criminalité élevés plutôt que faibles. Une présence policière accrue pourrait générer une augmentation du taux de signalement des crimes, ce qui pourrait donner l'impression que l'intervention (présence policière) exacerbe le problème (crime), même si elle a en réalité l'effet opposé. De façon similaire, si une intervention touchant la santé, comme un programme d'exercice, a tendance à attirer avant tout des gens plus sportifs et en meilleure santé, cela mènera à une estimation biaisée des impacts.

Dans ces cas, ce sont les résultats escomptés qui donnent lieu à l'intervention, et non le contraire. Il peut être utile de voir la causalité inverse comme un scénario de type « la poule ou l'œuf ». Il ne faudrait pas considérer cela comme un simple effet de causalité. Dans de tels cas, les données longitudinales portant sur les résultats au fil du temps permettent d'utiliser des techniques qui donnent une meilleure estimation de l'effet de causalité.

ERREUR DE MESURE SYSTÉMATIQUE

Les erreurs de mesure renvoient aux écarts qui existent entre les valeurs réelles et les valeurs enregistrées d'une mesure. Ces différences peuvent être aléatoires ou non aléatoires/*systématiques*. Par exemple, les individus ont souvent tendance à gonfler leur revenu, ce qui signifie que la valeur enregistrée de leur revenu est supérieure à leur revenu réel.

«Les erreurs de mesure renvoient aux écarts qui existent entre les valeurs réelles et les valeurs enregistrées d'une mesure. Ces différences peuvent être aléatoires ou non aléatoires/systématiques.»

Une erreur de mesure *aléatoire* ne pose pas de problème étant donné que les erreurs surviennent de façon aléatoire à la fois pour les sujets tests et les sujets de comparaison, et que ces anomalies mèneront en théorie à un équilibre entre les deux groupes. Ce sont plutôt les erreurs de mesure *systématiques* qui représentent un problème.

NON-CONFORMITÉ

Certains participants pourraient ne pas se conformer au groupe dans lequel ils ont été répartis. Par exemple, des participants du groupe de traitement pourraient refuser de suivre le programme ou abandonner le programme avant la fin (on appelle ce phénomène *attrition*), tout comme des participants du groupe témoin pourraient trouver une façon de retirer des bienfaits du traitement. La probabilité d'une non-conformité dépend grandement de la nature de l'intervention et de la mesure dans laquelle le comportement des participants est facile à surveiller. Une intervention où il faut visionner une vidéo de 20 minutes sur la santé et la sécurité dans une salle de cours est susceptible de générer un taux de conformité élevé. Or, dans le cadre d'une autre intervention où la même vidéo est envoyée par courriel aux participants, le taux de conformité pourrait être plus faible. Et si l'intervention exige une participation à un programme de formation de six mois, l'attrition pourrait devenir un problème.

La non-conformité n'influence pas l'estimation des effets du traitement si elle est aléatoire. Malheureusement, la non-conformité est rarement aléatoire, et des données probantes de nombreux secteurs tels que la formation professionnelle, la formation sur l'abus d'alcool ou d'autres drogues et la psychothérapie appuient ce fait (Shadish, Cook, & Campbell, 2002, pp. 323-324). Par exemple, les gens moins motivés ou qui ont des problèmes de santé sont plus susceptibles d'abandonner un programme de formation professionnelle. Ces cas de non-conformité non aléatoire représentent un problème parce qu'ils altèrent systématiquement la composition du groupe de traitement ou du groupe témoin, ce qui signifie qu'ils ne sont plus similaires (comparables). Dans ce cas, l'estimation des impacts est susceptible d'exagérer les véritables effets puisque les individus qui demeurent dans le groupe de traitement sont, en moyenne, plus motivés et en meilleure santé que ceux du groupe témoin en raison de l'attrition.

«Malheureusement, la non-conformité est rarement aléatoire, et des données probantes de nombreux secteurs tels que la formation professionnelle, la formation sur l'abus d'alcool ou d'autres drogues et la psychothérapie appuient ce fait (Shadish, Cook, & Campbell, 2002, pp. 323-324).»

Il est généralement impossible d'obtenir une conformité parfaite, autant pour des raisons éthiques (il est souvent impossible d'imposer une intervention de force à quelqu'un sans son consentement) que pratiques (cela exigerait de surveiller constamment un grand nombre d'individus sur une longue période), ou encore pour des raisons inévitables telles que le décès de participants. Une façon de contourner cette difficulté consiste à utiliser un modèle d'encouragement, décrit plus loin dans ce document.

EFFET PLACEBO

L'effet placebo découle de l'*expérience* de recevoir le traitement plutôt que du traitement lui-même. On l'observe souvent dans le domaine de la santé, où c'est un fait bien connu que les patients peuvent réagir à des comprimés de sucre lorsqu'ils croient qu'il s'agit d'un véritable médicament. Cet effet est habituellement contrôlé à l'aide d'une technique appelée « essai à insu », où les participants ne peuvent pas savoir s'ils ont été placés dans le groupe de traitement ou le groupe témoin, généralement parce que les deux groupes reçoivent une forme de traitement (p. ex., un véritable médicament ou un comprimé de sucre identique au médicament).

«L'effet placebo découle de l'expérience de recevoir le traitement plutôt que du traitement lui-même.»

L'effet placebo est souvent difficile à gérer dans les programmes sociaux puisqu'il nécessite l'administration de traitements fictifs ou placebo. Dans l'exemple du programme de formation professionnelle, cela supposerait en fait de faire participer le groupe de contrôle à un programme placebo soigneusement élaboré. Une solution serait de procéder à un essai à volets multiples, dans lequel le groupe de traitement et le groupe témoin reçoivent le traitement, mais où le groupe de traitement reçoit un soutien supplémentaire. Toutefois, il convient de noter que les impacts estimées renvoient aux impacts des options du programme plutôt qu'au programme lui-même. Dans les cas où un programme est modifié d'une quelconque façon, une autre approche à adopter serait d'utiliser le programme original comme référence pour le programme modifié, ce qui permettrait de comparer la nouvelle approche avec le statu quo.

Il faut aussi mentionner que dans certains cadres politiques, les effets placebo font souvent partie des effets souhaités et il n'est pas nécessaire d'en tenir compte. Par exemple, un programme de formation pourrait améliorer le rendement des participants, soit en leur permettant d'acquérir de nouvelles compétences grâce au contenu du cours, soit en augmentant leur motivation par le simple fait qu'ils participent au programme. Les deux cas représenteraient toutefois un résultat souhaité.

EFFETS COMPORTEMENTAUX

Dans le même ordre d'idées que l'effet placebo, le changement de comportement de certains participants au cours d'une évaluation des impacts pose problème. On peut observer ce genre de changement quand certaines personnes savent qu'elles font partie du groupe de traitement ou du groupe témoin lorsqu'aucun placebo n'est utilisé. Les deux principaux effets comportementaux sont les suivants :

- ***Effet Hawthorne :*** Les participants au programme modifient leur comportement parce qu'ils sont observés. Ce nom provient d'une étude réalisée à Hawthorne Works, une usine électrique située dans l'État de l'Illinois. Dans le cadre de cette étude, une augmentation du rendement des travailleurs au départ attribuée à un meilleur éclairage et à une plus grande propreté des postes de travail a par la suite été attribuée au fait que les sujets savaient qu'ils participaient à une étude et qu'ils portaient plus attention à leur travail pour cette raison. En résumé, le groupe de traitement peut obtenir de meilleurs résultats même si le programme n'a aucun effet et, par conséquent, l'effet Hawthorne mène à une surestimation biaisée des impacts.

- ***Effet John Henry :*** les membres du groupe de comparaison (non-participants) modifient leur comportement lorsqu'ils savent qu'ils ne reçoivent pas le traitement. Par dépit, les membres du groupe de comparaison pourraient être motivés à en faire plus que d'habitude pour améliorer leurs résultats afin de montrer aux administrateurs du programme qu'ils peuvent s'améliorer sans l'aide du programme. Dans ce cas, les résultats du groupe témoin peuvent s'améliorer et l'effet John Henry mène à une sous-estimation biaisée des impacts.

Comme pour l'effet placebo, il est recommandé, dans la mesure du possible, de mener des essais à l'insu des sujets, qui ne savent pas à quel groupe ils appartiennent et sont soumis à des conditions égales (les deux

groupes sont surveillés avec le même degré d'attention), afin d'éviter les effets comportementaux tels que l'effet Hawthorne et l'effet John Henry.

Dans le domaine des sciences médicales, la pratique exemplaire consiste à utiliser un protocole à double insu. Dans ce cas, le groupe de traitement et le groupe témoin, ainsi que les administrateurs du programme (ceux qui distribuent les comprimés), ne savent pas quels comprimés (véritable médicament ou placebo) ils prennent ou distribuent. Les essais à insu dans le domaine des sciences sociales sont généralement beaucoup plus difficiles à réaliser – par exemple, même dans le cas d'un programme de formation professionnelle fictif ou placebo pour un groupe témoin, les administrateurs du programme sauront quel programme ils donnent, ce qui pourrait fournir des indices aux participants du groupe de comparaison. En réalité, ces types de placebos sont exigeants sur le plan des ressources et ne sont généralement pas pratiques.

BIAIS DE L'OBSERVATEUR

L'évaluateur (observateur) pourrait consciemment ou inconsciemment s'attendre à différents résultats de la part des sujets traités et des sujets non traités. Cela pourrait mener à une évaluation biaisée de la part de l'évaluateur au moment de consigner les mesures (plus particulièrement, si l'évaluation est subjective, l'évaluateur pourrait juger instinctivement que le groupe traité a mieux réussi parce qu'il/elle *sait* que le groupe était traité), ou encore l'évaluateur pourrait envoyer différents signaux/indices aux sujets des deux groupes, ce qui pourrait influencer leur comportement. Lorsque les biais de l'observateur peuvent s'avérer problématiques, la solution typique consiste à mener un essai à double insu, où ni les participants ni les responsables de la saisie des données ne savent qui fait partie du groupe de traitement et du groupe de comparaison. L'affectation des participants à l'un ou l'autre des groupes est déterminée par un tiers qui ne connaît pas les caractéristiques des participants.

MATRICE DE PREUVES D'IMPACT CANADA

MATRICE DE PREUVES D'IMPACT CANADA

La hiérarchie des preuves (ou pyramide des preuves) est un principe qui a gagné en popularité au cours des dernières décennies. Cette hiérarchie classe les méthodes de mesure des impacts en fonction de leur validité interne. Les essais contrôlés randomisés (ECR) occupent généralement le sommet du classement, suivis des méthodes quasi expérimentales et des modèles de base (avant-après). Cette approche a beaucoup de mérite, bien qu'elle puisse entraîner une perception erronée que les ECR représentent une méthode « de haut calibre » universelle. Cela peut réduire le recours à d'autres méthodes rigoureuses qui pourraient s'avérer plus appropriées dans certains contextes. En outre, les ECR ne sont pas toujours réalisables sur le plan administratif. Les programmes de subventions et de contributions, par exemple, ne financent pas toujours les ECR étant donné qu'on considère qu'il y a chevauchement avec les sources de financement de la recherche.

Bien que la validité interne soit de toute évidence le principal critère pris en compte pour le choix de la méthode de mesure des impacts, d'autres préoccupations sont également pertinentes :

- ***Éthique*** **:** Dans de nombreux cas, la répartition aléatoire pourrait être (ou être perçue comme étant) contraire à l'éthique, tout comme le fait de refuser des services à un groupe afin de créer un groupe de comparaison. De nombreux fournisseurs de services travaillent avec des ressources limitées et ciblent délibérément les individus qui ont les plus grands besoins. En revanche, il peut être contraire à l'éthique d'attribuer certains éléments de façon aléatoire. Par exemple, fumer cause le cancer, mais aucune étude expérimentale ne le démontre puisqu'il est contraire à l'éthique d'attribuer du tabac de façon aléatoire.

