WHAT WORKS
In Youth Employment

Organisation internationale du Travail

# Mesurer l'emploi décent des jeunes

## Un guide sur le suivi, l'évaluation et les leçons des programmes du marché du travail

Note 5

# Méthodes d'évaluation d'impact des interventions sur l'emploi des jeunes

Mesurer l'emploi décent des jeunes

Un guide sur le suivi, l'évaluation et les leçons des programmes du marché du travail

# NOTE 5
# Méthodes d'évaluation d'impact des interventions sur l'emploi des jeunes

Première édition 2018



---

*Mesurer l'emploi décent des jeunes.*
*Un guide sur le suivi, l'évaluation et les leçons des programmes du marché du travail.*
*Note 5 : Méthodes d'évaluation d'impact des interventions sur l'emploi des jeunes /*
Bureau international du Travail. Genève, 2019.

ISBN : 978-92-2-133471-2 (imprimé)
978-92-2-133472-9 (Web pdf)

Également disponible en arabe: دليل قياس الوظائف اللائقة للشباب.الرصد والتقييم والتعلّم في برامج سوق العمل النشطة - نظرة عامة
ISBN 978-92-2-630801-5 (imprimé), 978-92-2-630802-2 (Web pdf), Genève, 2018. Et an anglais:
ISBN 978-92-2-131670-1 (print), 978-92-2-131671-8 (web pdf), Genève 2018

---



---

Conception et mise en page par le Centre international de formation de l'OIT, Turin Italie

Imprimé en Suisse

# Table des matières

# Méthodes d'évaluation d'impact des interventions sur l'emploi des jeunes

**Conditions préalables :**

Des connaissances de base sur les méthodes de recherche quantitative. Cette note décrit les principales méthodes d'évaluation d'impact et explique les avantages et limites de chacune, en tenant compte des considérations théoriques et pratiques.

**Objectifs d'apprentissage :**

À la fin de cette note, les lecteurs seront en mesure de :

- Comprendre les considérations et défis majeurs à prendre en compte dans la détermination de l'impact en se demandant : "Que serait-il arrivé aux mêmes personnes/foyers/communautés si l'intervention n'avait pas eu lieu ?"
- Construire une situation contrefactuelle pour évaluer quels changements des résultats peuvent être attribués à une intervention, et identifier les caractéristiques clés que les groupes de traitement et de comparaison doivent partager pour respecter la validité interne
- Peser les pour et les contre des différentes techniques d'évaluation et comment elles cherchent à éliminer les biais de sélection
- Bien comprendre différentes méthodes de recherche quantitative, des méthodes expérimentales (entièrement randomisées) à celles quasi-expérimentales, tels que le modèle de double différences *(DID)*, d'appariement, et le modèle de régression par discontinuité
- Utiliser les méthodes qualitatives pour savoir "ce qui s'est passé" – en déterminant l'effet du traitement moyen de l'intervention – mais aussi "pourquoi".

**Mots-clés:**

Attribution, comparaison avant et après, groupe témoin, contrefactuel, double différences, validité externe, validité interne, tirage aléatoire, traçage du processus, appariement par score de propension, essais controlés randomisés, modèle de promotion aléatoire, modèle de régression par discontinuité, groupes de traitement.

Cette note[1] donne aux praticiens une vue d'ensemble des différents outils disponibles pour l'évaluation d'impact et propose des orientations sur ceux qui doivent être sélectionnés dans des circonstances spécifiques, et comment les mettre en œuvre pour pouvoir évaluer les effets des interventions sur l'emploi des jeunes. Alors que les évaluations d'impact peuvent se fonder sur les méthodes qualitatives et quantitatives, cette note est essentiellement axée sur les méthodes quantitatives et introduit les méthodes qualitatives comme important complément dans le contexte de l'approche des méthodes mixtes.

## Le défi de l'attribution

Tout d'abord, il est nécessaire de clarifier ce que nous entendons par *impact*. Dans les notes précédentes de ce guide, nous avons utilisé le terme comme synonyme des objectifs ou résultats de haut niveau relatifs au changement de la situation de l'emploi des jeunes, tels que la réduction du chômage ou l'amélioration du bien-être des individus et des ménages. Cependant, dans le contexte des évaluations d'impact, nous réduisons la définition de l'*impact* au changement des résultats (ex: statut de l'emploi, temps de travail, revenus) pouvant être attribué à notre intervention.

Comme indiqué dans la Note 4, les évaluations d'impact tentent de répondre aux questions de causalité, c'est-à-dire, à savoir si l'intervention (la cause) a amélioré le résultat parmi les bénéficiaires (l'effet). Par exemple:

- Le changement observé sur la probabilité des stagiaires d'obtenir un emploi peut-il être attribué à notre intervention sur la formation professionnelle?
- Notre intervention en conseil d'emploi mène-t-elle à une augmentation du degré de satisfaction des employeurs, et à un taux plus élevé du maintien de l'effectif?
- Notre intervention sur l'encadrement des start-ups favorise-t-elle la création et la durabilité d'entreprises?

Les résultats du marché du travail qui nous intéressent sont déterminés par plusieurs facteurs complexes tels que le contexte global du développement économique et social, les changements politiques et/ou personnels, etc. Ainsi, c'est un véritable défi d'établir le degré selon lequel les changements de tels résultats peuvent être attribués à une intervention particulière. L'objectif de l'évaluation d'impact est précisément de surmonter cette difficulté d'**attribution** en mesurant le degré auquel un programme particulier, *et seulement ce programme*, contribue au changement des résultats concernés.

**DÉFINITION**

**Attribution: L'imputation d'un lien causal entre des changements observés (ou que l'on attend à être observés) et une intervention spécifique.**

[1] La note de ce guide se fonde sur une étude originellement développée par Duflo et al. (2006), Khandker et al. (2010), et Gertler et al. (2016), tout en adaptant certains des éléments et illustrations au domaine de l'emploi des jeunes et en fournissant une présentation plus concise des méthodes d'évaluation d'impact.

En d'autres termes, les évaluations d'impact tentent d'estimer dans quelle mesure et pourquoi les changements observés sur les résultats concernés peuvent être attribués à une intervention ou un projet.

Le point focal de cette note est ce qu'on appelle le cadre **contrefactuel** sur lequel les évaluations d'impact quantitatives sont typiquement basées.[2] Cette approche définit l'impact d'une intervention comme étant la différence entre les résultats observés sous l'intervention et le scénario dit contrefactuel : "Que serait-il arrivé aux mêmes personnes/ménages/communautés si l'intervention n'avait pas eu lieu ?". La Figure 5.1 illustre les concepts d'impact et de contrefactuel.

En pratique, le vrai contrefactuel est impossible à mesurer. Les méthodes d'évaluation d'impact tentent de quantifier les effets de causalité en estimant ou en construisant souvent le contrefactuel – mais pas toujours – par comparaison avec les groupes témoins, parfois appelés *groupes de contrôle*. Le groupe de participants est appelé le *groupe de traitement ou groupe de participants*.

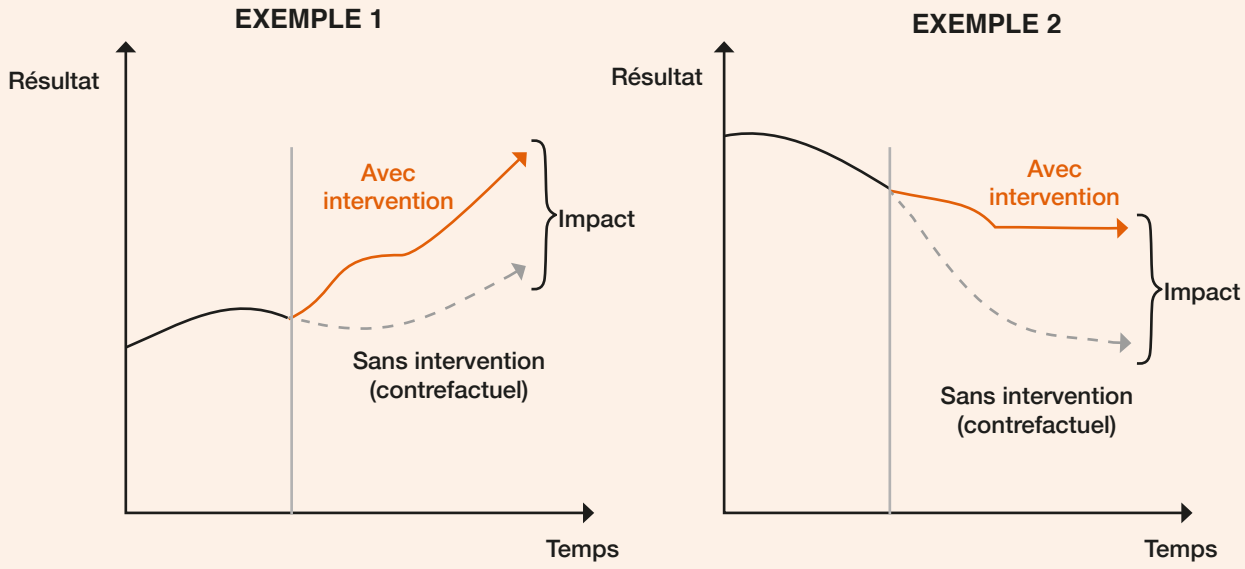
Les **groupes de traitement** et de **comparaison** devraient avoir les mêmes caractéristiques à au moins trois niveaux (Gertler et al., 2016):

[2] Selon l'intervention en cours d'évaluation et son contexte, la conception d'une évaluation d'impact qui mixe des méthodes qualitatives et quantitatives est généralement plus approprié, comme expliqué plus en détail ci-dessous.