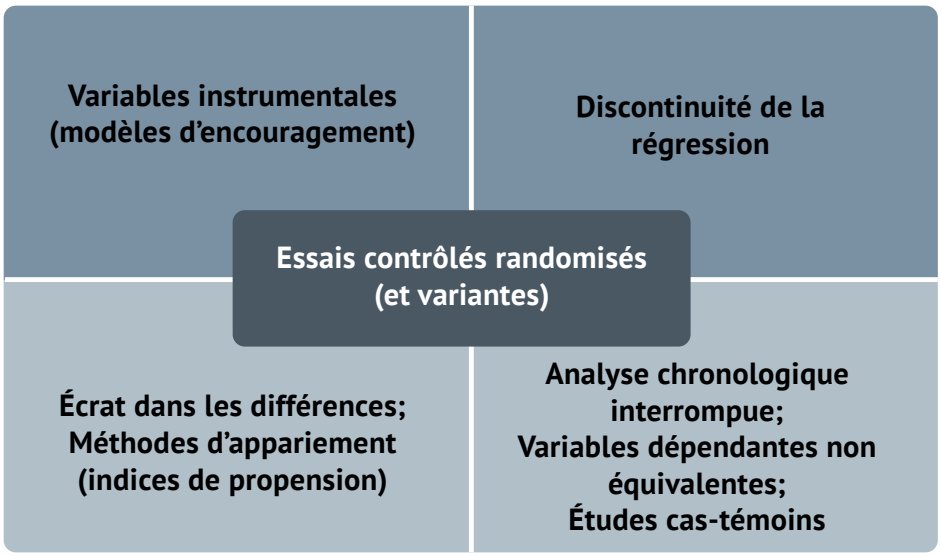
- ***Faisabilité*** **:** Nous pouvons nous intéresser aux impacts de nombreux éléments sur les résultats, sans toutefois pouvoir les attribuer de façon aléatoire. Le sexe ou les traits de personnalité en sont de bons exemples. Nous ne pouvons pas *manipuler* ces éléments, donc nous ne pouvons pas les attribuer de façon aléatoire.

- ***Impossibilité d'exclure*** **:** Certains types de programmes ou d'interventions se déroulent dans un contexte où il est impossible de créer un groupe de contrôle puisque tous les participants sont simultanément exposés à l'intervention. La législation, la réglementation et l'imposition en sont de bons exemples. Nous ne pouvons pas étudier les impacts de la légalisation du cannabis sur les Canadiens à l'aide d'un essai contrôlé randomisé puisque la loi s'applique de façon universelle à tous les Canadiens – il n'y a pas de groupe de contrôle.

Il n'est pas logique de laisser entendre qu'une « norme absolue » s'applique dans tous les cas. Dans les cas où il est possible d'établir des groupes témoins, les approches randomisées seront généralement les méthodes les plus efficaces étant donné qu'elles comportent le plus haut degré de validité interne.

Il est toutefois impossible de répondre à certaines questions sur les impacts à l'aide de la randomisation, c'est pourquoi les méthodes non randomisées (quasi-expériences), même si leur degré de validité interne est moins élevé, sont souvent notre meilleur choix. Pour Impact Canada, notre approche consiste à nous efforcer d'utiliser la méthode (validité interne) la plus rigoureuse possible étant donné les réalités opérationnelles du programme dont nous mesurons les impacts. Cela suppose d'utiliser des techniques expérimentales et quasi expérimentales, selon les besoins.

Figure 3



Au lieu d'une pyramide ou d'une hiérarchie traditionnelle, l'UII utilise une **matrice de preuves** qui présente un menu des options possibles pour mesurer les impacts et qui illustre trois principes clés :

- L'absence d'une hiérarchie claire indique qu'aucune méthode n'est universellement applicable à tous les cas.
- Cela étant dit, les méthodes randomisées, notamment les ECR, les études par étapes/avec liste d'attente et les plans multifactoriels, occupent une place unique dans le menu des options en tant que catégorie de méthodes présentant le plus haut degré de validité interne.
- L'efficacité d'une méthode sur le plan de la validité interne est généralement indiquée sur une échelle de couleurs, où les couleurs plus foncées indiquent une validité interne plus élevée.

«Pour Impact Canada, notre approche consiste à nous efforcer d'utiliser la méthode (validité interne) la plus rigoureuse possible étant donné les réalités opérationnelles du programme dont nous mesurons les impacts»

Le classement des différentes méthodes de mesure des impacts repose sur le principe clé selon lequel la rigueur s'observe dans la mise en œuvre plutôt que dans le nom. Un schéma d'appariement bien exécuté pourrait très bien s'avérer plus efficace qu'un modèle de discontinuité de la régression mal exécuté, c'est pourquoi il faut faire preuve de vigilance en ce qui concerne la qualité. Étant donné que ces méthodes peuvent être difficiles à concevoir et à mettre en œuvre, l'UII encourage la collaboration, dans la mesure du possible, avec des experts qui fourniront conseils et orientation pour qu'on puisse assurer l'exactitude des méthodes d'évaluation des impacts et des conclusions qui en découlent.

CHOISIR LE BON MODÈLE

Selon le principe de causalité et la définition de la mesure des impacts décrits dans le présent document, nous avons classifié les méthodes décrites ci-après dans ***l'Échelle des preuves d'Impact Canada***. Cette échelle classe les différentes méthodes par ordre d'efficacité au chapitre de la validité interne (c.-à-d. capacité à fournir des estimations rigoureuses des impacts). Comme nous l'avons vu précédemment, il convient de noter que ce classement ne tient pas compte des problèmes associés à la validité externe (c.-à-d. généralisabilité des résultats de l'étude). Cet aspect est abordé à la fin du présent guide.

Pour que le classement soit plus facile à visualiser et à comprendre, nous avons réparti les méthodes dans six niveaux de validité distincts. Pour chaque niveau de l'échelle, une liste des méthodes associées à la catégorie en question est fournie. Veuillez noter que certaines méthodes comportent plusieurs variantes (p. ex., les méthodes d'appariement peuvent être combinées avec la méthode « écart entre les différences »), mais elles sont classées au même niveau que la méthode à laquelle elles sont associées.

Conformément à une pratique exemplaire dans ce domaine (qui se fonde sur les seuils de preuve généralement acceptés par de nombreux gouvernements et organisations du secteur public dans les pays de l'OCDE), nous établissons le seuil minimal au niveau 4. L'utilisation de méthodes de niveau 5 et de niveau 6 n'est pas recommandée pour la mesure des impacts étant donné que ces méthodes ne satisfont pas à une norme fondamentale pour la mesure des impacts, à savoir une validité interne élevée. En règle générale, il convient de privilégier des méthodes de niveau aussi élevé que possible dans le classement, mais les ressources utilisées pour l'évaluation doivent être proportionnelles à la portée du programme évalué.

Au moment d'utiliser l'Échelle des preuves d'Impact Canada, il est important de ne pas oublier que la rigueur réside dans la mise en œuvre de la méthode plutôt que dans son nom. La rigueur d'une méthode est fonction de certaines prémisses. Certaines des méthodes les plus complexes, tout particulièrement les évaluations à l'aide de variables instrumentales, peuvent être très instables, et toute estimation établie à partir d'un instrument non valide ou inefficace peut générer d'importants biais et s'avérer considérablement moins efficace qu'une évaluation à l'aide d'une méthode de niveau 1.

« Au moment d'utiliser l'Échelle des preuves d'Impact Canada, il est important de ne pas oublier que la rigueur réside dans la mise en œuvre de la méthode plutôt que dans son nom. »

Un dernier point qui doit être mentionné ici est que le niveau associé à la méthode n'est d'aucune façon lié au niveau d'efforts ou de ressources requis pour mener l'étude, donc le budget ne devrait pas être pris en compte au moment de déterminer l'étude à réaliser. Par exemple, une étude de l'écart entre les différences ou une étude cas-témoins peut exiger plus de temps et d'argent qu'une expérience, qui exige une bonne planification, mais moins d'analyse statistique qu'une étude des écarts entre les différences ou une étude cas-témoins.

Les sections suivantes présentent les principales caractéristiques de chaque méthode de mesure des impacts, traitent des différents contextes dans lesquels ces méthodes sont susceptibles d'être le plus efficaces et abordent certains éléments clés à considérer pour mesurer la qualité des évaluations des impacts réalisées à l'aide de ces méthodes. Ces sections sont suivies d'une brève section traitant des avantages uniques des méthodes qualitatives utilisées pour mesurer les impacts.

ÉCHELLE DES PREUVES D'IMPACT CANADA

NIVEAU	DESCRIPTION	MÉTHODES
1	L'intervention est menée de façon aléatoire ou « considérée comme étant menée de façon aléatoire » dans le cadre de processus naturels, notamment lorsque le groupe de traitement et le groupe témoin présentent des caractéristiques de base en moyenne identiques au chapitre des variables observables et non observables. Cela signifie que les groupes sont identiques (en expectation) et obtiendraient les mêmes résultats en l'absence de l'intervention. Par conséquent, tout écart dans les résultats observés après l'intervention est uniquement attribuable à celle-ci. Ces modèles d'étude donnent lieu à des estimations non biaisées de la causalité présentant les niveaux les plus élevés de validité interne.	Essais contrôlés randomisés (ECR)/Expérimentation
2	L'intervention est menée de façon aléatoire ou une variation indépendante dans l'intervention peut être isolée après l'application des méthodes d'estimation respectives. Toutefois, certains facteurs de confusion liés à l'administration de l'intervention (comme des effets placebo ou des effets découlant du fait d'être observé) pourraient biaiser l'estimation des effets du traitement. Des preuves convaincantes ont été présentées pour appuyer l'argument selon lequel les hypothèses émises à l'aide de la méthode d'estimation sont valables.	Introduction par étapes Estimation à l'aide de variables instrumentales (VI) Modèle de discontinuité de la régression
3	L'intervention n'est pas menée de façon aléatoire. La méthode d'estimation permet de contrôler les biais de sélection découlant de facteurs de confusion observables et au minimum quelques-uns des biais découlant de facteurs de confusion non observables. Des preuves raisonnables ont été présentées pour appuyer l'argument selon lequel les hypothèses émises à l'aide de la méthode d'estimation sont susceptibles d'être valables. Nous ne pouvons pas conclure avec certitude qu'il y a un effet de causalité. Toutefois, les méthodes de niveaux 2, 3 et 4 se sont révélées efficaces (Concato, Shah, & Howritz, 2000) et constituent la forme d'évaluation dominante pour l'analyse des politiques.	Écart entre les différences Interruption du traitement et variables dépendantes non équivalentes
4	L'intervention n'est pas menée de façon aléatoire. La méthode d'estimation permet de contrôler les biais de sélection découlant de facteurs de confusion observables, mais pas les biais découlant de facteurs de confusion non observables. Il y a de fortes chances que les hypothèses émises à l'aide de la méthode d'estimation ne sont pas valables. Nous ne pouvons donc pas conclure avec certitude qu'il y a un effet de causalité. Toutefois, les méthodes de niveaux 2, 3 et 4 se sont révélées efficaces et constituent la forme d'évaluation dominante pour l'analyse des politiques.	Appariement Études cas-témoins

5

L'intervention n'est pas menée de façon aléatoire, mais les données sont connues pour le groupe traité et le groupe non traité. Les méthodes d'estimation ne sont pas à l'origine du biais de sélection. Au mieux, ces méthodes peuvent montrer une corrélation entre l'intervention et les résultats, mais la corrélation n'est pas la causalité étant donné qu'elle pourrait être attribuable à une multitude d'autres raisons : les résultats pourraient être à l'origine de la participation à l'intervention (causalité inverse); les résultats pourraient être attribuables à la situation et aux caractéristiques particulières du groupe traité (biais de sélection et régression vers la moyenne); les résultats pourraient être attribuables à un certain nombre d'autres événements (effets historiques). Ces méthodes présentent des niveaux minimaux de validité interne et seront généralement extrêmement biaisées. Le biais sera généralement ascendant, c'est-à-dire que les évaluations des impacts fondées sur ces modèles d'étude surestimeront systématiquement les impacts de l'intervention. Par conséquent, ces méthodes ne devraient pas être utilisées pour mesurer les impacts , mais plutôt pour formuler des hypothèses et fournir des preuves appuyant les études de rang plus élevé dépassant le seuil de validité de niveau 4.

Écart dans les moyennes

Estimation avant-après

Analyse comparative à l'aide de données agrégées

6

Aucun groupe de comparaison (groupe témoin ou groupe de traitement avant la réception du traitement) ne permet de comparer les observations relatives au traitement. Parallèlement, les conclusions de l'étude ne sont pas fondées sur des données individuelles, mais plutôt sur des arguments théoriques ou des opinions subjectives qui ne sont pas appuyés par des preuves et qui ne tiennent pas compte des éventuelles menaces touchant la validité. Ces méthodes ne devraient pas être utilisées pour mesurer les impacts en soi, mais plutôt pour formuler des hypothèses et fournir des preuves appuyant les études de rang plus élevé.

Énoncés de lien de causalité non appuyés par des données ou d'autres études

MISES EN GARDE ET AVERTISSEMENTS

1. La classification de l'Échelle des preuves d'Impact Canada se fonde uniquement sur la validité interne des méthodes et ne tient pas compte de la généralisabilité des résultats (validité externe).

2. L'échelle s'applique aux évaluations individuelles. Certaines échelles de preuves dans le domaine des sciences médicales comportent des niveaux plus rigoureux selon que les résultats ont été ou non reproduits dans des études de suivi. La reproduction des résultats est alors utilisée pour renforcer les conclusions d'une étude.

3. Le contexte dans lequel l'étude a été réalisée est également important. Le contexte peut suggérer que certaines hypothèses ne sont pas valables ou que le budget, les ressources et les conditions initiales ne conviennent pas à une méthodologie donnée. Si les hypothèses ne sont pas confirmées ou que la méthodologie ne peut pas être adéquatement appliquée, les méthodes plus élevées dans le classement ne se révéleront alors pas plus efficaces que toute autre méthode de l'échelle. Étant donné que le contexte dans lequel l'étude est réalisée est important, les échelles de preuves comme celle décrite ci-dessus devraient être vues comme des guides pour évaluer la validité des différentes méthodes plutôt que comme une règle absolue.

MÉTHODES D'ESTIMATION DES IMPACTS

MÉTHODES D'ESTIMATION DES IMPACTS

Les sections suivantes fournissent un aperçu des principales méthodes de mesure des impacts. Elles décrivent ces méthodes en détail et les structurent en ordre de rigueur décroissant en fonction de leur capacité à cerner adéquatement les résultats contrefactuels. L'ordre de présentation des méthodes dans une même sous-section est arbitraire. Un bref aperçu de ces méthodes, ainsi que de leurs avantages et désavantages, est fourni dans une annexe à titre de référence. Une section subséquente aborde le rôle complémentaire important que jouent les méthodes qualitatives dans la mesure des impacts.

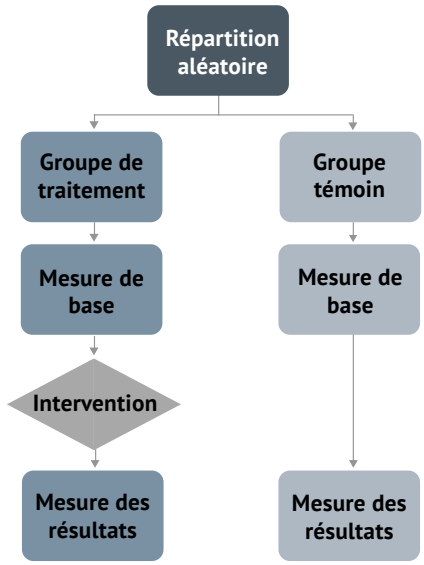
L'ESSAI CONTRÔLÉ RANDOMISÉ ET L'APPROCHE EXPÉRIMENTALE

L'*expérimentation* (un terme souvent utilisé de façon interchangeable avec le terme « *essai contrôlé randomisé* ») est la meilleure façon de respecter le principe *ceteris paribus*. Lorsqu'elles sont bien exécutées, les expérimentations sont les méthodes qui comportent le plus haut niveau de validité interne, ce qui signifie qu'elles constituent d'excellentes méthodes de mesure des impacts là où elles s'appliquent.

La caractéristique clé des expérimentations est la répartition aléatoire dans le groupe de traitement et le groupe témoin. La répartition aléatoire dans les deux groupes signifie que chaque individu est associé à un groupe uniquement par hasard. Cette méthode comporte deux grands avantages :

1. Elle permet de s'assurer que les deux groupes sont équilibrés (au chapitre des attentes) en ce qui a trait aux éléments observables et aux éléments non observables. Les deux groupes seront aussi semblables que possible. Étant donné que les deux groupes ne sont pas intrinsèquement différents, aucun élément n'est susceptible d'exercer une influence différente sur les résultats.
2. Cela assure également que toute menace touchant la validité est répartie de façon aléatoire entre les deux groupes. Ainsi, tous les autres facteurs susceptibles de modifier les résultats sont eux aussi équilibrés entre les deux groupes, donc ces effets « s'annulent ». Par conséquent, toute variation observée dans les résultats peut être directement attribuée au programme, à l'exclusion de tout autre facteur.[2]

Figure 4



[2] Pour une discussion plus detailée de la théorie de la répartition aléatoire, voir (Shadish, Cook, & Campbell, 2002, pp. 248-251)

Cela signifie que la répartition aléatoire est une façon extrêmement efficace d'écarter d'autres explications pour les résultats changeants. Avec la répartition aléatoire, tout changement dans les résultats observés dans les deux groupes ne peut pas être attribué aux écarts entre les deux groupes étant donné qu'il n'y a pas d'écarts, ni à d'autres causes étant donné que toute cause existante a un effet équivalent sur les résultats des deux groupes.

L'autre caractéristique clé de l'expérimentation réside dans les « mesures de contrôle ». Nous pouvons considérer les mesures de contrôle comme des caractéristiques qui préservent les avantages de la répartition aléatoire. Contrôler une étude signifie préserver la composition du groupe test et du groupe témoin pour éviter qu'elle change au fil du temps. Cela signifie habituellement :

- S'assurer qu'il n'y a pas d'***attrition*** dans les groupes. Nous savons que l'attrition (abandon) n'est généralement pas aléatoire, et qu'il est donc très possible que l'étude commence avec des groupes répartis de façon aléatoire, mais se termine avec des groupes qui ne sont plus comparables en raison de l'attrition;

- S'assurer qu'il n'y a pas de ***contamination***, c'est-à-dire que tous les participants du groupe test sont effectivement exposés au programme/au traitement/à l'intervention et qu'aucun participant du groupe témoin n'y est exposé;

- Protéger les groupes des ***effets comportementaux***. Comme il est indiqué précédemment, les sujets d'études contrôlées peuvent modifier leur comportement lorsqu'ils sont conscients qu'ils sont exposés à un traitement. De façon similaire, les administrateurs du programme ou encore les individus qui évaluent les impacts pourraient se comporter différemment envers le groupe test et envers le groupe témoin. Ces « changements » de comportement ont souvent une influence sur les résultats qui nous intéressent, ce qui signifie que tout changement observé dans les résultats ne peut pas être entièrement attribué au programme lui-même. Les meilleures évaluations des impacts sont menées à l'insu de tous, ce qui signifie qu'aucune des personnes en cause (participants, administrateurs du programme, évaluateurs, etc.) ne peut distinguer le groupe de traitement du groupe témoin. En pratique, il est difficile de mener des évaluations des impacts à l'insu de tous (voir l'explication ci-dessus).

Il faut savoir que les expérimentations sont bien plus que de simples évaluations randomisées – d'autres caractéristiques sont nécessaires pour préserver le modèle de répartition aléatoire tout au long du processus d'évaluation des impacts. Nous devons être conscients qu'une expérimentation qui s'accompagne de mesures de contrôle inadéquates pourrait ne pas respecter le principe *ceteris paribus*, et que les résultats de telles évaluations des impacts devraien être interprétés avec prudence.

VARIANTES DES ECR

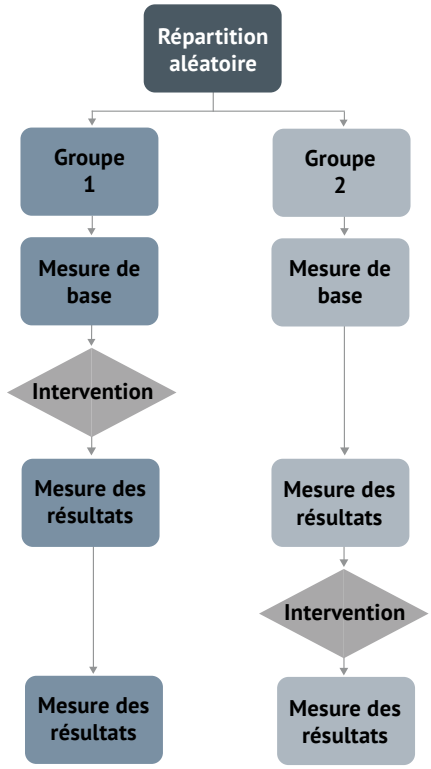
L'approche traditionnelle associée aux ECR consiste à répartir simultanément et aléatoirement les participants dans le groupe de traitement et le groupe témoin, et à mesurer les résultats après le traitement. Toutefois, cela n'est pas toujours possible et dans ces cas, il y a des façons d'utiliser la répartition aléatoire de façon créative pour obtenir à peu près le même effet. L'introduction par étapes, les essais à volets multiples et les plans factoriels sont quelques-unes des approches couramment utilisées dans les modèles de répartition aléatoire pour l'évaluation pragmatique des résultats. Ces différentes variantes du modèle d'ECR classique peuvent être attrayantes puisqu'elles peuvent très bien s'intégrer dans les modes d'opération naturelle de nombreux programmes et dans de nombreux contextes de prestation de services.

INTRODUCTION PAR ÉTAPES

Aussi appelé ***essai par étapes***, ***liste d'attente*** ou ***comparaison de pipeline,*** ce modèle est une adaptation de l'ECR, où une intervention est déployée en deux étapes ou plus auprès des participants, et où les participants sont répartis de façon aléatoire dans les différentes phases. Cette approche peut s'appliquer dans le cadre de programmes à court terme, dont les résultats sont observables peu de temps après la fin de l'intervention. Cette méthode est souvent employée lorsqu'il y a des objections éthiques à refuser le traitement à certains participants étant donné qu'au bout du compte, tous les participants se seront vu offrir le traitement.

La principale caractéristique de ce modèle est que le groupe témoin reçoit l'intervention après le groupe de traitement, plutôt que de ne pas le recevoir du tout. Il est donc possible d'identifier les effets sur le groupe de traitement en le comparant au groupe témoin avant que celui-ci reçoive l'intervention, ou simplement en observant les effets découlant du fait qu'un des groupes a été exposé au traitement plus longtemps que l'autre (pour les interventions à plus long terme). La répartition dans les phases plus précoces ou plus tardives du traitement doit tout de même demeurer aléatoire, c'est-à-dire indépendante de toute caractéristique individuelle, pour que l'étude soit efficace et mène à des estimations présentant un haut degré de validité interne.

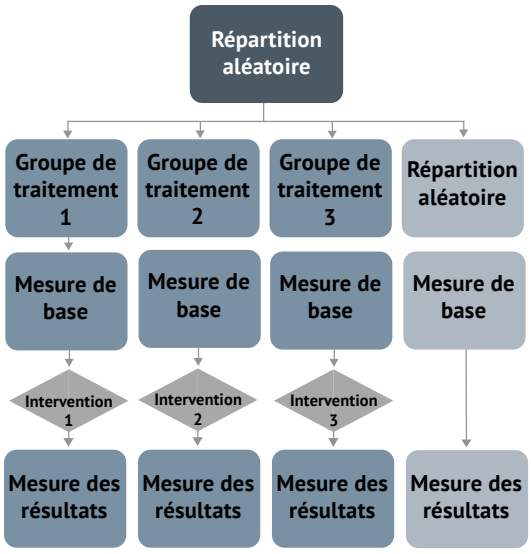
Figure 5



MODÈLES À VOLETS MULTIPLES ET PLANS FACTORIELS

Les ECR à volets multiples testent simultanément plusieurs options d'un programme. Dans ce modèle, les participants sont répartis de façon aléatoire dans un programme parmi plusieurs autres, ce qui permet de comparer les résultats des différentes options visées par l'essai. Plus particulièrement, il est possible d'utiliser un ECR à volets multiples qui ne comprend pas de groupe témoin, mais qui compare plutôt les résultats de différents programmes. Ce modèle peut également être utile dans les cas où il est impossible de refuser des traitements aux participants admissibles. Parallèlement, lorsqu'on hésite à former un véritable groupe témoin (aucun programme), on peut intégrer un « volet » dans lequel les participants bénéficient d'une intervention du programme à un degré minime, qui peut ensuite être utilisée dans l'analyse pour estimer les résultats qu'auraient obtenus les participants s'ils n'avaient pas participé au programme.

Figure 6



QUAND PEUT-ON UTILISER LES MÉTHODES RANDOMISÉES?

Les méthodes dans cette catégorie se caractérisent par le fait que l'administrateur/l'évaluateur du programme peut déterminer à l'avance la répartition des participants au programme/à l'intervention. Ces méthodes pourraient s'appliquer dans les circonstances suivantes :

- Les programmes sont extrêmement populaires. Lorsque le nombre de participants admissibles dépasse le nombre de places offertes dans le cadre du programme, l'approche la plus juste ou la plus éthique est fort probablement la sélection aléatoire.
- Le programme est offert en phases ou en cohortes. Ce type de programme permet l'utilisation de modèles « par étapes », et les participants peuvent être répartis de façon aléatoire dans les différentes phases.
- Lorsqu'un programme est « ajusté » ou que des variantes d'un programme sont mises à l'essai. Les participants peuvent être répartis dans les différentes versions du programme, ce qui permet d'appliquer des modèles à volets multiples/plans factoriels.