**DÉFINITION**

**Contrefactuel:** Le contrefactuel décrit ce qu'un certain résultat aurait été pour un participant à un programme en l'absence du programme. Par définition, le contrefactuel ne peut pas être directement observé. En conséquence, il doit être estimé, par exemple, en recourant à des groupes de comparaison.

FIGURE 5.1 : ILLUSTRATION DE L'IMPACT D'UNE INTERVENTION



1. Ils devraient être similaires en termes de caractéristiques observables et non-observables : Les **Caractéristiques observables** peuvent inclure l'âge, le genre, le niveau d'éducation, le statut socio-économique, les caractéristiques familiales, le statut de l'emploi, etc. Les **Caractéristiques non-observables** peuvent inclure la motivation, l'intérêt, les valeurs et idéologies, et le niveau de soutien familial, parmi d'autres facteurs. Toutes les personnes du groupe de traitement ne doivent pas nécessairement être identiques à toutes celles du groupe de comparaison, mais les deux groupes doivent en moyenne avoir les mêmes caractéristiques.
2. Les groupes de traitement et de comparaison devraient réagir à l'intervention de manière similaire : par exemple, les résultats tels que les compétences ou le revenu, devraient être susceptibles d'augmenter pour les membres du groupe de traitement comme pour celui du groupe de comparaison.
3. Les groupes de traitement et de comparaison devraient avoir des niveaux similaires d'exposition aux autres interventions : par exemple, les deux groupes devraient avoir le même accès aux services de soutien fournis par le gouvernement local, les ONG, etc.
4. Quand les groupes de traitement et de comparaisons ont les mêmes similarités citées ci-dessus, on peut raisonnablement en déduire que toute différence observée sur les résultats des deux groupes peuvent être attribués à l'intervention. D'autre part, si le groupe de comparaison est substantiellement différent du groupe de traitement, la comparaison des résultats entre les groupes de traitement et de comparaison reflètera non seulement l'impact de l'intervention, mais aussi les conséquences de ces différences. C'est ce que l'on appelle un biais de sélection.

**Le biais de sélection** se produit généralement lorsque les participants et non-participants à l'intervention diffèrent en caractéristiques non-observées, qui affectent la probabilité des individus à participer à (et/ou terminer) l'intervention ainsi que les résultats d'intérêt.

Dans la plupart des programmes d'emploi des jeunes, il est vraisemblable que ceux qui font une demande de participation diffèrent considérablement de ceux qui ne font pas une telle demande, et que ces différences ne puissent pas être facilement observées. Par exemple, les participants à un projet de conseil en emploi pourraient être plus motivés et avoir accès à une meilleure information sur les moyens de trouver un emploi que les non-participants, même avant le début de l'intervention. Dans un tel cas, les participants pourraient avoir un plus grand taux de réussite en termes de résultats sur le marché du travail mais on ne saurait pas vraiment si c'est dû à l'intervention ou à leur avantage initial relatives aux conditions de départ.

Un des objectifs clés des techniques d'évaluation présentées ici est d'éliminer le biais de sélection. Lorsqu'il n'y a pas de

**DEFINITION**

**Un groupe de comparaison est un groupe qui sert à estimer le contrefactuel dans une évaluation d'impact. Contrairement aux membres d'un groupe de traitement, les membres du groupe de comparaison n'ont pas été exposés à l'intervention que l'on veut évaluer. Les termes "groupe de comparaison" et "groupe de contrôle" sont souvent utilisés indistinctement. Dans ce document, nous utiliserons le terme générique groupe de comparaison.**

**DEFINITION**

**Groupe de traitement : le groupe de personnes qui participent activement à l'intervention est appelé le groupe de traitement ou groupe participant.**

biais de sélection, les différences observées sur les résultats entre groupes de traitement et de comparaison peuvent être attribuées à l'intervention.

Tout comme les contrefactuels, les biais de sélection ne sont pas observables. À moins de faire très attention à la sélection de groupes de traitement et de comparaison, une simple comparaison des résultats du marché du travail inclura aussi bien l'impact de l'intervention que le biais de sélection. En général, il est impossible de savoir dans quelle mesure ils sont causés par l'un ou par l'autre.

Les techniques d'évaluation présentées ci-dessous ont pour objectif de permettre la sélection de groupes de traitement et de comparaison en élimant le biais de sélection afin que les comparaisons entre groupes de traitement et de comparaison ne reflètent que l'impact de l'intervention.

Un bon groupe de comparaison est essentiel pour la **validité interne** de l'évaluation qui détermine la fiabilité et la crédibilité de ses résultats. La validité externe entre en jeu lorsque l'on considère la transférabilité de ces résultats : l'intervention devrait-elle être élargie à d'autres communautés ou mise en œuvre au niveau régional ou national ? Peut-on s'attendre à des résultats similaires si l'on conçoit ce programme pour d'autres contextes ou une population cible différente ? Ce sont là des questions qui sont généralement d'intérêt pour les décideurs politiques.

Il est important de garder à l'esprit que les conditions ne seront jamais exactement les mêmes lorsque l'on reproduit ou intensifie une intervention. En conséquence, pour obtenir une validité externe, il est crucial de comprendre les aspects complexes du programme dans le contexte, l'endroit et le moment spécifiques de sa mise en œuvre, et leur influence possible sur les résultats de l'évaluation. Par exemple, les services de l'emploi pour jeunes diplômés pourraient être effectifs dans une

**DEFINITION**

**Caractéristiques observables et non-observables :** Les caractéristiques observables peuvent être mesurées par des méthodes de collecte de données appropriées (telles que les enquêtes). Elles incluent souvent l'âge, le genre, le niveau d'éducation, le statut socio-économique, les caractéristiques familiales, le statut de l'emploi, etc. Les caractéristiques non-observables sont ces facteurs qui ne peuvent pas être, ou ne sont pas, mesurés dans une évaluation (d'impact) et pourraient inclure la motivation, l'intérêt, les valeurs et idéologies, et le niveau de soutien familial. Pour plusieurs de ces caractéristiques non-observables, des mesures indirectes (imparfaites) ont été développées.

**DEFINITION**

**Biais de sélection :** Le biais de sélection se produit lorsque les raisons de la participation d'un individu au programme sont corrélées aux résultats. Ce biais se produit souvent lorsque le groupe de comparaison se retire de lui-même du programme (par exemple, les abandons).

**DEFINITION**

**Validité interne :** pour avoir une validité interne, une évaluation d'impact doit avoir un groupe de comparaison qui fournit une estimation valide du contrefactuel. Une évaluation d'impact validée à l'interne sera capable d'attribuer clairement les changements des résultats de l'intervention en contrôlant toutes les différences possibles entre les groupes de traitement et de comparaison. Ceci peut être accompli en appliquant des techniques expérimentales ou quasi-expérimentales appropriées.

région puisqu'ils correspondent à une demande de la part de l'économie locale de ce groupe cible, car les institutions d'éducation ont une bonne réputation, ou du fait que le programme ait été mis en œuvre pendant une saison à forte demande de travail des employeurs. Mettre en œuvre la même intervention dans d'autres régions et pendant toute l'année pourrait s'avérer bien différent.

Pour comprendre non seulement *si* quelque chose marche, mais également *pourquoi et dans quel contexte* elle est censée marcher, il est nécessaire d'analyser les mécanismes de causalité qui sous-tendent les résultats observés. Les méthodes qualitatives, par exemple, celles appliquées dans le contexte de l'évaluation fondée sur la théorie, sont d'une importance fondamentale pour ce travail.

Au fil de cette note, diverses méthodes quantitatives d'évaluation d'impact seront introduites, suivies par un exemple de méthode qualitative et de remarques sur comment une approche de méthodes mixtes peut contribuer à accomplir une validité interne et externe des résultats de l'évaluation.

**DEFINITION**

**Validité externe : Dans l'évaluation d'impact, la validité externe signifie que l'impact causal observé peut être généralisé à tous les individus éligibles. Donc, pour qu'une évaluation soit valide en externe, il est nécessaire que l'échantillon de l'évaluation soit un échantillon représentatif de tous les individus éligibles.**

## Encadré 5.1: Appui de l'OIT à l'évaluation d'impact

Le Département de l'Évaluation de l'OIT (alias "EVAL") a développé plusieurs ressources pour soutenir l'évaluation d'impact (IE, *impact evaluation*) :

- Un *Cadre d'Evaluation de l'Impact*: EVAL a développé un papier indiquant comment, quand et pourquoi les IE devraient être considérées et mises en œuvre, à travers des avis du personnel de l'OIT. Le document traite de questions centrales telles que l'usage et l'objectif spécifique de l'IE ; l'appariement entre les questions de recherche de l'évaluation et la méthodologie appropriée ; l'usage d'une gamme de méthodologies complémentaires et disponibles ; la faisabilité et la valeur des IE ; et le besoin de non seulement identifier l'impact (le quoi) mais aussi le comment et pourquoi.
- Une *Facilité de Revue des Évaluations d'Impact (IERF, Impact Evaluation Review Facility)*: EVAL a établi un mécanisme de revue qui permet au personnel de l'OIT de poser des questions et demander des revues de documents de réflexion, de propositions complètes, de plans et rapports pour soutenir la planification, la conception ou la mise en œuvre des IE (EVAL_impact@ilo.org). Une Note d'Information sur le fonctionnement ce service est disponible.
- Un *inventaire d'évaluations d'impact menées à l'OIT*: L'inventaire facilite l'accès au savoir institutionnel dans plusieurs zones d'interventions.
- *Une expertise de qualité des évaluations d'impact de l'OIT*: Afin de pouvoir faire le suivi et établir un rapport sur l'avancée de l'OIT sur son utilisation et la qualité des IE, EVAL commandera périodiquement une expertise-qualité des IE dans toute l'organisation.
- Une *communauté de pratique* formant un *Réseau Informel d'Evaluation d'Impact*: Ce groupe informel de collègues impliqués et intéressés par les IE se réunit régulièrement pour partager ses expériences et, au besoin, fournir une revue par les pairs.