POINTS À CONSIDÉRER

- L'intervention doit être attribuée de façon aléatoire.
- Les participants dovent se conformer au groupe dans lequel ils ont été reepartis aleeatoirement (ils ne doivent pas recevoir l'intervention s'ils ont éte placés dans le groupe témoin, tout comme ils doivent la recevoir s'ils sont dans le groupe de traitement).
- L'évaluateur doit s'assurer que l'estimation des impacts n'est pas liée au simple fait que les participants savent qu'ils sont observés ou qu'ils savent a quel groupe ils appartiennent (groupe de traitement ou groupe témoin).

APPROCHES QUASI EXPÉRIMENTALES

Il arrive souvent que les administrateurs des programmes ne peuvent pas, ou ne veulent pas, utiliser un modèle de répartition aléatoire, ce qui rend le modèle expérimental impossible à appliquer. Dans ces cas, il y a des façons d'estimer les résultats contrefactuels en procédant à des quasi-expériences. Les quasi-expériences sont définies comme des modèles « caractérisés par une répartition non aléatoire » (Dunning, 2013, p. 19), mais qui tentent de reproduire un grand nombre des avantages de l'ECR. La caractéristique clé des approches quasi expérimentales est, par conséquent, qu'elles ne permettent pas de déterminer la répartition dans le cadre du programme ou de l'intervention.

«Les approches quasi expérimentales se définissent comme des modèles « caractérisés par une répartition non aléatoire.»»

Les quasi-expériences s'accompagnent d'une grande diversité d'approches, mais elles ont toutes comme objectif commun de tenter de reproduire autant que possible la rigueur de l'ECR dans les scénarios où la répartition aléatoire n'est ni réalisable ni souhaitable. Dans le présent document, elles sont divisées en méthodes quasi expérimentales de « niveau supérieur » et de « niveau inférieur ».

- Par nature, les méthodes de niveau supérieur permettent de contrôler une plus grande variété de facteurs, notamment des caractéristiques non observables. Dans les bonnes conditions, elles peuvent mener à des estimations plus justes des causes et des effets, à des niveaux d'efficacité semblables à ceux des ECR.

- Les méthodes quasi expérimentales de niveau inférieur permettent uniquement de contrôler les caractéristiques observables et, par conséquent, sont uniquement valides pour les hypothèses où seuls des facteurs observables affectent les résultats et l'état du traitement.

- Néanmoins, comme dans toute approche, ces méthodes comportent des avantages et des désavantages, et même les méthodes de niveau supérieur s'accompagnent d'hypothèses fondamentales qui peuvent être réfutées. Autant les méthodes de niveau supérieur que les méthodes de niveau inférieur peuvent mener à des estimations non biaisées des impacts dans certains cas, et à des estimations fortement biaisées dans d'autres.

MÉTHODES QUASI EXPÉRIMENTALES DE NIVEAU SUPÉRIEUR

Les méthodes quasi expérimentales ci-dessous sont classées au niveau supérieur et sont plus efficaces que d'autres méthodes quasi expérimentales étant donné qu'elles peuvent faire ressortir et contrôler des écarts observables et des écarts non observables entre le groupe de traitement et le groupe de comparaison. Cela signifie que, lorsqu'elles sont bien exécutées, elles représentent des méthodes de mesure des impacts efficaces.

VARIABLES INSTRUMENTALES

L'approche par variables instrumentales (VI) tire ses origines du domaine de l'économie, où elle est utilisée depuis environ cinquante ans. De nos jours, elle gagne en popularité dans le domaine des sciences sociales en général, et dans la mesure des impacts en particulier. L'approche par VI mise sur une source aléatoire externe de variation d'une variable pour établir son effet de causalité sur une variable de résultats. Par exemple, il est difficile d'isoler l'effet de causalité du revenu sur le niveau de bonheur puisque les gens plus heureux pourraient gagner plus d'argent (causalité inverse). Toutefois, les gains à la loterie sont aléatoires et représentent une source de variation aléatoire dans le revenu d'une personne; nous pouvons donc utiliser cette variation aléatoire du revenu (attribuable à un gain à la loterie) pour estimer l'effet de causalité du revenu sur le niveau de bonheur.

Dans l'exemple ci-dessus, les gains à la loterie sont un exemple de variable instrumentale (VI), utilisée comme instrument pour le revenu. Donc, qu'est-ce qu'une VI? En langage simple, une VI peut se décrire comme tout élément qui influence le facteur dont nous souhaitons mesurer les effets, mais qui est en soi exempt de biais relatifs, entre autres, à la sélection, aux effets du temps et à la causalité inverse. Cela se produit généralement lorsque la VI est aléatoire (comme les gains à la loterie ci-dessus) ou attribuée de façon aléatoire par l'évaluateur (voir la section Modèle d'encouragement ci-après).

Deux hypothèses fondamentales s'imposent pour qu'un instrument soit considéré comme valide. Un bon instrument doit :

1. affecter la probabilité qu'un individu participe à un programme; mais

2. ne pas avoir d'incidence directe sur les résultats de ce participant, sauf par l'intermédiaire de son influence sur la participation de cet individu.

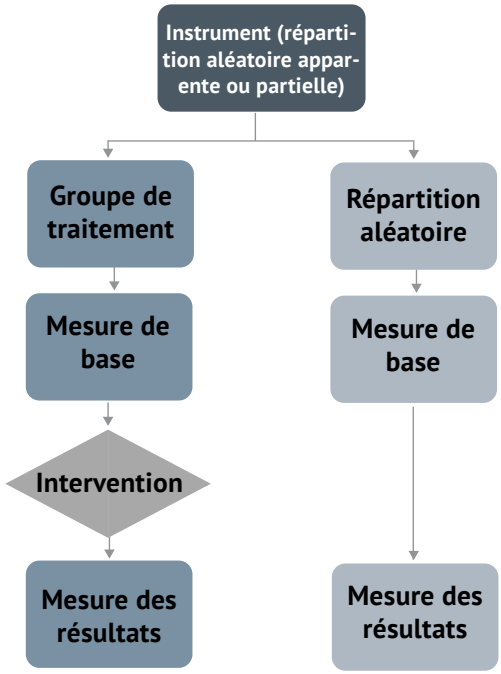
«L'approche par VI mise sur une source aléatoire externe de variation d'une variable pour établir son effet de causalité sur une variable de résultats.»

Le première hypothèse renvoie à la ***pertinence au premier stade***, et une VI qui ne répond pas à ce critère est considérée comme une VI *faible*. La seconde hypothèse renvoie à la ***restriction en matière d'exclusion***.

Si ces hypothèses sont valides, comme dans le cas d'un ECR, tous les autres facteurs (observables et non observables) autres que le traitement lui-même sont contrôlés, donc nous pouvons obtenir un effet de traitement qui présente une validité interne élevée au moyen de la VI.

C'est pour cette raison que la VI se situe dans le niveau supérieur de la catégorie des méthodes quasi expérimentales. Toutefois, si ces hypothèses sont pas valables, les estimations peuvent être extrêmement biaisées. La VI ne devrait par conséquent pas être utilisée sans argument plausible en faveur de ces deux hypothèses. Les tests statistiques destinés à vérifier la validité des hypothèses devraient toujours être réalisés et combinés avec un raisonnement théorique adéquat.

Figure 7



IV est courant dans ce que l'on appelle les « expériences naturelles ». Elles sont « naturelles » en ce sens qu'une certaine forme de répartition aléatoire se produit naturellement, et non par l'intermédiaire des efforts d'un évaluateur. Par exemple, une étude bien connue aux États-Unis (Angrist & Lavy [1999]) a examiné l'effet des années de scolarité sur les gains ultérieurs en utilisant la date de naissance comme instrument pour les années de scolarité. Aux États-Unis, les lois sur la scolarité obligatoire exigeaient que les élèves restent à l'école au moins jusqu'à leur seizième anniversaire. La plupart des États exigent également que les élèves entrent à l'école au cours de l'année civile où ils atteignent l'âge de six ans, ce qui signifie que les élèves entrent à l'école à des âges différents. Les élèves qui sont entrés à l'école à un âge légèrement plus avancé (qui sont généralement nés plus tôt dans l'année) ont atteint l'âge légal d'abandon scolaire après avoir terminé moins de scolarité que leurs homologues plus jeunes. Comme la date de naissance n'est probablement pas liée à d'autres attributs personnels qui déterminent le revenu (ce qui le rend aléatoire), cela a introduit une variation exogène dans le niveau de scolarité atteint, ce qui a permis aux auteurs d'évaluer l'impact causal de la scolarité sur le revenu.

MODÈLE D'ENCOURAGEMENT

Les modèles d'encouragement sont un cas spécial de VI particulièrement utile pour la mesure des impacts. Dans les cas où on ne peut pas choisir aléatoirement les personnes qui bénéficieront de l'intervention ou du traitement (peut-être parce que la prestation de l'intervention n'est pas entièrement du ressort de l'évaluateur), ou lorsque la probabilité que les participants abandonnent le traitement est élevée, une autre option consiste « encourager » de manière aléatoire certaines personnes à participer à l'intervention[3]. En pratique, dans les modèles d'encouragement, tous les participants potentiels sont informés de la disponibilité d'un programme, mais une sélection aléatoire de participants se voit offrir un incitatif supplémentaire à s'y inscrire. Cet incitatif peut prendre la forme d'un appel téléphonique ou d'un rappel par courriel, d'un bon de réduction ou d'un autre mode d'encouragement financier ou non financier, qui rendra le groupe ayant reçu les encouragements plus susceptible de s'inscrire au programme.

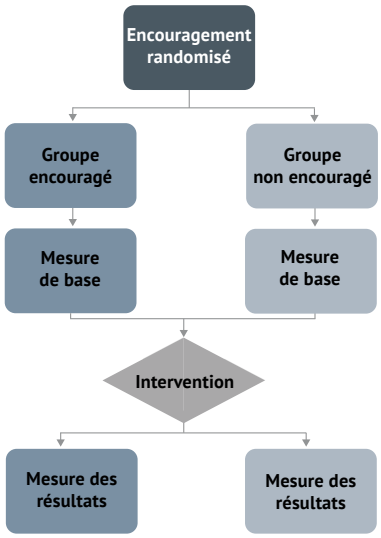
Même si les personnes qui bénéficieront de l'intervention ne sont pas choisies de manière aléatoire, l'existence d'une source de variation aléatoire intrinsèque – l'encouragement – permet d'utiliser une approche de VI pour extraire une estimation non biaisée des impacts. Dans les cas où

[3] Ce méthode est également connu comme répartition aléatoire en aval

une personne bénéficie de l'intervention par hasard, mais qu'il y a ensuite attrition ou non-conformité, une VI peut également être utilisée si la répartition elle-même est vue comme une façon d'encourager la participation à l'intervention.

Il est important de mentionner ici que les résultats d'un modèle d'encouragement randomisé sont moins généralisables que ceux d'un ECR traditionnel puisqu'ils s'appliquent uniquement aux individus qui ont modifié leur comportement après avoir été encouragés à participer (groupe conformiste). Toutefois, les modèles d'encouragement sont souvent utiles pour la mesure des impacts étant donné qu'ils sont plus étroitement liés aux éléments sur lesquels les organismes gouvernementaux exercent un contrôle. Souvent, les ministères/organismes gouvernementaux peuvent uniquement encourager des individus à faire quelque chose plutôt que les y obliger.

Figure 8



Supposons qu'une décideuse politique souhaite comprendre les impacts du tutorat parascolaire sur le niveau de scolarité; elle pourrait alors offrir du tutorat au groupe de traitement et au groupe témoin afin de mesurer les résultats. Toutefois, étant donné les enjeux éthiques que cela pourrait soulever, une solution pourrait être d'offrir le tutorat à tous les élèves admissibles, mais encourager de façon aléatoire les parents à inscrire leurs enfants. Cette technique est similaire à celle utilisée dans une étude bien connue (Bogatz & Ball, 1971) qui visait à évaluer les impacts de l'émission Sesame Street sur les résultats scolaires des jeunes enfants. Dans le cadre de cette étude, un groupe de parents était encouragé à laisser les enfants regarder l'émission, et cet encouragement était utilisé comme « instrument » pour estimer l'effet de causalité de Sesame Street sur les résultats scolaires, comparativement au groupe qui n'avait pas reçu d'encouragements.