Le but de ces ressources est d'aider l'OIT à améliorer sa capacité dans le domaine de l'utilisation des IE en documentant ce qui marche et pour qui, ainsi qu'en estimant l'impact.

# Défis spécifiques à l'évaluation des programmes actifs du marché du travail axés sur les jeunes[3]

La nature des programmes actifs du marché du travail (PAMT), particulièrement ceux qui ciblent la population jeune, affecte plusieurs aspects de la conception d'une évaluation valide. En tant que toile de fond de la discussion plus détaillée des questions de conception de la Note 6, cette section présente une vue d'ensemble d'une partie des caractéristiques les plus communes des PAMT axés sur les jeunes et décrit quelques traits de la conception d'évaluation qui se révèlent particulièrement pertinents pour ce type de programme. Comprendre lesquels de ces traits sont susceptibles d'être présents dans un environnement donné aidera à formuler le modèle d'évaluation approprié.

## Programmes obligatoires ou volontaires

Une des caractéristiques fondamentales d'une intervention sur l'emploi des jeunes est la nature obligatoire ou volontaire du programme. Les programmes obligatoires sont incorporés dans plusieurs services d'emploi publics, y compris ceux liés aux assurances-chômage et aux programmes de formation. Dans ces contextes, les jeunes doivent participer à un PAMT qui est rattaché à une allocation de chômage. Ceci dit, et comme nous le verrons dans les sections suivantes, la participation obligatoire à une intervention pour l'emploi des jeunes pose certains défis à l'évaluation d'impact, où des estimations d'impact valides nécessitent typiquement un groupe de traitement et un groupe équivalent de comparaison. Certaines méthodes d'évaluation d'impact peuvent seulement être appliquées aux programmes volontaires qui recrutent des participants à partir d'un plus grand groupe de candidats qui peuvent décider s'ils veulent participer ou non.

## Non-conformité : Défecteurs et décrocheurs

Dans beaucoup de programmes volontaires d'intervention sur l'emploi des jeunes, une fraction considérable de personnes assignées au programme ne s'enregistrera pas au programme (les défecteurs) ou abandonnera avant d'avoir complété le programme (les décrocheurs). Ce défi est particulièrement pertinent chez les jeunes qui sont très mobiles, qui ont tendance à changer fréquemment d'adresse et de lieu de travail et qui alternent entre le travail et les études.

En effet, Card et al. (2011) déclarent que :

> Il est rare d'atteindre un taux de réussite d'un programme au-dessus de 80 pour cent, des taux aussi bas que 50 pour cent sont courants et le manque d'anticipation des problèmes causés par les défections et les décrochages est une des causes majeures du modèle inadéquat des évaluations de PAMT (2011, p. 13).

Bien que le non-respect des membres du groupe du programme ou du groupe témoin n'invalide pas un modèle d'évaluation en soi, il complique l'interprétation des résultats, et implique que l'évaluation soit obligée de collecter des données sur les taux actuels de participation du programme par le groupe de traitement et le groupe de comparaison.

3 Cette section est basée sur Card et al. (2011).

La validité d'un modèle randomisé repose de façon cruciale sur l'équivalence entre les résultats observés du groupe de comparaison et les résultats contrefactuels du groupe de traitement. Dans la plupart des cas, cette équivalence est compromise lorsque les membres d'un groupe ou de l'autre abandonnent. Pour cette raison, l'analyse d'un modèle randomisé devrait être fondée sur la comparaison des groupes de traitement et de comparaison tels qu'ils ont été initialement affectés, utilisant les données de toutes les personnes initialement assignées à ces groupes. Dans la littérature de l'évaluation expérimentale, cet aspect est appelé une "analyse de l'intention de traiter".

## Recrutement et filtrage

Puisque seulement quelques-uns des jeunes recrutés pour l'évaluation d'impact sont effectivement assignés au programme, l'apport à une évaluation peut entraver le flux normal de clients d'un programme en cours.[4] Ceci n'est pas particulièrement un souci dans les contextes où il y a beaucoup plus de candidats que de places disponibles : dans ces cas-là, la sélection aléatoire est un dispositif objectif et pratique de sélection. Cependant, dans les situations où le flux normal de recrues est nécessaire pour remplir les places du programme en cours, les opérateurs du programme pourraient objecter d'avoir certains de leurs potentiels clients affectés au groupe de comparaison et pourraient tenter d'outrepasser le processus d'affectation. C'est extrêmement important de savoir à l'avance si une telle situation est susceptible de se produire. Si tel est le cas, la planification de l'évaluation pourrait nécessiter un budget pour augmenter les efforts de recrutements et accroître le flux de nouveaux clients, ainsi que des ressources supplémentaires pour faire un suivi rapproché de la conformité aux protocoles de recrutement. Par exemple, les PAMT pour jeunes pourraient être limités aux hommes et femmes sans emploi entre 16 et 30 ans. Normalement, les mêmes procédés et règles de filtrage d'éligibilité devraient être employés pour sélectionner les participants à l'évaluation.

## Taille des échantillons

Les orientations sur les tailles d'échantillons nécessaires à une évaluation d'ALMP sont fondées sur un calcul de puissance normalisé. L'ingrédient principal pour un tel calcul est une estimation de la taille du programme pour l'effet plausible (ex : l'effet du programme sur le résultat d'intérêt, exprimé en tant que fraction de l'écart type de ce résultat). Étant donné cette valeur, et les choix normalisés pour le niveau de signification statistique (ex : 5 pour cent) ainsi que l'adéquation de la puissance du modèle (ex : 0,80), le calcul de la taille appropriée des groupes de traitement et de comparaison d'un modèle randomisé à groupes de même taille est simple. Card et al. (2011) ont développé une orientation (voir tableau 5.1) qui précise quelle est la taille de l'échantillonnage requise pour mesurer une gamme d'impacts. Chaque ligne montre le taux d'emploi du groupe de comparaison, et chaque colonne représente la différence entre les groupes de traitement et de comparaison. Par exemple, si le taux d'emploi est de 50 pour cent dans le groupe de comparaison, pour détecter un impact significatif de 2,5 points de pourcentage dans l'emploi, la taille de l'échantillonnage requis est de 6.354 participants, avec le même nombre de non-participants.

[4] Par exemple, si 100 nouveaux clients se présentent sur le site du programme chaque mois, mais qu'il y a 80 places disponibles mensuellement, alors 40 personnes au plus peuvent recrutées pour l'évaluation : 20 seront affectées au programme (avec les 60 nouveaux clients qui ne font pas partie de l'évaluation) et 20 au groupe de comparaison.

**Tableau 5.1: Exemple de taille d'échantillon requise pour détecter les impacts significatifs**

| | | Impact du programme | | | | | |
|---|---|---|---|---|---|---|---|
| | | **2,5%** | **5,0%** | **7,5%** | **10,0%** | **12,5%** | **15,0%** |
| **Taux d'emploi du groupe de comparaison** | 30% | 5.475 | 1.417 | 650 | 376 | 247 | 176 |
| | 35% | 5.883 | 1.511 | 688 | 396 | 259 | 183 |
| | 40% | 6.166 | 1.574 | 713 | 408 | 265 | 186 |
| | 45% | 6.323 | 1.605 | 723 | 412 | 266 | 186 |
| | 50% | 6.354 | 1.605 | 719 | 408 | 262 | 183 |
| | 55% | 6.260 | 1.574 | 702 | 396 | 254 | 176 |
| | 60% | 6.040 | 1.511 | 671 | 376 | 240 | 165 |
| | 65% | 5.695 | 1.417 | 625 | 349 | 221 | 151 |

Source: Card et al. (2011)
* Sous les hypothèses standard (puissance = 0,8, significativité = 0,5, groupes de tailles égales, en utilisant la commande sampsi du programme Stata

Lorsque l'on considère la taille de l'effet considéré pour une PAMT, Card et al. (2011) recommandent de remettre ces programmes dans leur contexte. Ils déclarent :

> Un grand nombre de recherches ont démontré que, dans la plupart des pays à l'échelle mondiale, chaque année additionnelle de scolarisation formelle est associée à un gain de revenu d'environ 10 pour cent. On pourrait argumenter qu'une PAMT typique représente un plus petit investissement qu'une année typique de scolarisation formelle, donc, une taille d'effet de moins de 10 pour cent est raisonnable, et pour les programmes moins intensifs, des tailles d'effets de pas plus de 5 pour cent pourraient être plausibles (2011, p.19).

## Calendrier des enquêtes de suivi

Le calendrier d'une (ou plusieurs) enquête(s) de suivi est une décision importante en termes de garantie des impacts du programme. Plusieurs évaluations de PAMT mènent une enquête de suivi après un an, en partie parce que les termes du contrat d'évaluation nécessitent souvent un rapport final dans les deux à trois années. D'un autre côté, la littérature de PAMT existante suggère que l'impact des programmes plus intensifs, tels que les programmes de formation en salle de classe et formation professionnelle sur le tas, ne semblent se manifester que deux à trois ans après l'entrée en vigueur de la PAMT, plutôt que juste une année après (Card et al., 2011). Sur la base de ces études, et en tenant compte des effets d'interruption de beaucoup de PAMT, un horizon post-programme d'au moins deux ans est souhaitable pour les PAMT de plus longue durée.

Il y a toutefois un compromis à faire entre la capacité à observer des impacts à (plus) long terme et garantir la validité d'un modèle d'évaluation d'impact puisque les jeunes sont largement mobiles et peuvent se retrouver à l'autre bout du pays (ou même migrer) après avoir complété leurs programmes d'éducation ou de formation, il devient de plus en plus difficile au fil du temps de localiser un nombre suffisant de bénéficiaires de programme. Lorsqu'un grand nombre de jeunes ayant participé à l'enquête de référence d'une évaluation d'impact ne peut être contacté pour l'(es) enquête(s) de suivi (taux d'abandon élevé), il devient de plus en plus difficile, ou même impossible, de détecter et quantifier les impacts de façon fiable, dû à la puissance statistique réduite et au biais possible.

# Méthodes quantitatives de l'évaluation d'impact

Les méthodes quantitatives d'évaluation d'impact recommandées permettent d'obtenir une validation interne et évitent le biais de sélection en comparant les groupes avec ou sans traitement, lesquels ne sont idéalement différents que dans cette mesure. Le meilleur moyen d'accomplir cela est en détenant le contrôle sur qui reçoit l'intervention et qui ne la reçoit pas. Dans un tel cas, les modèles d'**évaluations expérimentales** sont possibles, et la plus courante de ces méthodes est l'**essai contrôlé aléatoire**. Si l'affectation au groupe de traitement et de comparaison est entièrement aléatoire, les deux groupes seront en moyenne très similaires avant le début du programme et nous serons ainsi dans la bonne direction pour assurer la validité interne.

Pour plusieurs raisons dont nous parlerons plus loin, la randomisation n'est pas toujours possible ou souhaitable. Dans ce cas, d'autres méthodes visant à entreprendre des comparaisons valides à l'interne peuvent être utilisées par la création d'un contrefactuel valide. On les appelle les **méthodes d'évaluation d'impact quasi-expérimentales**. Celles qui sont le plus utilisées sont les double différences (DID, *difference-in-differences*), l'appariement par score de propension (PSM, *propensity score matching*) et le modèle de régression par discontinuité (RDD, *regression discontinuity design*), que nous introduirons toutes brièvement dans cette section. Il est généralement fortement recommandé de considérer tous les efforts d'évaluation d'impact possibles et d'en peser avec attention les avantages et limites avant d'engager une évaluation. Le Tableau 5.2 fournit une vue d'ensemble sur diverses méthodes quantitatives d'évaluation d'impact.

FIGURE 5.2: CONSIDÉREZ TOUTES LES MÉTHODES POSSIBLES D'ÉVALUATION PENDANT LA PLANIFICATION


Source: www.freshspectrum.com

**DÉFINITION**

**Modèle expérimental :** les modèles expérimentaux utilisent la randomisation dans l'affectation des participants aux groupes de traitement et de comparaison. Ils peuvent produire des estimations d'impacts très crédibles mais sont souvent chers et, pour certaines interventions, difficiles à mettre en œuvre.

**Un essai contrôlé aléatoire** est une étude dans laquelle les personnes sont affectées au hasard (par pure chance) pour recevoir un traitement, tel que participer à une intervention spécifique.

**Modèle quasi-expérimental :** les méthodologies quasi-expérimentales sont utilisées pour construire un groupe de comparaison valide en utilisant des moyens statistiques pour le contrôle des différences entre les individus traités avec le programme à évaluer et ceux non-traités.

**Tableau 5.2 : Vue d'ensemble de diverses méthodes d'évaluation d'impact**

| Méthodologie | Description | Qui est dans le groupe de comparaison ? | Hypothèse requises | Données requises |
|---|---|---|---|---|
| Avant et après | Mesure comment les participants au programme s'améliorent dans le temps | Les participants au programme eux-mêmes – avant de participer au programme | Le programme était le seul facteur à influencer tout changement dans les résultats mesurés sur la période donnée | Données avant-et-après pour les participants du programme |
| Comparer les participants aux non-participants | Mesure la différence entre participants et non-participants au programme après la fin du programme | Les non-participants au programme (quelle qu'en soit la raison) mais dont les données ont été collectées après la fin du programme | Les non-participants sont identiques aux participants sauf en ce qui concerne la participation au programme | Données post-programme pour les participants et non-participants |
| Double différences | Mesure l'amélioration (changement) des participants au programme sur une période de temps relative à l'amélioration (changement) des non-participants | Les non-participants au programme (quelle qu'en soit la raison) mais dont les données ont été collectées avant et après la fin du programme | Si le programme n'existait pas, les deux groupes auraient eu des trajectoires identiques sur la période donnée (auraient partagé les mêmes tendances périodiques « communes ») | Données avant-et-après pour les participants et non-participants du programme |
| Appariement par score de propension | Personnes du groupe de traitement appariées aux non-participants ayant des caractéristiques observables similaires. La différence moyenne des résultats entre personnes appariées est l'impact estimé | Les non-participants ayant une combinaison de caractéristiques qui prédit qu'ils seraient aussi susceptibles de participer que les participants | Les facteurs ayant été exclus (car ils sont non-observables et/ou n'ont pas été mesurés) ne biaisent pas les résultats car ils sont soit non corrélées aux résultats ou ne diffèrent pas entre les participants et les non-participants | Résultats et "variables d'appariement" des participants et non-participants |
| Modèle de régression par discontinuité | Les personnes sont classées sur la base de critères spécifiques mesurables. Il y a un seuil limite pour déterminer qui est éligible pour participer. L'impact est mesuré en comparant les résultats des participants et non-participants proches de ce seuil. | Les personnes proches du seuil limite, mais qui se trouvent du côté de ceux qui vont (juste) rater l'admission au programme | Après contrôle des critères (et autres mesures de choix), les différences restantes entre personnes immédiatement en-dessous et immédiatement au-dessus du seuil limite ne sont pas statistiquement significatives et ne biaiseront pas les résultats. Il est nécessaire pour que cela soit valide qu'il y ait une adhérence stricte aux critères du seuil limite | Résultats et données de critères de classification (ex : âge, indice, etc.). Les variables du contexte socio-économique sont hautement souhaitables . |
| Evaluation randomisée | Un échantillon de personnes éligibles est aléatoirement assigné en deux groupes : ceux qui reçoivent l'intervention et ceux qui ne la reçoivent pas. L'impact est la différence des résultats entre les deux groupes. Il y a plusieurs façons de randomiser | Les participants sont aléatoirement affectés aux groupes de traitement et de comparaison | La randomisation est réussie et respectée ; c'est-à-dire que les deux groupes sont statistiquement identiques (en termes de facteurs observables et non-observables) | Données de résultat pour les groupes de comparaison et de traitement. Les situations de référence et variables de contexte sont souhaitables |

# Randomisation – Le Modèle de loterie

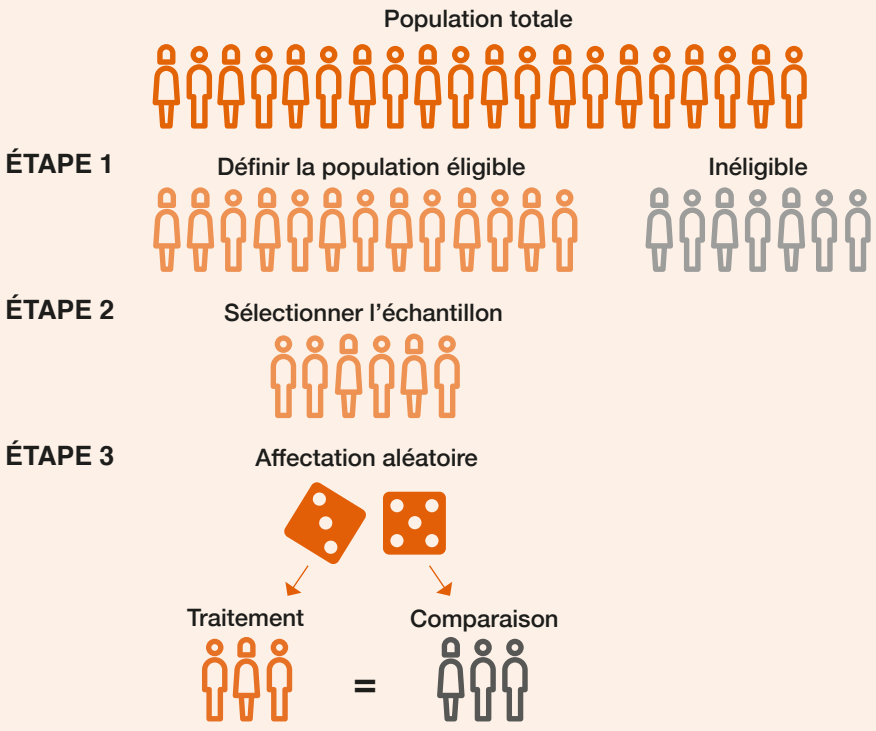
Une loterie est un moyen simple et transparent d'affecter des jeunes aux groupes qui recevront nos services (le groupe de traitement) et ceux qui ne les recevront pas (le groupe de comparaison). C'est la méthode utilisée par les modèles d'essais contrôlés aléatoires. Si un échantillon suffisamment large de personnes d'une même population d'intérêt est aléatoirement affecté à un des deux groupes, alors les deux groupes auront, en moyenne, des caractéristiques observables similaires (âge, genre, hauteur, niveau d'éducation, etc.). De façon tout aussi importante, ils auront aussi, en moyenne, les mêmes caractéristiques non-observables (telles que la motivation et l'état d'esprit).