QUAND PEUT-ON UTILISER LES APPROCHES PAR VARIABLES INSTRUMENTALES?

Les approches par VI peuvent être utilisées :

- Lorsqu'il y a présence d'une source de variation, comme une loterie ou une exposition « apparemment » aléatoire à un programme;
- Dans un contexte de programme ouvert à tous, où l'inscription est volontaire. Cela suppose qu'aucun participant ne peut se voir refuser l'inscription, et que les participants s'inscrivent de leur propre gré. Ce type de programme permet l'utilisation de modèles d'encouragement.

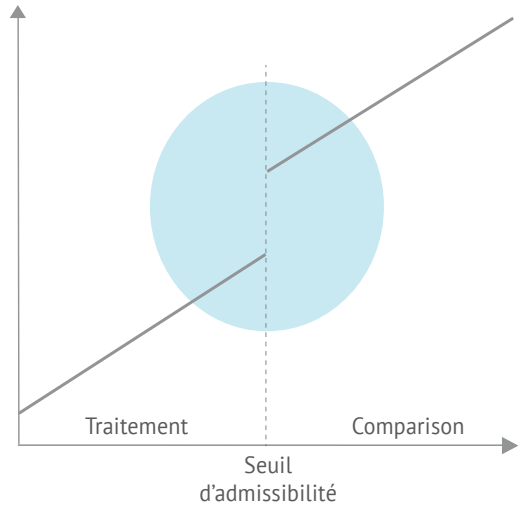
POINTS À CONSIDÉRER

- La restriction en matière d'exclusion doit être valable.
- L'instrument doit être pertinent. Un instrument faible mènerait à une estimation très instable et amplifierait le biais découlant de la violation de la restriction en matière d'exclusion.
- (Dans le cas d'un modèle d'encouragement) Les individus qui tirent parti des encouragements ne doivent pas être évalués différemment du reste de la population.

MODÈLE DE DISCONTINUITÉ DE LA RÉGRESSION

Le modèle de discontinuité de la régression (MDR) s'applique aux programmes qui déterminent l'admissibilité des participants à l'aide de certains critères minimaux (quantifiables). Certains exemples pourraient comprendre le revenu familial, l'âge ou la moyenne pondérée cumulative, où les individus qui se situent au-dessus ou en dessous d'une certaine valeur sont ciblés par le traitement. Dans les faits, les programmes et les services sociaux utilisent souvent ce modèle pour établir l'admissibilité des participants, donc les occasions d'utiliser cette méthode pour comprendre les impacts ne manquent pas. Certains intervenants estiment même que le MDR est une méthode largement sous-utilisée (Moscoe, Bor, & Barnighausen, 2015; Shadish, Cook, & Campbell, 2002, p. 208).

Figure 9



Prenons pour exemple un programme de soutien parascolaire destiné aux enfants qui éprouvent des difficultés d'apprentissage, où l'admissibilité au programme est établie à partir des plus récents résultats en mathématiques des élèves. Supposons que le programme de soutien est offert aux élèves qui ont obtenu une note inférieure à 40 %. Dans les faits, les élèves qui ont obtenu un résultat proche de ce seuil (c.-à-d. une note entre 37 % et 43 %) présentent des caractéristiques très similaires. Toutefois, en raison de quelques points retranchés de façon arbitraire, les élèves qui ont obtenu une note entre 37 % et 39 % recevront le traitement, contrairement à ceux qui ont des capacités très similaires, mais dont le résultat se situe entre 40 % et 43 %.

Le MDR se fonde donc sur l'hypothèse voulant que le groupe « tout juste inadmissible » et le groupe « tout juste admissible » soient effectivement comparables puisque le léger écart dans leur admissibilité respective est probablement attribuable à des variations aléatoires de circonstances. Dans le scénario hypothétique ci-dessus, l'écart entre les élèves qui ont obtenu 39 %, 40 % ou 41 % le jour de l'examen est probablement une question de chance, notamment parce que certains élèves se sentaient bien ce jour-là ou qu'ils ont réussi à deviner certaines réponses.

En effet, nous avons un MDR qui se situe autour du seuil d'admissibilité, ce qui signifie que nous pouvons tenir pour acquis que toutes les caractéristiques (observables et non observables) sont équilibrées dans le groupe de traitement et le groupe témoin, et par conséquent que tout écart dans les résultats observés peut être attribué au programme/traitement lui-même. Si les hypothèses établies à l'aide du MDR sont valables, l'estimation des impacts aura alors une validité interne élevée. C'est pour cette raison que le MDR est une méthode quasi expérimentale de niveau supérieur.

QUAND PEUT-ON UTILISER LES MODÈLES DE DISCONTINUITÉ DE LA RÉGRESSION?

Les modèles de discontinuité de la régression sont utiles dans le cas des programmes :

- Qui établissent l'admissibilité à partir de critères quantifiables, comme les résultats scolaires, l'âge ou une quelconque note;
- Dont les critères d'admissibilité ne sont liés à aucun autre élément pouvant expliquer les résultats;
- Qui comprennent suffisamment de « cas » (participants) répondant aux critères pour permettre une analyse (généralement, pour cette raison, les échantillons sont plus grands dans le cas des expérimentations

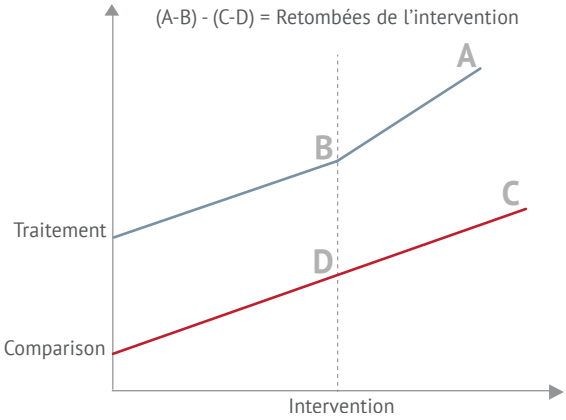
POINTS À CONSIDÉRER

- Si d'autres éléments changent soudainement au point de seuil, le MDR ne fera pas de distinction entre les effets de ces changements et les effets du programme. Par exemple, un MDR utilisé pour évaluer les effets d'un crédit d'impôt sur le bien-être financier de personnes de 65 ans et plus pourrait s'avérer problématique, par exemple si les gens ont tendance à prendre leur retraite à cet âge ou qu'ils deviennent admissibles à d'autres avantages réservés aux aînés à cet âge.
- Le choix de la « fourchette » associée au point de seuil est important, c'est-à-dire qu'il doit y avoir suffisamment de cas des deux côtés du seuil choisi pour assurer une analyse statistique significative. Toutefois, les fourchettes trop larges peuvent empêcher de comparer adéquatement les deux groupes de chaque côté.
- Le MDR peut fournir des estimations efficaces et un haut degré de validité interne, mais étant donné qu'il n'analyse qu'une petite partie des cas se trouvant à proximité du seuil, il est plus limité au chapitre de la validité externe. Dans l'exemple du programme d'aide en mathématiques ci-dessus, l'estimation des impacts établie à partir du MDR est moins susceptible de s'appliquer aux élèves de la classe dont les résultats sont très faibles (par exemple, ceux qui ont obtenu une note inférieure à 30 %).

ÉCART ENTRE LES DIFFÉRENCES

La méthode de l'écart entre les différences est une méthode qui se fonde sur des données de type avant-après recueillies à propos d'un groupe de traitement et d'un groupe de comparaison afin de comparer les changements dans les résultats respectifs. La méthode de l'écart entre les ***différences*** mise sur ***l'hypothèse des tendances parallèles***, c'est-à-dire l'hypothèse selon laquelle la tendance dans les résultats du groupe de comparaison représente bien ce que serait la tendance dans les résultats du groupe de traitement en l'absence de l'intervention (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2016, p. 134).

Figure 10



Si l'hypothèse des tendances parallèles est valable, la méthode d'écart entre les différences permet alors d'estimer un effet de causalité avec un haut degré de validité interne pour l'ensemble des caractéristiques observables et non observables. La validité de l'hypothèse des tendances parallèles peut en fait uniquement être testée à l'aide d'une analyse des tendances historiques dans les résultats du groupe de traitement et du groupe de comparaison. Si le groupe de traitement et le groupe de comparaison avaient tendance à présenter des résultats très similaires pendant une longue période avant la mise en œuvre du programme, cela représente une preuve à l'appui de l'hypothèse des tendances parallèles. Plus nous avons de données historiques, plus nous pouvons obtenir de preuves pour appuyer ou réfuter cette hypothèse. Toutefois, nous ne pouvons pas écarter la possibilité que l'hypothèse des tendances parallèles ne s'applique pas même si les tendances historiques sont quasi identiques. Par exemple, une nouvelle politique inattendue pourrait être mise en œuvre et changer subitement la tendance observée dans les résultats du groupe de comparaison.

La méthode d'écart entre les différences suit les étapes ci-dessous :

1. Trouver un groupe de comparaison qui présente des tendances historiques similaires à l'égard de la variable des résultats.
2. Observer les résultats du groupe de traitement et du groupe de comparaison, avant et après le programme.
3. Calculer l'écart dans les résultats de chaque groupe au fil du temps. La première différence s'observe à cet endroit.
4. Calculer l'écart entre les deux écarts calculés à l'étape deux. Il s'agit de la deuxième différence, et de l'estimation des impacts du programme.

QUAND PEUT-ON UTILISER LE MODÈLE D'ÉCART ENTRE LES DIFFÉRENCES?

Le modèle d'écart entre les différences peut être utilisé dans les cas où :

- La répartition aléatoire est impossible;
- Un groupe de comparaison peut être identifié et lorsque les données sur les tendances historiques confirment que ses résultats ont évolué au même rythme que ceux du groupe de traitement;
- Il est raisonnable de tenir pour acquis que l'« hypothèse des tendances parallèles » est valable.

MÉTHODES QUASI EXPÉRIMENTALES DE NIVEAU INFÉRIEUR

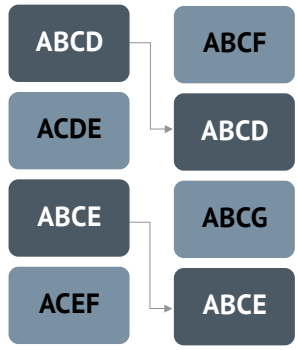
La série de méthodes ci-dessous exige que l'analyste tienne pour acquis que des résultats contrefactuels valides peuvent être établis uniquement à partir de caractéristiques observables. De par leur nature, ces méthodes ne permettent pas de tenir compte des écarts entre les éventuelles caractéristiques non observables du groupe de traitement et du groupe de comparaison. Pour cette raison, elles sont considérées comme moins efficaces que les méthodes décrites ci-dessus.

APPARIEMENT

Les techniques d'appariement sont des façons de former un groupe de comparaison à partir de caractéristiques observables. À l'aide des caractéristiques observables qui se dégagent des données, on recherche à la base (avant le début du programme) à trouver pour chaque participant un « équivalent » parmi les non-participants afin de former un groupe de comparaison le plus similaire possible au groupe de traitement.
Cet « équivalent » doit présenter des valeurs aussi similaires que possible pour une série de variables de confusion, comme l'âge, le sexe, le revenu et le niveau d'éducation. Sur le plan intuitif, un « équivalent » est une « copie » de l'unité traitée qui ressemble en tous points à celle-ci, sauf qu'il ne participe pas au traitement. S'il est possible de créer un groupe de comparaison équivalent, l'écart dans les niveaux de résultats entre le groupe de traitement et le groupe de comparaison peut alors fournir une bonne estimation des effets du programme. La théorie sous-jacente suggère qu'étant donné que les bons appariements présentent des caractéristiques de base similaires, ils devraient présenter des niveaux de résultats semblables, et ces résultats peuvent par conséquent être utilisés comme base contrefactuelle.

Une des principales difficultés de la méthode d'appariement réside dans le fait que plus le nombre de variables que nous tentons de jumeler est élevé, plus il devient difficile de trouver un équivalent proche dans le groupe de comparaison. Il en résulte donc un problème d'appui commun – c'est-à-dire que le groupe de traitement et le groupe de comparaison se situent dans la même fourchette de valeurs pour toutes les variables de contrôle pertinentes. D'un autre côté, il sera impossible de trouver un équivalent adéquat pour certaines unités puisqu'aucune unité n'est suffisamment similaire dans le groupe opposé. L'image à droite illustre ce problème : si chaque lettre représente une caractéristique, il peut être difficile de trouver un équivalent exact dans le groupe de comparaison puisque le nombre de caractéristiques à associer augmente.

Figure 11



En outre, l'hypothèse la plus importante pour la validité de l'estimation de l'appariement est que l'écart entre le groupe de traitement et le groupe de comparaison est observable (c.-à-d. que le programme ne devrait pas comprendre de sources de sélection non observables). C'est une hypothèse très difficile à confirmer. En pratique, il peut être difficile de savoir exactement quelles sont ces variables, et les évaluateurs sont souvent limités par les conclusions des études précédentes ainsi que par les données existantes.

Dans ce contexte, l'appariement figure parmi les méthodes quasi expérimentales de niveau inférieur puisqu'il est uniquement réalisable à partir de caractéristiques observables. L'appariement repose grandement sur une hypothèse avancée, et il présente un risque que d'importantes différences non observables ne soient pas prises en compte entre le groupe de traitement et le groupe témoin alors que ces différences peuvent contribuer à des résultats indépendants de l'intervention ou du programme.

APPARIEMENT DES SCORES LIÉS À LA PROPENSION

Comme il a été indiqué plus tôt, l'appariement fondé sur un grand nombre de caractéristiques de base est souhaitable, toutefois cela peut entraîner ce qu'on appelle un ***problème dimensionnel***. À mesure que la liste de variables à apparier s'allonge, il devient plus difficile de trouver un équivalent adéquat pour chaque unité. L'appariement peut rapidement devenir impossible dans le cas des grands ensembles de données pour lesquels il y a de nombreuses variables de contrôle.

«l'appariement fondé sur un grand nombre de caractéristiques de base est souhaitable, toutefois cela peut entraîner ce qu'on appelle un problème dimensionnel. À mesure que la liste de variables à apparier s'allonge, il devient plus difficile de trouver un équivalent adéquat pour chaque unité.»

La méthode d'appariement par score de propension est conçue pour contourner ce problème. Plutôt que de fonder l'appariement sur les caractéristiques mêmes, on peut adopter une méthode d'appariement par score de propension, laquelle mise sur l'estimation de la propension d'un individu (vraisemblance ou probabilité) à s'inscrire à partir d'une série de caractéristiques observables qui prédisent l'inscription, attribue un score sommaire entre zéro et un, puis associe ces cas aux cas du groupe non participant qui présentent des scores identiques ou quasi identiques. Ce groupe « équivalent » constitue le groupe de comparaison. En pratique, cela permet d'établir une comparaison entre les participants et les non-participants, qui avaient une propension similaire à participer au programme, mais qui n'y ont pas participé pour une quelconque raison.

Essentiellement, l'appariement par score de propension consiste à tenter de reproduire les avantages uniques de la randomisation. Lorsque nous répartissons aléatoirement les participants dans deux groupes, nous nous assurons qu'ils ont des probabilités égales de se retrouver dans le groupe de traitement ou le groupe de comparaison – c'est ce qui crée un équilibre. En faisant un appariement par score de propension, nous faisons quelque chose de similaire : nous créons deux groupes qui ont des probabilités similaires de se retrouver dans le groupe de traitement ou le groupe de comparaison. Il s'agit-là d'une autre façon – bien qu'imparfaite – de créer un équilibre entre les deux groupes afin de les rendre comparables.

QUAND PEUT-ON UTILISER LES TECHNIQUES D'APPARIEMENT?

Les techniques d'appariement sont polyvalentes et peuvent s'appliquer dans différents contextes où la répartition aléatoire est impossible. Elles conviennent dans les cas où :

- Des données administratives existent au sujet des participants et des non-participants, ou pourraient être créées;
- Les facteurs qui prédisent la participation au programme (pour l'appariement par score de propension) sont bien connus;
- On peut raisonnablement tenir pour acquis qu'aucune caractéristique non observable ne prédit la participation ou, à tout le moins, que l'effet des caractéristiques non observables est minime.

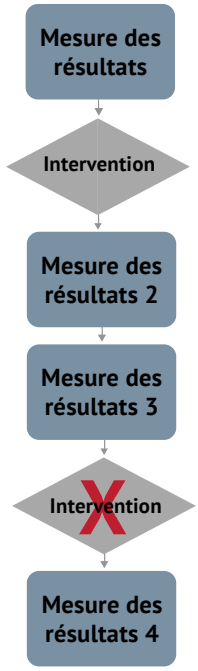
POINTS À CONSIDÉRER

- L'appariement exige de grands ensembles de données pour accroître la probabilité que de bonnes équivalences soient établies pour les participants au programme.
- Une mesure d'écart doit être définie pour déterminer le degré de « proximité » entre deux observations, étant donné que celles-ci varient en fonction de plusieurs facteurs (âges différents, revenus différents et ainsi de suite).
- Dans l'ensemble, les sous-échantillons du groupe de traitement et du groupe de comparaison doivent être comparables (c.-à-d. qu'il serait problématique que tous les individus du groupe de traitement aient un faible revenu et que tous les individus du groupe de comparaison aient un revenu élevé. Dans ce cas, aucune équivalence ne pourrait être établie.)
- Toutes les variables entraînant un biais doivent être recensées et incluses dans l'algorithme d'appariement. Des arguments doivent être élaborés pour appuyer le fait qu'aucune variable pertinente pour les résultats n'a été omise.

MODÈLES DE TRAITEMENT SUPPRIMÉ/ INTERROMPU

L'objectif principal des modèles de traitement supprimé/interrompu n'est pas d'observer un groupe témoin qui n'est jamais soumis au traitement, mais plutôt de recueillir des données sur le groupe de traitement au fil de plusieurs périodes, où le traitement commence et s'interrompt à plusieurs reprises. L'évolution des résultats au cours de la période où le traitement est interrompu sert alors de base contrefactuelle pour l'intervention. Si les valeurs propres aux résultats reproduisent ce modèle cyclique, nous avons des motifs raisonnables de croire que le traitement a un effet de causalité. Seules les autres variables qui suivent également ce modèle ont un potentiel de confusion; nous pouvons tenter de les contrôler par l'intermédiaire d'une régression. Cette méthode, de par sa nature, ne s'applique pas aux interventions ponctuelles, comme les projets d'infrastructure.

Figure 12



QUAND PEUT-ON UTILISER LES MODÈLES DE TRAITEMENT SUPPRIMÉ/INTERROMPU?

Ces modèles conviennent dans les cas suivants :

- Il n'était pas possible de former un groupe de comparaison.
- Les résultats peuvent être mesurés à de multiples reprises avant et après l'intervention.
- Le type d'intervention se prête bien à l'interruption (p. ex., une intervention mettant en cause des règles ou des frais plutôt qu'une intervention sur le plan scolaire pouvant avoir des effets à long terme).
- L'intervention a des effets relativement immédiats plutôt que des effets à retardement.

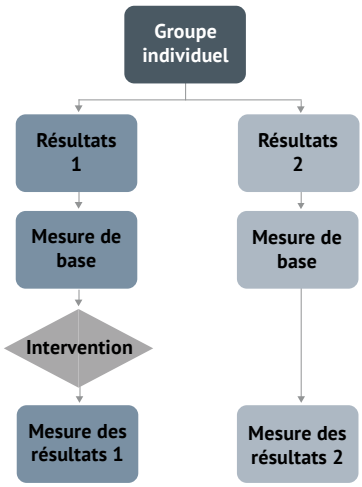
POINTS À CONSIDÉRER

- L'intervention doit avoir commencé et avoir été interrompue à plusieurs reprises.
- La variable des résultats doit réagir assez rapidement à l'intervention. Si les effets de l'intervention se manifestent à retardement, il devient plus difficile d'isoler les périodes visées par l'intervention des périodes non visées.

VARIABLES DÉPENDANTES NON ÉQUIVALENTES

Une autre façon de compenser l'absence d'un groupe témoin consiste à trouver une autre variable qui évolue de façon similaire à la variable des résultats d'intérêt, mais qui n'est pas affectée par le traitement. Les tendances touchant cette variable peuvent alors être utilisées comme base contrefactuelle.

Figure 13



Reichardt (Reichardt, 2011) donne l'exemple hypothétique d'une émission éducative pour enfants. L'émission enseigne les lettres de l'alphabet aux enfants d'âge préscolaire, à raison d'une lettre par semaine. Après la treizième semaine, une fois que les enfants ont appris la première, mais pas la seconde moitié de l'alphabet, leurs connaissances sont mises à l'épreuve pour l'ensemble des vingt-six lettres. Dans ce scénario, les véritables impacts de l'émission télévisée sont l'écart entre la connaissance qu'ont les enfants des treize premières lettres et leur connaissance des treize dernières lettres, parce que nous pouvons tenir pour acquis qu'en l'absence du programme l'évolution des connaissances liées aux treize premières serait naturellement la même que l'évolution des connaissances liées aux treize dernières lettres.

S'assurer que la variable de contrôle n'est pas affectée par le traitement est une exigence difficile à respecter qui ne peut pas être vérifiée au-delà du raisonnement théorique. En outre, la variable de contrôle doit être mesurée sur la même échelle que la variable de traitement pour fournir une base contrefactuelle adéquate.

Dans les versions améliorées de ce modèle, il est question de choisir plusieurs variables de contrôle qui seront fort probablement affectées par le traitement à différents degrés. L'evaluateur peut ensuite vérifier si les changements subséquents touchant les variables de résultats et les variables de contrôle s'inscrivent dans les tendances prévues. Dans la mesure du possible, de telles approches devraient s'inspirer d'autres études, notamment en ce qui concerne les estimations rigoureuses relatives aux incidences causales du traitement sur les variables de contrôle.

QUAND PEUT-ON UTILISER LES VARIABLES DÉPENDANTES NON ÉQUIVALENTES?

Les variables dépendantes non équivalentes peuvent être intégrées à n'importe quel modèle pour en augmenter l'inférence causale, dans la mesure où :

- une variable s'applique aux mêmes éléments que ceux du résultat d'intérêt, à l'exception du traitement lui-même.

POINTS À CONSIDÉRER

- Il faut qu'il y ait une variable sur laquelle l'intervention n'aura aucun effet, et qui sera évaluée sur la même base que la variable des résultats.

ÉTUDES CAS-TÉMOINS

Les études cas-témoins tirent leur origine du domaine de l'épidémiologie. Une façon simple de comprendre la logique des études cas-témoins est de les voir comme une « rétrospective », ou comme des études qui fonctionnent de façon contraire aux expériences précédemment décrites. N'oublions pas que l'expérience consiste à répartir aléatoirement les individus dans deux groupes, à exposer un des groupes à une intervention, puis à observer l'écart dans les résultats entre les deux groupes afin de déterminer si l'intervention s'est avérée efficace ou non. Par contraste, dans le cas d'une étude cas-témoin, nous travaillons à l'envers en observant le résultat avant de déterminer la cause (probable). Ces études sont généralement utilisées lorsque le résultat d'intérêt est « binaire » (il est soit présent soit absent, comme dans le cas d'une maladie en particulier ou du chômage).

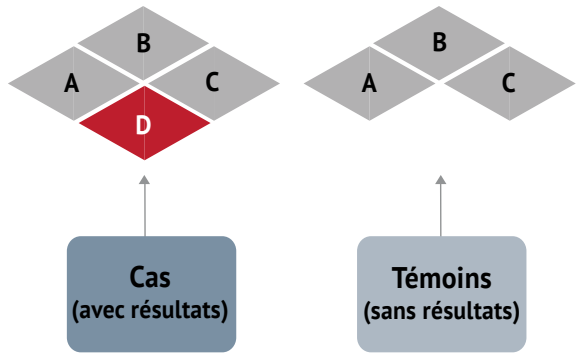
«... dans le cas d'une étude cas-témoin, nous travaillons à l'envers en observant le résultat avant de déterminer la cause (probable).»

Ces méthodes sont efficaces dans deux contextes. Tout d'abord, lorsque la répartition aléatoire est contraire à l'éthique – comme lorsqu'on cherche à comprendre les effets du tabac sur le cancer. Aucune expérimentation de ce type n'existe. La majorité des études qui établissent ce lien se fondent sur des études cas-témoins, où l'on compare des individus qui présentent le résultat à d'autres qui ne le présentent pas (dans ce cas-ci, le cancer en tant que résultat négatif) et où l'on essaie d'établir les différences entre les deux groupes pouvant expliquer le résultat (dans ce cas-ci, l'exposition à la fumée de tabac).

L'autre contexte où cette méthode peut s'avérer utile est lorsque le résultat qui nous intéresse est rare, ce qui rend la répartition aléatoire impossible puisque dans une expérience, nous risquerions d'affecter au groupe témoin les individus les plus susceptibles d'atteindre le résultat, ce qui rendrait nos efforts inutiles.