Par le fait de la randomisation, la différence de résultats que nous observons entre les deux groupes à la fin de l'intervention peut être attribuée à celle-ci, car tous les autres facteurs pouvant influencer les résultats sont, en général, égaux.

## COMMENT ÇA MARCHE?

Il y a trois étapes dans le modèle de loterie (voir figure 5.3).

FIGURE 5.3: ÉTAPES DANS UN MODÈLE DE LOTERIE



## Étape 1 : Définir la population éligible

La première étape d'un essai contrôlé aléatoire est de trouver un groupe de jeunes éligibles pour l'intervention. Si une chercheuse en médecine étudie les effets d'un médicament contre une maladie infantile, elle recherchera un groupe spécifique d'enfants et ne prendra pas d'adultes ou de personnes âgées dans l'intervention. Similairement, un programme de formation professionnelle peut cibler les jeunes des villes à la rue se trouvant dans un intervalle d'âge spécifique, et n'inclura donc pas d'adultes ou de jeunes des villages. Ce qui est important ici est d'avoir des critères très clairs et transparents (âge, genre, niveau de revenus, statut de l'emploi, etc.) et d'être capable de communiquer sur qui sera éligible à l'intervention et qui ne le sera pas.

## Étape 2 : Sélectionner un échantillon pour l'évaluation

Pour évaluer une intervention, nous n'avons pas besoin de tester tous les participants à l'intervention. Il faut juste choisir un groupe de personnes représentatif en nombre suffisant pour les objectifs de l'évaluation ; c'est ce que l'on appelle l'**échantillon**. Il s'agit des jeunes sur lesquels les données seront collectées. La Note 6 fournit plus de détails sur la façon de déterminer l'échantillon et sa taille, toutefois, la taille typique de l'échantillon d'une intervention d'emploi pour jeunes évaluée par un modèle de loterie est quelque part entre 500 et 2.000 participants à l'étude (généralement divisés également entre les groupes de traitement et de comparaison).

Le choix de l'échantillon pour l'évaluation peut être effectué de deux façons, selon si l'intervention est conséquente ou petite. Une petite intervention peut déterminer qu'il y a 10.000 bénéficiaires éligibles, tels que les jeunes urbains des rues âgés entre 16 et 24 ans. L'intervention pourrait n'avoir de budget que pour aider 500 d'entre eux. Idéalement, un groupe de comparaison devra être similaire en taille au groupe de traitement, donc il faudra sélectionner 1.000 des 10.000 jeunes des rues pour l'intervention et l'évaluation (voir figure 5.4, image de gauche).

Les grands programmes peuvent être plus larges que la taille d'échantillon requise pour une évaluation. Si la formation professionnelle peut bénéficier à 4.000 jeunes, il n'est pas nécessaire de trouver 4.000 jeunes de plus pour la comparaison. Seulement 1.000 pourraient suffire. L'intervention peut alors identifier un échantillon de 5.000 jeunes dans la population totale de 10.000. De ces derniers, 3.000 jeunes pourront avoir une garantie d'admission à l'intervention. Les 2.000 autres seront ainsi aléatoirement divisés entre le groupe d'intervention et le groupe de comparaison (figure 5.4, image de droite).

Pour que la sélection soit représentative de la population éligible totale de 10.000 jeunes des rues, l'échantillon (que ce soit les 1.000 du premier cas ou les 5.000 du deuxième) devrait être sélectionné au hasard à partir de la population éligible. En faisant une sélection aléatoire, les participants auront, en moyenne, des caractéristiques similaires à celles de la population éligible totale. Bien que l'on inclût seulement un nombre limité de jeunes dans l'étude, l'impact potentiel de l'intervention peut être extrapolé à toute la population éligible, dans ce cas, à 10.000 jeunes.

**DÉFINITION**

**Un échantillon est un sous-ensemble de la population. Vu qu'il est généralement impossible ou non-pratique de collecter des informations sur l'ensemble de la population d'intérêt, on peut plutôt collecter des informations sur un sous-ensemble d'une taille plus gérable. Si le sous-ensemble est bien choisi, alors il est possible d'extrapoler les résultats à la population entière.**

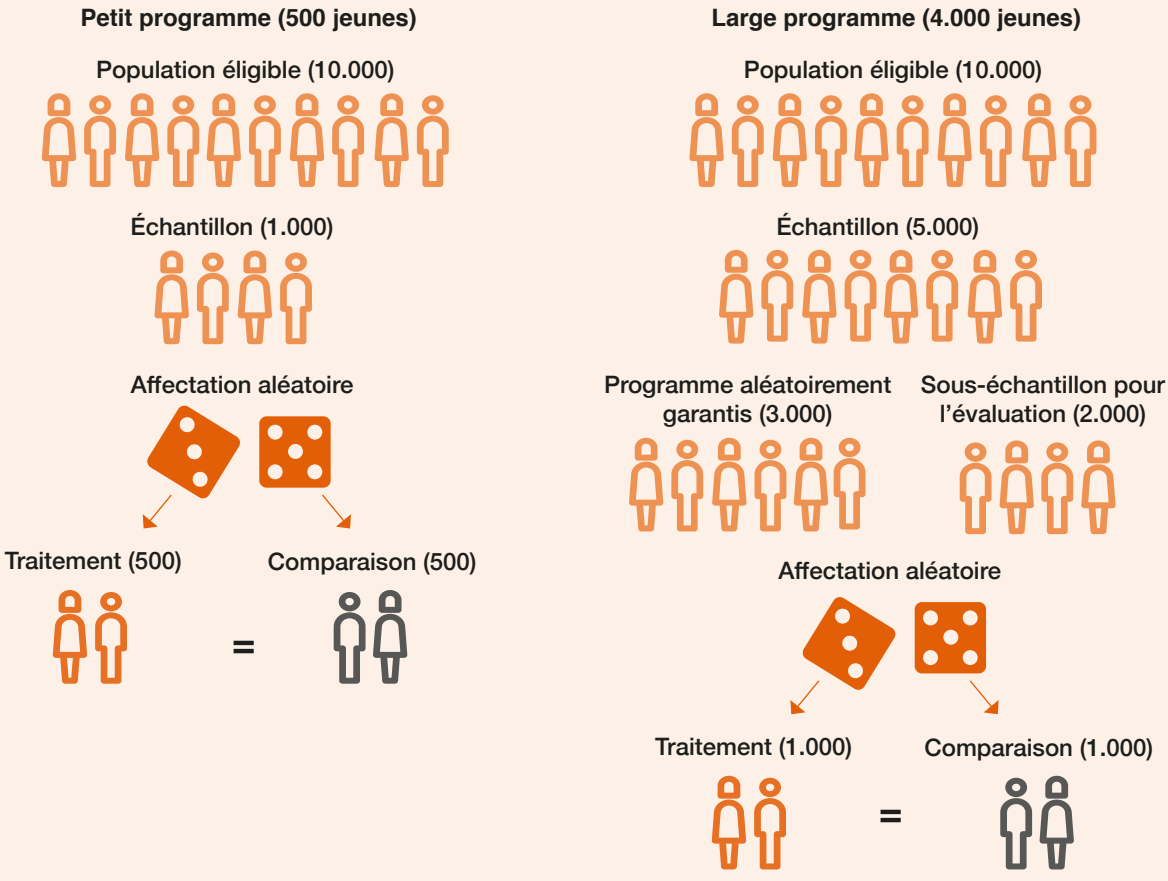
## Étape 3 : Randomiser l'affectation

L'étape suivante est d'affecter l'échantillon de jeunes sélectionnés aux groupes de traitement et de comparaison qui sont sensiblement égaux en termes de taille. Dans les essais contrôlés aléatoires, chaque jeune a la même chance de recevoir l'intervention. La randomisation peut être effectuée en utilisant des techniques traditionnelles, telles qu'un pile-ou-face d'une pièce de monnaie, lancer les dés, ou tirer des noms d'une urne. La randomisation peut être faite publiquement si l'échantillon est assez petit (tirer 2.000 noms d'une urne, par exemple, ne serait pas très pratique). Sinon, si le nombre de personnes est large, on peut randomiser par logiciel informatique tel que Microsoft Excel. La randomisation peut avoir lieu à plusieurs

**CONSEIL**

**Une façon d'obtenir un échantillon aléatoire de jeunes est de se procurer la liste d'une population totale de jeunes des rues à partir d'un recensement, des listes électorales ou toute autre base de données, ou autre, et sélectionner aléatoirement à partir de cette liste. Si ce n'est pas possible, cibler au hasard des endroits d'interactions de jeunes, tels que les centres urbains, pourra fournir un échantillon aléatoire. S'il est connu que les jeunes sont dispersés dans 50 divers centres de la ville ou du pays, sélectionner au hasard des centres, puis une portion de jeunes dans ces centres pour participer à l'étude est susceptible de résulter en une sélection de jeunes avec un biais minimal. La Note 6 abordera l'échantillonnage plus en détail.**

FIGURE 5.4 : CHOISIR DES ÉCHANTILLONS POUR PETITS ET LARGES PROGRAMMES



niveaux. En affectant aléatoirement aux groupes de traitement et de comparaison, les participants sont sélectionnés aléatoirement et un contrefactuel est développé : si la taille d'échantillonnage est assez large, les jeunes du groupe de traitement ont, en moyenne, les mêmes caractéristiques observables et non-observables que ceux du groupe de comparaison.

## QUAND PEUT-ON UTILISER UN MODÈLE DE LOTERIE ?

Une évaluation de loterie aléatoire peut être utilisée quand l'évaluation est planifiée en amont de la mise en œuvre (prospection) et quand l'intervention ne peut bénéficié qu'à une fraction des jeunes éligibles. Tant que des contraintes de ressources empêchent l'intervention de servir l'entière population éligible, il n'y a aucun souci d'éthique dans le fait d'avoir un groupe de comparaison, car un sous-ensemble de la population sera nécessairement exclu de l'intervention. Dans une telle situation, les groupes de comparaison peuvent être maintenus pour mesurer les impacts à court, moyen et long terme de l'intervention (Gertler et al., 2016). Il est important de comprendre que l'avantage principal des randomisations – soit, que les groupes de traitement et de comparaison, en moyenne, aient les mêmes caractéristiques – ne sera maintenu que si nous arrivons à faire un suivi avec (presque) tous les membres des groupes de traitement et de comparaison. Les taux élevés d'abandon posent une menace sévère à la validité interne des résultats pour chaque méthode d'évaluation d'impact, y compris les méthodes qui utilisent les techniques de randomisation.

## AVANTAGES

- Un modèle de loterie est la méthode la plus solide pour développer un contrefactuel car elle mène à un groupe de comparaison très bien apparié (qui se base sur moins d'hypothèses que les autres méthodes). Il est ainsi considéré comme le modèle le plus crédible pour mesurer l'impact.
- Il est de loin la plus simple des méthodes d'évaluation en termes analytiques. L'impact de l'intervention dans un essai aléatoire est simplement la différence moyenne des résultats entre les groupes de traitement et de comparaison.
- Il permet aux communautés d'être directement impliquées dans le processus de sélection, résultant ainsi en une allocation des bénéfices équitable et transparente.
- Il est facile à mettre en œuvre et à communiquer au personnel du programme.

## LIMITES

- Faire une expérience aléatoire peut être très coûteux et prendre beaucoup de temps.
- Aucune mise en œuvre ex-post de cette méthode n'est possible. La planification de l'évaluation doit faire partie de la planification de l'intervention (ce qui est une bonne pratique dans tous les cas, mais ne représente pas toujours la réalité du projet).
- Il requiert que le groupe de comparaison soit exclu de l'intervention pendant la durée de l'évaluation d'impact. Des questions politiques et/ou éthiques pourraient être

soulevées malgré le critère d'affectation transparente de la randomisation (en savoir plus dans la section "Adapter les modèles aléatoires aux divers contextes" ci-dessous).

- Les organisations doivent obtenir l'accord des partenaires et parties prenantes locales.
- La validité interne d'un modèle de loterie dépend du fait que la randomisation marche et est maintenue pendant l'étude, ce qui peut s'avérer difficile à accomplir. Cette condition peut être menacée si la randomisation n'est pas correctement mise en œuvre, si les groupes de traitement ou de comparaisons ne se conforment pas à leur statut (c.à.d., les individus du traitement ne prennent pas l'intervention ou ceux de la comparaison reçoivent le programme), si les participants abandonnent l'étude avant la fin ou s'il y a des **effets d'entraînement** : par exemple, les jeunes qui reçoivent la formation professionnelle peuvent transférer les compétences et connaissances acquises à leurs pairs, rendant ainsi la séparation entre groupes de traitement et de comparaisons plus floue. Ces cas sont très problématiques car ils peuvent fortement biaiser les résultats et ainsi menacer la validité globale de l'évaluation.

**DEFINITION**

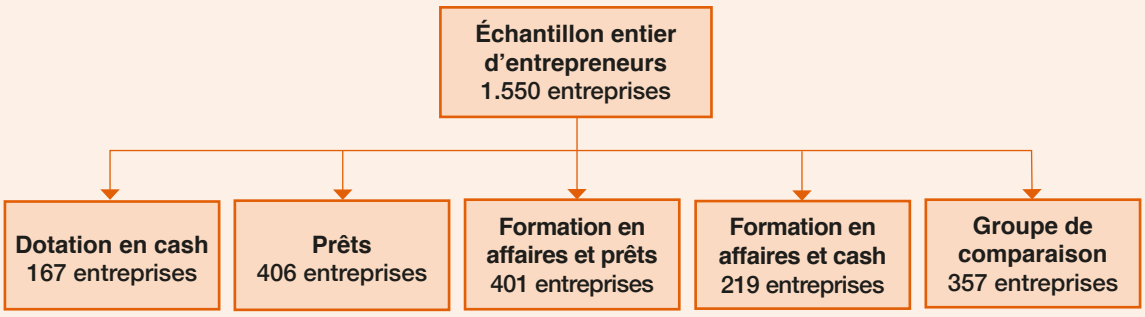
**Effets d'entraînement : ce sont les effets de retombées d'une intervention sur les non-participants ; par exemple, lorsque les compétences d'une formation technique sont disséminées dans le village, même à ceux qui n'ont pas suivi les cours.**

## Encadré 5.2: Évaluation de l'intervention de l'OIT "Gérez mieux votre entreprise (GERME)"

L'évaluation a été conçue pour tester si l'expansion de l'accès au capital par les subventions ou les prêts augmentait les profits des microentreprises des hommes et des femmes, et si la formation en entreprenariat GERME de l'OIT pouvait accroître encore plus les impacts.

L'équipe de recherche a mené une enquête sur 4.637 microentreprises d'un recensement de commerce et a sélectionné 1.550 entrepreneurs à inclure dans l'échantillon d'évaluation – sur la base, parmi d'autres critères, de l'expression de leur intérêt à recevoir une formation de l'OIT et à participer à l'intervention de prêts. L'échantillon incluait de petits entrepreneurs intéressés à améliorer leur commerce (par exemple, salons de coiffure, magasins de détail et tailleurs). L'échantillon était aléatoirement divisé en cinq branches de traitement qui ont reçu les interventions suivantes : (1) un prêt ; (2) une dotation en cash ; (3) une formation en affaires et un prêt ; (4) une formation en affaires et une dotation en cash ; et (5) aucune intervention (le groupe de comparaison) (figure 5.5).

FIGURE 5.5 : MODÈLE DE L'ÉVALUATION



Les interventions principales auprès des propriétaires de commerces étaient :

- Formation en affaires : le programme de formation Démarre ton affaire (SYB, *Start Your Business*) cible les nouveaux entrepreneurs et consiste en un stage de formation de 5 jours, suivi par un travail sur le terrain et des séances de conseil en groupe et individuellement. Les stagiaires préparent leur plan d'affaires et plan d'action bancables détaillés (voir www.ilo.org/siyb).
- Des subventions en cash inconditionnels, valorisés à 200 USD, ont été livrés sur des comptes en banque gratuits d'une institution de microfinance (IMF) locale. Les propriétaires de commerce avaient le choix libre de l'usage de leur prêt.
- Des prêts semi-conditionnels, valorisés de 180 à 220 USD, ont été offerts à un taux annuel réduit de 20 pour cent par l'IMF. Les prêts devaient être remboursés à l'IMF, mais il n'y avait pas de sanction en cas d'utilisation abusive des fonds.

La taille des dons et des prêts est égale à approximativement 1,5 fois les profits mensuels des commerces moyens.

Pour vérifier si la randomisation a "marché", l'évaluation a comparé les propriétaires de commerce du groupe de traitement à ceux du groupe de comparaison au regard de 26 variables différentes et a trouvé que, pour quasiment chaque caractéristique, les entreprises traitées et non-traitées étaient en moyenne similaires avant l'intervention.

Les propriétaires de commerce ont été enquêtés avant l'intervention (enquête de référence) et à six mois, neuf mois et deux ans après l'intervention (trois enquêtes de suivi).

Le principal résultat de la variable d'intérêt était le profit du commerce et l'évaluation a révélé une augmentation considérable des revenus après l'intervention, mais seulement pour les entrepreneurs masculins. Aucune des interventions n'a résulté en une augmentation soutenue des profits chez les femmes entrepreneurs. Les femmes qui avaient initialement de hauts profits ont aussi eu des effets négatifs dans toutes les interventions. Alors que la réponse initiale aux subventions était positive, cette augmentation s'est complètement évaporée et est même devenue négative avec le temps. Les femmes qui ont reçu la subvention ont fait 35 pour cent moins de profit que celles qui n'avaient reçu aucune intervention. Après neuf mois, les femmes étaient soit en moins bonne posture ou même dans une plus mauvaise position que leurs pairs dans le groupe de comparaison.

L'évaluation a aussi trouvé que la proximité des membres de la famille constitue une force positive sur le commerce pour les hommes et une force négative pour les femmes. Les femmes mariées avec des familles vivant dans le même district ont subi de grosses diminutions de profits.

Source: Fiala, 2015.

# Adapter des modèles randomisés à différents contextes

Certains programmeurs sont réticents à l'idée d'affecter aléatoirement des bénéficiaires potentiels aux groupes de traitement et de comparaison. Le souci en général est que l'évaluation cause la rétention de bénéfices apparemment évidents (tels que les opportunités de formation) chez des personnes qui en ont besoin, ce qui serait contraire à l'éthique. Il n'en reste pas moins que, pour beaucoup d'interventions, la demande excède de loin ce qui peut être offert et, comme signifié dans l'encadré 5.3, la randomisation pourrait en fait être plus éthique que les autres méthodes de sélection.