Les études cas-témoins se situent relativement bas sur l'échelle des preuves d'Impact Canada, mais représentent un choix pratique afin de générer des preuves initiales pour justifier la tenue d'une évaluation plus rigoureuse.

Figure 14



QUAND PEUT-ON UTILISER LES ÉTUDES CAS-TÉMOINS?

Les études cas-témoins peuvent être utilisées :

- Lorsque les résultats d'intérêt sont rares (p. ex., développement d'une maladie rare, programmes d'entraînement d'athlètes de haut niveau);
- Lorsqu'une théorie plausible explique pourquoi de tels résultats sont observés, ce qui permet de comparer les résultats de l'exposition pour les cas et les témoins.

MÉTHODES QUANTITATIVES NON EXPÉRIMENTALES (EXPLORATOIRES)

Les méthodes non expérimentales se caractérisent par le fait qu'on ne peut pas déterminer qui a été visé par le programme ou l'intervention et par le fait qu'elles ne tentent pas de reproduire une expérience. C'est dans le deuxième contexte qu'elles se distinguent grandement des méthodes quasi expérimentales. À cet égard, les méthodes non expérimentales ne tiennent pas compte de l'importance des résultats contrefactuels. Par conséquent, étant donné que ces méthodes ne font pas une utilisation rigoureuse des groupes témoins ou des comparisons, elles présentent de faibles degrés de validité interne. Elles sont toutefois très efficaces en tant que méthodes exploratoires pour justifier la nécessité d'obtenir ou de fournir du financement pour réaliser une évaluation plus rigoureuse à l'aide d'ECR ou de méthodes quasi expérimentales.

ÉCART DANS LES MOYENNES

La façon la plus simple d'estimer les effets d'un traitement consiste à mesurer l'écart dans les moyennes de la variable des résultats entre le groupe de traitement et le groupe témoin. Cette méthode est simple, parce qu'il n'est pas nécessaire de tenir compte de l'information sur les autres caractéristiques des unités dans l'échantillon. Par contre, elle est très susceptible de comporter des biais, comme des biais de sélection et des effets de causalité inverse. Pour cette raison, l'estimation de l'écart dans les moyennes est souvent qualifiée d'« estimation naïve » étant donné qu'elle tient pour acquis que le groupe de traitement et le groupe témoin sont équivalents au chapitre des caractéristiques de base. Il faut noter que c'est uniquement lorsque le groupe de traitement et le groupe témoin présentent des caractéristiques de base en moyenne identiques (p. ex., dans un ECR parfait) que l'estimation naïve représente une méthode de mesure des impacts valide.

COMPARAISON AVANT-APRÈS

Les études avant-après sont un modèle d'évaluation courant. Elles mesurent les impacts à partir des écarts dans les résultats moyens du groupe de traitement avant et après le programme. Autrement dit, elles tiennent pour acquis que les impacts du programme s'observent dans l'évolution des résultats des participants au fil du temps. Cette méthode est extrêmement problématique étant donné que le changement avant-après dans les résultats sera attribuable non seulement au programme, mais à l'effet de tous les autres éléments qui ont affecté le groupe traité au cours de la même période. Autrement dit, cette méthode comporte une lacune en raison du biais lié au décalage de temps (effets historiques). Tout changement positif dans les résultats au fil du temps peut être attribuable à une multitude d'autres raisons non liées au programme lui-même : les participants ont vieilli, ils ont acquis plus d'expérience, la situation économique locale a changé, etc. Dans le cadre de cette méthode, ces autres facteurs sont impossibles à isoler (comme il est mentionné plus tôt, un groupe témoin valide est nécessaire pour les isoler). Par conséquent, l'estimation des résultats avant-après ne constitue pas une preuve acceptable de l'effet de causalité d'une intervention ou d'un programme.

ANALYSE COMPARATIVE À L'AIDE DE DONNÉES AGRÉGÉES

Dans les cas où il n'y a pas de groupe témoin valide, les évaluateurs utilisent parfois l'analyse comparative pour établir une projection des résultats contrefactuels. Par

exemple, si nous connaissons la proportion moyenne de personnes d'un groupe traité travaillant après un programme de formation professionnelle, nous pourrions utiliser le niveau d'emploi moyen à l'échelle nationale pour une cohorte similaire comme référence pour un groupe témoin. Ces analyses comparatives peuvent se baser sur les moyennes nationales ou régionales pour les résultats en question. Cela permet de calculer l'effet « traitement moins référence ». Ces méthodes ont généralement un faible degré de validité interne et n'indiquent pas la causalité puisque l'échantillon traité pourrait présenter des caractéristiques différentes de l'échantillon de référence. L'ajustement de ces écarts (c.-à-d. contrôler les écarts entre le groupe de traitement et le groupe de personnes sur lequel se fonde l'analyse comparative) peut aider à accroître la rigueur de l'approche. Cet ajustement est uniquement possible si des données individuelles figurent dans les données de référence (p. ex., pour le programme de formation professionnelle, nous avons des données sur les individus de la région qui nous communiquent leur situation professionnelle et des données de base telles que l'âge, le sexe, le niveau d'éducation et l'état de santé). En règle générale, l'analyse comparative devrait uniquement être utilisée pour fournir une estimation initiale très sommaire au moment d'explorer les résultats d'une intervention.

MÉTHODES QUALITATIVES

Les méthodes qualitatives jouent un rôle important dans la compréhension des processus à l'origine des impacts. Même si les méthodes quantitatives décrites dans le présent document peuvent donner lieu à des estimations présentant un haut degré de validité interne, elles comportent de par leur nature un certain degré d'inconnu. Elles peuvent permettre d'établir un lien entre un changement dans les résultats et un programme, toutefois ces méthodes n'en disent souvent pas assez sur la raison pour laquelle le programme a eu les impacts observées. Les méthodes qualitatives peuvent fournir des perspectives sur le déroulement du programme et l'expérience des participants, des éléments utiles pour la conception et l'amélioration des programmes. Par exemple, si une évaluation quantitative rigoureuse a permis d'établir qu'un programme de formation professionnelle n'a eu aucun effet positif sur l'accès des participants au marché du travail, une étude qualitative pourrait aider à en comprendre la raison (p. ex., les participants pourraient s'être sentis intimidés par le formateur ou avoir eu l'impression que le matériel ne répondait pas à leurs besoins). Ce sont là des renseignements additionnels très pertinents pour les analystes de politiques et les concepteurs de programmes et d'interventions politiques.

En ce qui concerne la validité interne, se fier uniquement aux approches qualitatives est problématique pour un certain nombre de raisons. Tout d'abord, elles ont tendance à se fonder sur de plus petits échantillons, de sorte qu'il est impossible de mener des vérifications statistiques des données et qu'il est difficile de généraliser les conclusions. Ensuite, les conclusions/résultats eux-mêmes sont souvent touchés par différents types de biais. Plusieurs raisons expliquent ce fait, notamment :

- Les preuves démontrent que lorsqu'on leur pose une question sur les impacts au fil du temps, les gens (participants à un programme ou experts tiers) auront tendance à utiliser une comparaison de type avant-après pour évaluer l'efficacité d'un programme. Ils évalueront donc leurs résultats avant le programme et attribueront au programme tous les changements à cet égard, sans tenir compte des éléments contrefactuels. Par conséquent, les méthodes qualitatives ont tendance à présenter les mêmes problèmes que les études avant-après et mèneront généralement à une surestimation des impacts.

- En outre, le biais de désirabilité sociale – le fait que les participants à l'étude veulent « plaire » aux administrateurs du programme en offrant des réactions positives – mènera à une surestimation des impacts. Lorsque l'administrateur d'un programme leur demande si le programme a eu un effet, les gens peuvent difficilement dire que non. De façon similaire, la dissonance cognitive peut jouer un rôle. Cette théorie de la science comportementale suggère que lorsqu'il y a contradiction entre une croyance et un comportement, les gens auront tendance à modifier leurs croyances pour ne pas donner l'impression d'incohérence.

En règle générale, en ce qui concerne les impacts, les études qualitatives ne devraient pas servir d'unique méthodologie. Combinées aux études quantitatives, les études qualitatives ont un rôle à jouer dans la compréhension des processus et des expériences des participants à un programme afin d'appuyer l'amélioration de l'élaboration et de l'application des politiques.

PRINCIPE DE VALIDITÉ EXTERNE

Les sections précédentes se sont majoritairement concentrées sur la validité interne des modèles de recherche. Une fois qu'un bon degré de validité interne a été confirmé (à l'aide de l'Échelle des preuves d'Impact Canada), la validité externe devient un élément important pour les applications et les modèles futurs.

La validité externe est la mesure dans laquelle les conclusions d'une étude peuvent être généralisées à d'autres cas, régions, périodes et individus. Les effets d'une intervention politique varieront souvent en fonction de l'environnement dans lequel la politique est appliquée, de la population cible, du moment où l'intervention a lieu et des détails relatifs à la mise en œuvre, et d'autres facteurs propres au contexte. Une évaluation rigoureuse devrait permettre de démontrer que les effets estimés du traitement sont transférables dans d'autres contextes et à l'échelle de sous-populations définies par différents niveaux de facteurs préexistants (Cook & Campbell, 1979).

« La validité externe est la mesure dans laquelle les conclusions d'une étude peuvent être généralisées à d'autres cas, régions, périodes et individus. »

Les menaces à la validité externe peuvent s'observer autant sur le plan de la collecte des données que du modèle de recherche, de même que sur le plan économétrique. Sur le plan de la collecte des données, les principales menaces potentielles sont le manque de généralisabilité entre les différents cas et le manque de généralisabilité entre les individus. La première menace s'observe lorsque le contexte ou les détails de l'expérience poussent les participants à adopter un comportement autre que celui qu'ils auraient adopté dans une situation réelle. Par exemple, une expérience en laboratoire où les participants manipulent de l'argent fictif et prennent des décisions de consommation lorsqu'il n'y a pas de véritables biens et services entraînera une distorsion des motivations. Il est fort probable que les participants ne prendront pas les mêmes décisions que dans une situation réelle. Par ailleurs, il peut y avoir un manque de généralisabilité entre les individus si l'échantillon n'est pas représentatif. Par exemple, les répondants à un sondage en ligne sont plus susceptibles d'être des jeunes qui utilisent beaucoup les technologies et qui vivent dans des régions urbaines. Leur réaction au programme pourrait être différente de celle du reste de la population.

Les résultats d'une étude (considérée comme ayant un haut degré de validité interne) devraient uniquement être extrapolés à d'autres cas présentant un contexte similaire (p. ex., caractéristiques régionales similaires, participants similaires, programmes similaires sur les plans du contenu et de la durée).

Une question d'ordre plus technique touchant la validité externe s'observe dans le concept d'effets hétérogènes d'un traitement, c'est-à-dire que des individus différents au sein d'une population réagissent différemment à une intervention. Les retombées moyennes sur la population pourraient être différentes des retombées moyennes sur les individus qui choisissent de participer au programme puisque les gens qui se sont inscrits s'attendent à retirer des avantages du programme. Ce concept est très proche, bien que différent, du biais de sélection, c'est-à-dire que deux individus peuvent obtenir le même résultat contrefactuel sans l'intervention, mais qu'un des deux peut en retirer plus d'avantages que l'autre. En raison de la répartition aléatoire, les méthodes expérimentales (ECR) annulent les retombées de la sélection sur les gains, et l'estimation des retombées

qui en découle est une mesure généralisable qui présente un haut degré de validité externe puisqu'elle représente les retombées auxquelles on pourrait s'attendre en moyenne pour l'échantillon visé par l'étude.

Toutefois, les résultats établis à l'aide de certaines méthodes sont encore plus complexes et difficiles à interpréter étant donné qu'ils tiennent uniquement compte des retombées sur une partie de l'échantillon visé par l'étude. Par exemple et comme il est mentionné plus tôt, les modèles d'encouragement estiment uniquement les retombées pour la sous-population qui bénéficie des encouragements. Si la conformité n'est pas aléatoire, les retombées moyennes sur l'ensemble de la population seront différentes des retombées moyennes pour la sous-population conformiste. Les estimations établies à l'aide d'un MDR tiennent également compte des retombées à l'échelle locale, cette fois-ci pour la sous-population proche du point de seuil, qui pourrait ne pas être représentative de l'ensemble de la population cible.