Néanmoins, créer un pur groupe de comparaison par l'affectation aléatoire d'une loterie dans laquelle les jeunes ne reçoivent aucune intervention est parfois impossible. Le modèle d'instauration graduelle randomisé (randomized phase-in design)

## Encadré 5.3 : La randomisation est-elle éthique ?

Parfois, affecter aléatoirement des bénéficiaires potentiels aux groupes de traitement et de comparaison est considéré comme contraire à l'éthique. Ces soucis pourraient être valides dans certains cas, par exemple lorsqu'une politique ou intervention susceptible de marcher ou qui a déjà fait ses preuves peut être élargie à moindre coût à une grande population. Toutefois, il arrive souvent qu'une des situations suivantes se présente :

- **Incertitude de l'impact du projet.** Pour la plupart des programmes, on ne sait pas si l'intervention a un impact positif et significatif sur l'individu et la communauté qui justifie les ressources dépensées. Par exemple, les programmes destinés aux filles et excluant les garçons peuvent augmenter la violence fondée sur le genre. Une intervention de microfinance pour jeunes peut mener les participants à une pire situation s'ils ne peuvent pas payer leurs dettes. Un programme de formation mal conçu pourrait diminuer les chances d'emplois. Une augmentation des revenus (ex : 100 USD par participant) pourrait coûter très cher (ex : 1.000 USD par personne). Ainsi, dans le cas des interventions dont l'impact et la structure coût-bénéfice n'ont pas encore fait leurs preuves, il est bien justifié d'évaluer l'intervention sur la base des affectations aléatoires aux groupes de traitement et de comparaison.
- **Contraintes budgétaires.** En réalité, à cause des ressources limitées, il est rarement possible de servir tous ceux dans le besoin. C'est-à-dire, la plupart des programmes fournissent des bénéfices et des services seulement à un nombre limité de bénéficiaires, excluant ainsi les autres, que ce soit fait de façon explicite ou non. Par exemple, si une intervention de formation de jeunes a un nombre limité de places disponibles, alors certains jeunes recevront la formation et d'autres pas. Similairement, si une intervention est mise en œuvre dans un district particulier, les jeunes éligibles des autres districts sont exclus. La randomisation permet aux managers de programmes d'attribuer les places limitées de leurs interventions d'une façon équitable qui donne les mêmes chances de participation à tous. Si la randomisation est faite de façon ouverte (par exemple, par une loterie lors d'un évènement public), elle apporte aussi la transparence dans le processus de sélection et peut réduire les craintes de la population que la sélection se soit faite sur des préférences personnelles ou politiques.

et le modèle de promotion aléatoire évitent la séparation stricte en groupes de traitement et de comparaison et pourraient être une alternative viable pour un modèle d'évaluation d'impact expérimental lorsque les modèles de loterie ne sont pas faisables ou souhaitables.

## MODÈLE D'INSTAURATION GRADUELLE RANDOMISÉ

Dans les cas où plusieurs programmes seraient actifs dans une communauté pendant des années, ne jamais donner l'intervention à un groupe de personnes dans le besoin peut s'avérer difficile sur les plans politique et programmatique. Une variation du modèle de loterie est le modèle d'instauration graduelle randomisé (randomized phase-in design). La différence principale entre ces deux modèles est la méthode d'affectation des personnes aux groupes de traitement et de comparaison. En pratique, les bénéficiaires potentiels sont aléatoirement divisés en deux ou plusieurs groupes. L'intervention est alors étalée dans le temps, de sorte que les personnes du premier groupe participent à l'intervention en premier, suivis par ceux du deuxième groupe, puis du troisième groupe, et ainsi de suite. Pendant que les groupes sont sur la liste d'attente, ils peuvent être utilisés comme groupes de comparaison jusqu'à ce qu'ils reçoivent l'intervention.

Par exemple, une organisation non-gouvernementale (ONG) peut avoir un budget pour la formation de 1.500 jeunes, mais peut ne pas avoir la capacité de donner la formation à tous simultanément. Elle pourrait ainsi choisir de former 500 jeunes par an sur une période de trois ans. Si elle peut identifier tous les 1.500 participants au départ, une randomisation par phases pourrait être la meilleure méthode d'évaluation à adopter. Les 1.500 jeunes sont aléatoirement divisés en trois groupes. La première année, pendant que le groupe 1 reçoit la formation, les groupes 2 et 3 sont sur la liste d'attente et servent comme groupe de comparaison. La deuxième année, seulement le groupe 3 sert de groupe de comparaison. La troisième année, les trois groupes auront reçu la formation.

**TIP**

**Avec une approche par phases, il est crucial d'avoir assez de temps entre chacune des phases pour permettre à l'intervention de produire ses effets. Si, par exemple, le responsable d'intervention estime qu'il faudra deux ans pour que l'impact de l'intervention produise des effets, la première et la dernière phase doivent être à au moins deux ans d'écart. Cette approche pourrait ne pas être adéquate pour les petits programmes ou les programmes de court terme.**

Comme les personnes sont sélectionnées aléatoirement pour les différents groupes, il est possible de comparer ceux qui ont été traités en premier avec ceux qui l'ont été en dernier. Cette méthode convient au déploiement naturel de beaucoup de programmes.

Cependant, vu que tout le monde bénéficiera du programme, le modèle d'instauration graduelle randomisé (randomized phase-in design) n'est généralement pas idéal pour déterminer l'impact à long terme d'une intervention car il finit par ne plus y avoir de groupe de comparaison. Même les larges programmes en place depuis longtemps auront des difficultés à demander aux participants d'attendre leur tour pendant trois ou quatre ans, donc l'échéance des résultats est souvent limitée à un ou deux ans. De plus, il y a le risque que les participants puissent changer de comportement pendant leur période d'attente, ce qui pourrait invalider leur capacité à servir de bon groupe de comparaison. Par exemple, en anticipation à la formation professionnelle de l'intervention, ils pourraient arrêter de chercher du travail.

# MODÈLE DE PROMOTION ALÉATOIRE / D'ENCOURAGEMENT

Il peut y avoir des cas où il n'est pas possible ou souhaitable d'exclure des bénéficiaires potentiels et dans lesquels l'intervention n'est pas étalée dans le temps. Dans de tels cas, la méthode de promotion aléatoire (appelée aussi modèle d'encouragement) pourrait convenir. Lorsqu'il n'est pas possible d'affecter aléatoirement les jeunes dans un groupe qui reçoit les bénéfices et un groupe qui ne les reçoit pas, il peut être possible de faire une promotion randomisée de l'intervention. C'est-à-dire, au lieu de randomiser ceux qui *reçoivent* les bénéfices, on randomise ceux qui sont *encouragés à recevoir* ces bénéfices.

L'encouragement aléatoire peut prendre diverses formes. Dans le cas de comptes d'épargne pour jeunes, on peut aléatoirement promouvoir l'initiative dans des écoles sélectionnées. Pour un programme de formation, on pourrait embaucher un travailleur social pour aléatoirement visiter les foyers de jeunes sans emploi, décrire le programme et offrir l'inscription des jeunes sur le champ. Dans le cas d'une campagne d'éducation financière, on pourrait vouloir envoyer aléatoirement des messages-textes à une partie de l'audience ciblée mais pas à une autre. Dans tous les cas, il y aura toujours des personnes dans le groupe promu qui ne prendront pas notre intervention tout comme il y aura des personnes du groupe non promu qui la prendront. Mais l'idée est que, si l'encouragement est efficace, le taux d'inscription parmi le groupe promu devrait être plus élevé que celui du groupe qui n'a pas reçu la promotion.

Pour évaluer l'impact de l'intervention, on ne peut malheureusement pas simplement comparer les résultats de ceux qui ont participé à l'intervention à ceux qui n'y ont pas participé. Les personnes qui choisissent de participer à une intervention sont presque toujours différentes de celles qui choisissent de ne pas y participer, et plusieurs de ces différences peuvent ne pas être observables ou mesurables. Même si la promotion est randomisée, la participation à l'intervention ne sera pas aléatoire, donc comparer participants et non-participants serait semblable à comparer des pommes à des oranges.

On peut, toutefois, comparer les résultats entre tous ceux qui ont reçu l'encouragement et tous les jeunes du groupe de comparaison. Vu que la promotion est affectée aléatoirement, les groupes promus et non-promus ont, en moyenne, des caractéristiques égales. Ainsi, la différence des résultats moyens que nous observons entre les deux groupes peut être attribuée au fait que ces personnes ne se sont inscrites à l'intervention que parce qu'elles ont reçu la promotion.

Un des avantages principaux de ce modèle est que les campagnes de promotions randomisées ne refusent jamais le programme à qui que ce soit, mais plutôt permettent aux personnes de prendre leurs propres décisions de recevoir l'intervention ou pas. Cependant, ces études nécessitent souvent de larges échantillons pour pouvoir fournir des estimations d'impact fiables, ce qui augmente les coûts en conséquence.

**CONSEIL**

**L'évaluation d'une promotion aléatoire peut être appropriée pour les interventions qui :**

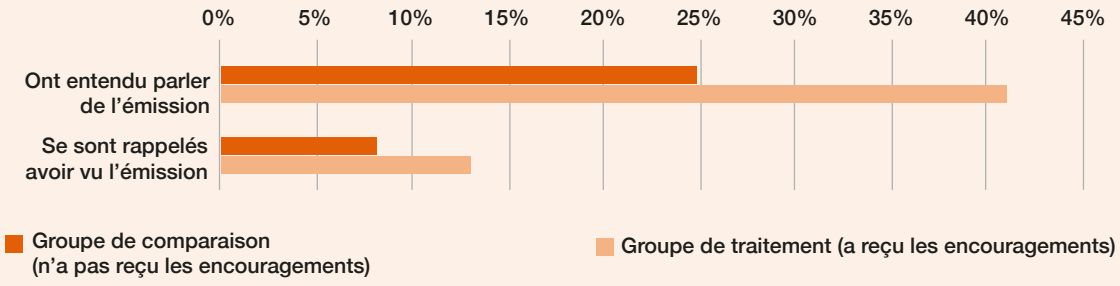
- Offrent des coupons de formation
- Encouragent les jeunes à ouvrir des comptes d'épargne
- Utilisent des campagnes médiatiques

## Encadré 5.4: Évaluer une intervention ludo-éducative en Égypte

El Mashroua est une émission de télé-réalité, conçue pour promouvoir l'entreprenariat parmi les jeunes adultes de l'audience, qui est diffusée sur l'une des chaines de télévision égyptienne les plus populaires. Pour évaluer l'impact de l'émission, une équipe de recherche, soutenue par l'OIT, a choisi un modèle de promotion aléatoire. Un groupe de traitement sélectionné aléatoirement à partir de l'échantillon de 9.277 personnes de l'étude a reçu par SMS des rappels pour les encourager à regarder l'émission.