Certaines méthodes mathématiques peuvent tenter de neutraliser certaines menaces en matière de validité externe, moyennant le respect de certaines conditions (Pearl & Bareinboim, 2014). Les poids d'échantillonnage sont des exemples particuliers de ces méthodes, où chaque unité de l'échantillon est pondérée par le ratio du groupe de population que représente l'unité et la taille que représente le groupe en question dans l'échantillon. Cela facilite l'extrapolation à la population générale des résultats de l'échantillon visé par l'étude. Enfin, un argument universellement applicable pour appuyer la validité externe des résultats d'une étude est lorsque les résultats ont été reproduits dans d'autres études à partir de contextes ou d'échantillons différents. À ce titre, il est important de constituer un ensemble de preuves de grande envergure à partir d'études réalisées à l'aide de méthodologies rigoureuses dans une grande variété de contextes. Cet élément est essentiel pour mener des examens approfondis de la documentation relative à des programmes et des interventions similaires au cours de l'évaluation des retombées, comme précédemment mentionné.

PRÉPARER ET MENER UNE ÉTUDE DE MESURE DES IMPACTS

Nous décrivons ici une procédure et une série d'étapes simples à suivre par les analystes qui mènent des évaluations des impacts.

1. Déterminer le programme et le groupe cible.
2. Mener un examen de la documentation propre aux évaluations de programmes similaires.
3. Choisir la méthode présentant le niveau le plus élevé possible dans l'échelle des preuves (en fonction de l'applicabilité et des contraintes liées aux données).
4. Établir des hypothèses à l'aide de l'approche et s'assurer qu'elles sont valables.
5. Mener l'évaluation, si possible à l'aide de méthodes quantitatives et qualitatives.
6. Préciser les résultats et formuler des mises en garde, s'il y a lieu. Discuter de la rigueur (validité interne) des résultats et les comparer aux conclusions énoncées dans la documentation. Remarque : cette étape peut être effectuée en partenariat avec des experts en évaluation du gouvernement du Canada ou des experts en évaluation et des organismes de recherche.
7. inalement, discuter de la généralisabilité (validité externe) des résultats.

ANNEXE A – SOMMAIRE DES MÉTHODES

Méthode	Description	Hypothèses	Avantages	Inconvénients
Essai contrôlé randomisé (ECR)/ Expérimentation	Répartition d'un échantillon de façon aléatoire dans un groupe de traitement et un groupe témoin pour éviter tout biais de sélection	Répartition aléatoire des personnes qui bénéficient du traitement; conformité parfaite; aucun biais de l'observateur; taille de l'échantillon suffisante	La meilleure approche pratique pour générer une estimation non biaisée de l'effet de causalité	Peut coûter cher; souvent impossible à réaliser
Estimation à l'aide de variables instrumentales (VI) et modèle d'encouragement	Utilisation d'une variation aléatoire d'une autre variable qui augmente la probabilité qu'une personne bénéficie de l'intervention	L'instrument n'influe pas sur le résultat directement, mais uniquement par l'intermédiaire de l'intervention; l'instrument est pertinent pour l'intervention	Contrôle des biais observables et non observables de façon similaire à un ECR, à condition que la méthode soit menée de façon adéquate	Les instruments valides peuvent être difficiles à trouver; les encouragements peuvent être inefficaces; la population conformiste peut ne pas être représentative
Modèle de discontinuité de la régression (MDR)	L'intervention est déterminée par un seuil, comparaison des observations tout juste au-dessus ou tout juste en dessous du seuil	Aucun élément ne change de façon abrupte au seuil, sauf que certaines personnes bénéficient de l'intervention	Contrôle des biais non observables	Exige un contexte précis (où un seuil détermine qui bénéficiera d'une intervention); peut ne pas être généralisable à l'ensemble de la population
Estimation des écarts entre les différences	Comparaison avant-après des résultats du groupe de traitement et des résultats du groupe comparison	Les résultats des deux groupes auraient évolué en parallèle sans l'intervention	Facile à comprendre de façon conceptuelle et permet de contrôler certaines caractéristiques non observables	Exige de suivre les mêmes individus au fil du temps
Appariement	Appariement d'une unité traitée et d'une unité comparison qui présentent des caractéristiques de base similaires	Tous les facteurs pertinents pour le résultat ont été recensés et observés; le groupe de traitement et le groupe comparison sont comparables à l'égard de ces facteurs	S'applique à tout type de relation entre les variables de résultats et les variables de contrôle (la relation ne doit pas nécessairement être linéaire)	Exige de grands ensembles de données; ne permet pas de tenir compte des biais pour les variables omises

Méthode	Description	Hypothèses	Avantages	Inconvénients
Modèle de traitement supprimé	Le traitement est administré et interrompu à plusieurs reprises; observation des changements dans la variable des résultats	L'administration ou l'interruption du traitement influe immédiatement sur le résultat	Une approche créative pour compenser l'absence d'un groupe témoin	Uniquement applicable à des cas très précis
Variables dépendantes non équivalentes	Compare les changements dans la variable des résultats aux changements touchant d'autres variables similaires	La variable utilisée pour le contrôle n'est pas affectée par le traitement, mais elle est évaluée sur la même base que la variable des résultats	Une approche créative pour compenser l'absence d'un groupe témoin	Uniquement applicable à des cas très précis
Études cas-témoins	Sépare les unités de l'échantillon qui ont obtenu des résultats de celles qui n'en ont pas obtenu, et compare leur exposition au traitement	Tous les facteurs pertinents pour le résultat ont été identifiés et observés	Permet de surmonter les problèmes liés aux petits échantillons lorsque les résultats sont rares	Ne permet pas d'établir les impacts causales de l'estimation, mais seulement un rapport des probabilités
Écart dans les moyennes	Comparaison des résultats moyens pour le groupe traité et le groupe comparison	Le traitement doit être administré de façon aléatoire (comme dans un ECR), autrement l'estimation sera faussée par un biais de sélection	Exige très peu de données et se calcule rapidement; peut être utilisée comme méthode d'analyse exploratoire	Ne permet pas de tenir compte des biais attribuables à la sélection ou à la causalité inverse
Comparaison avant-après	Comparaison des résultats moyens pour le groupe traité avant et après l'intervention	Rien à l'exception de l'intervention n'a changé entre les deux mesures – ce modèle ne s'applique jamais en pratique	Exige très peu de données et se calcule rapidement; peut être utilisée comme méthode d'analyse exploratoire	Ne permet pas de tenir compte des biais attribuables aux effets du temps/effets historiques
Analyse comparative	Compare les résultats moyens pour le groupe traité avec une moyenne régionale ou nationale	Le groupe traité présente les mêmes caractéristiques de base pertinentes que la moyenne nationale	Exige très peu de données et se calcule rapidement; peut être utilisée comme méthode d'analyse exploratoire	Le groupe traité est susceptible d'être différent de la moyenne nationale/régionale, et cet élément ne peut pas être corrigé

ANNEXE B – ILLUSTRATION MATHÉMATIQUE DE LA MESURE DES IMPACTS

La v**ariable des résultats** est ici définie par Y, et la **variable de traitement** par D. Notre principal objectif est de trouver l'effet de causalité de D sur Y. Dans le cas le plus simple, D est une variable binaire qui prend la valeur de 1 si l'unité est traitée et de 0 si elle ne l'est pas. Les unités qui ont la valeur D_i=1 forment le **groupe de traitement**, tandis que celles qui ont la valeur D_i=0 forment le **groupe témoin**. Pour chaque unité i observée, nous définissons les valeurs (hypothétiques) Y_{1i} and Y_{0i} pour illustrer le **résultat potentiel du groupe traité** et le **résultat potentiel du groupe non traité**, respectivement. Ces deux valeurs renvoient exactement à la même unité au même point dans le temps et dans le même contexte, sauf en ce qui concerne le traitement. L'écart restant, Y_{1i} - Y_{0i} , est l'effet de causalité du traitement, ou l'**effet du traitement**.

Désignons ensuite Y_i comme la valeur du résultat réellement observé pour i. Il s'ensuit que si i est traité, Y_i=Y_{1i} et Y_{1i} ne sont pas observés. Inversement, si i n'est pas traité, Y_i = Y_{0i} and Y_{1i} ne sont pas observés. L'effet moyen du traitement sur la population est $E(Y_{1i} - Y_{0i})$, alors que l'écart naïf entre les résultats du groupe traité et ceux du groupe non traité est $E(Y_i|D_i=1) - E(Y_i|D_i=0)$. Il peut être déconstruit de la façon suivante :

$$
\begin{aligned}
E(Y_i|D_i=1) - E(Y_i|D_i=0) &= E(Y_{1i}|D_i=1) - E(Y_{0i}|D_i=0) \\
&= E(Y_{1i}|D_i=1) - E(Y_{0i}|D_i=1) + E(Y_{0i}|D_i=1) - E(Y_{0i}|D_i=0) \\
&= E(Y_{1i} - Y_{0i}|D_i=1) + E(Y_{0i}|D_i=1) - E(Y_{0i}|D_i=0)
\end{aligned}
$$

Effets du traitement sur le groupe traité

Biais de sélection

Comme nous le verrons plus en détail dans cette section, pour que le terme de droite indiquant un biais de sélection disparaisse et pour que l'estimation naïve soit égale à l'effet (de causalité) du traitement $E(Y_{0i}|D_i=1)= E(Y_{0i}|D_i=0)]$, il faut que le traitement soit administré indépendamment des résultats potentiels.

BIBLIOGRAPHIE

Angrist, J., & Krueger, A. (1991). Does compulsary school attendance affect schooling and earnings? The Quarterly Journal of Economics, 106(4), 979-1014.

Bogatz, G., & Ball, S. (1971). *The Second Year of Sesame Street: A Continuing.* Princeton, NJ: Educational Testing Service.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297-312.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings.* Chicago: Rand McNally College Publishing Company.

Dunning, T. (2013). *Natural Experiments in the Social Sciences: A Design-Based Approach.* New York, New York: Cambridge University Press.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2016). *Impact Evaluation in Practice - Second Edition.* Washington, D.C.: Inter-American Development Bank and World Bank. Retrieved from http://www.worldbank.org/en/programs/sief-trust-fund/publication/impact-evaluation-in-practice

Jamison, J. C. (May 2017). *The Entry of Randomized Assignment into the Social Sciences.* Policy Research Working Paper, World Bank Group, Development Policy Department.

Moscoe, E., Bor, J., & Barnighausen, T. (2015). Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. *Journal of Clinical Epidemiology, 68*, 132-143.

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4), 579-595.

Reichardt, C. S. (2011). Evaluating Methods for Estimating Program Effects. *American Journal of Evaluation, 32*, 246-272.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Belmont, CA: Wadsworth Cengage Learning.

REMERCIEMENTS

L'Unité de l'impact et de l'innovation souhaite souligner les contributions des personnes ayant travaillé à la production de *Modèles de mesure des impacts*. Les coauteurs sont les suivants :

Craig M. Joyce
conseiller principal (UII); coauteur et chef de projet

Daniel Fujiwara
directeur (Simetrica); coauteur et réviseur

Iulian Gramatki
économètre (Simetrica); coauteur et réviseur

Les auteurs aimeraient souligner l'apport des collègues de l'UII pour leurs révisions et leurs commentaires relativement aux diverses versions, notamment Elizabeth Hardy, directrice principale, Étude des comportements; David Donovan, directeur, Politiques stratégiques et financement novateur; Julie Greene, directrice, Capacités et partenariats; Victoria Carlan, directrice, Mesure des répercussions; Haris Khan, conseiller, Étude des comportements; Alyssa Whalen, conseillère, Étude des comportements; ainsi que les fellows de l'UII Amanda Desnoyers, Meera Paleja, Lauren Conway, et Guillaume Beaulac.

Les auteurs sont également reconnaissants des contributions du Isabelle Agier, Analyste, Agence de la santé publique du Canada, ainsi que celles du group de travail technique sur l'évaluation de l'incidence de l'UII, dont les membres ont offert une excellente rétroaction, notamment Sonia Ben Amor, gestionnaire en évaluation, Statistique Canada; Anne-Renee Blaise, scientifique de la défense, ministère de la Défense nationale; Greg Bridgett, conseiller principal, Secrétariat du Conseil du Trésor; Geneviève Boudrias, agente principale de l'évaluation, Commission canadienne de sûreté nucléaire; Kristina Guiguet, analyse en politiques, Emploi et Développement social Canada; Francis Jobin, analyste principal, Agence de promotion économique du Canada atlantique; Chantal Langevin, directrice, Santé Canada.

Et, enfin, Laurie Bennett, agente de communications multimédias, a produit avec doigté la production graphique finale.

impact canada

/impact_innovfr

impact-and-innovation-unit

iiu - uii

impact.canada.ca