L'enquête de suivi menée environ un an et demi après l'émission a clairement démontré que les jeunes du groupe de traitement (ceux qui ont reçu les messages) étaient plus susceptibles d'avoir entendu parler de l'émission et d'avoir regardé au moins un épisode, comparé au groupe de comparaison (qui n'a pas reçu les rappels). Voir aussi figure 5.6.

### FIGURE 5.6: ADOPTION DE L'ÉMISSION EL MASHROU3




Ces différences statistiquement significatives peuvent être exploitées pour estimer les impacts actuels de l'émission sur le comportement de l'audience et des résultats du marché du travail. Cela est possible grâce à deux hypothèses que l'on peut faire:

1. À cause de la randomisation, les groupes de traitement et de comparaison ne diffèrent pas systématiquement pour toute caractéristique observable ou non-observable qui pourrait être corrélée aux variables du résultat.

2. Parce que le seul fait d'avoir reçu les messages n'affecte ni les comportements ni les résultats du marché du travail, toute différence entre les groupes de traitement et de comparaison peut être attribuée à la différence de la probabilité d'avoir regardé l'émission.

L'étude a déterminé que le fait d'avoir regardé l'émission n'a pas eu d'impacts sur la propension des jeunes à lancer leur propre affaire, mais a considérablement réduit les attitudes de discrimination des hommes contre les femmes basées sur le genre.

Source: Barsoum et al., 2017.

# La méthode des doubles différences (DID, *Difference-in-differences*)

Pour des raisons déjà expliquées, il n'est parfois pas possible ou souhaitable d'employer des méthodes d'évaluation expérimentales. Dans ce cas, il y a une gamme de méthodes d'évaluation d'impact quasi-expérimentales qui peuvent aussi produire des résultats robustes et valides à l'interne.

Une des techniques les plus couramment utilisées est l'approche de double différences (DID, *difference-in-differences*) qui compare le changement des résultats du groupe de traitement à celui des résultats du groupe de comparaison.

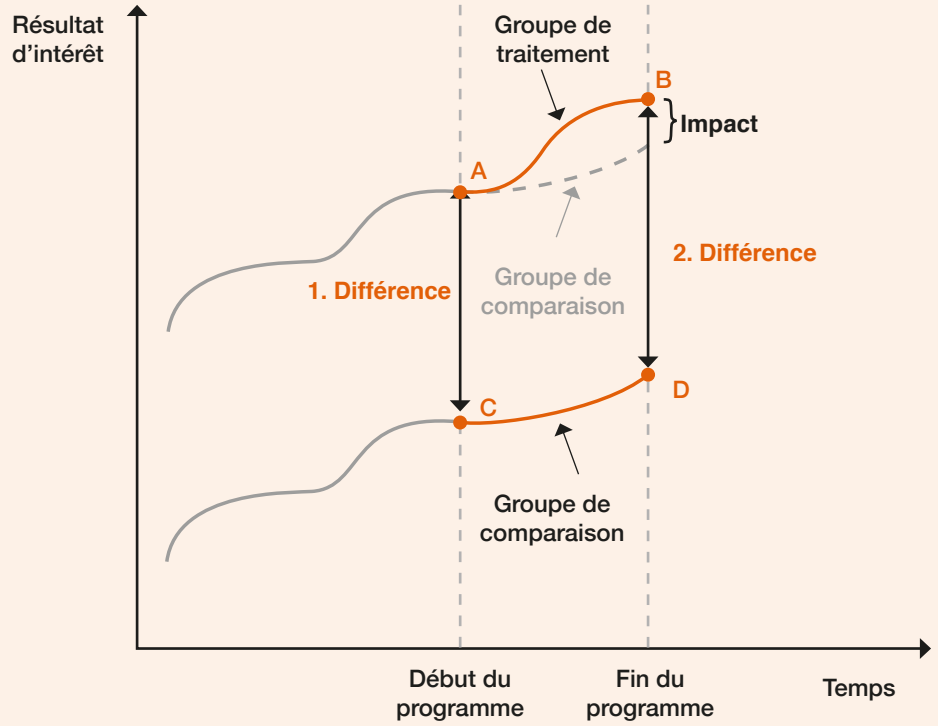
## COMMENT ÇA MARCHE ?

**Identifier le groupe de comparaison :** les modèles DID s'appuient sur la mise en place d'un groupe de comparaison dont nous pouvons raisonnablement assumer que le développement des résultats d'intérêt clés serait le même que celui du groupe de traitement pendant la période de l'intervention. Pour ce faire, il est souhaitable de choisir des groupes à caractéristiques similaires.

Imaginons une intervention de formation professionnelle de six mois pour jeunes, pour laquelle nous voulons évaluer les impacts sur les résultats du marché du travail. Distribuer aléatoirement des places de formation n'est pas possible. Donc, on prend à la place un échantillon de jeunes des mêmes âge, niveau d'éducation, contexte socio-économique et situation de marché de travail d'une autre communauté, comme groupe de comparaison.

**Estimer l'impact :** pour appliquer la technique d'évaluation DID, il faut (a) mesurer les résultats d'intérêt (par exemple, statut du marché de travail, voir Note 2) pour les groupes de traitement et de comparaison *avant* que ne débute l'intervention de formation professionnelle et (b) mesurer les résultats des deux groupes à un instant donné, *après* que l'intervention ait eu lieu. Bien que nous nous soyons attelés à identifier un groupe de comparaison de jeunes paraissant similaire au groupe de traitement, il est vraisemblable qu'il y ait des différences entre les deux groupes avant la formation, et que ces différences persistent après la formation. La figure 5.7 illustre une situation dans laquelle le groupe de comparaison a un indicateur de résultat considérablement inférieur (disons, le statut d'emploi) comme situation de référence. Cependant, cela n'affecte pas la méthode. La technique DID compare la différence des résultats entre les deux groupes à la fin de l'intervention (point B moins D) et l'ajuste en fonction de la différence de résultats entre les deux groupes au début (A moins C). Soustraire ces différences l'une de l'autre (c'est-à-dire, établir une différence à partir de deux différences – d'où son nom) donne une idée de l'impact du programme. Cela montre si, et dans quelle mesure, l'intervention de la formation a augmenté le statut de l'emploi des participants par rapport à celui des non-participants. Le scénario de la figure 5.7 indique un impact positif modéré de l'intervention de la formation professionnelle.

FIGURE 5.7 : EXEMPLE D'ANALYSE D'ÉCART DANS LES DIFFÉRENCES




**Impact = ( 2. Différence ) - ( 1. Différence )**

Source: Adapté de Gertler et al., 2016.

**L'hypothèse de la "tendance commune" :** l'hypothèse qui sous-tend cette méthode est que, bien que les caractéristiques observables et non-observables des groupes de traitement et de comparaisons puissent être différentes (reflétées dans les niveaux de revenus différents au début de l'intervention), leurs *différences sont constantes dans le temps*, ou invariantes avec le temps. Ceci permet d'utiliser la tendance du groupe de comparaison comme estimation de ce qui serait arrivé au groupe de traitement s'il n'y avait pas eu l'intervention. En conséquence, il n'est pas nécessaire d'assumer que, sans l'intervention, les résultats seraient restés constants, mais plutôt que les groupes de traitement et de comparaison ont la même *tendance* dans le temps. C'est ce que nous appelons l'hypothèse de la « tendance commune ».

Revenant à notre exemple de formation professionnelle ci-dessus, pour pouvoir appliquer la méthode DID, nous devons être sûrs que dans les six mois qui suivent il n'y aura pas de facteurs qui influenceront systématiquement les résultats des jeunes de la « communauté traitée » de manière différente qu'ils influenceront ceux des jeunes de la « communauté de comparaison », à part l'affectation à la formation. Par exemple, une croissance économique plus rapide, une nouvelle politique locale offrant des incitations aux sociétés pour l'emploi des jeunes ou un gros employeur arrêtant ses opérations dans seulement une des deux communautés bafoueraient cette l'hypothèse et biaiseraient en conséquence les résultats de l'évaluation.

Un bon test pour déterminer s'il est réaliste d'assumer des tendances communes entre

participants et non-participants est de comparer leurs changements de résultats avant que l'intervention ne soit mise en œuvre. Cette approche requiert plusieurs points de données avant l'intervention. Comme plusieurs enquêtes de référence peuvent rapidement devenir très onéreuses, ce test peut plus facilement être effectué si les données administratives sur nos indicateurs-clés de résultats sont disponibles à moindre coût (par exemple, le statut de l'emploi des services publics d'emploi ou des notes de classe des années scolaires précédentes). Si les résultats des deux communautés ont évolué en tandem avant le début de l'intervention, nous pouvons être plus confiants sur le fait que ces résultats continueront à afficher cette tendance pendant l'intervention. Si, toutefois, les tendances pré-intervention sont différentes, l'hypothèse de la tendance commune peut s'avérer incorrecte.

## QUAND UTILISER UN MODÈLE DID ?

Vu qu'elle assume que les différences entre participants et non-participants sont constantes dans le temps, cette méthode est la plus utile lorsqu'il y a de bonnes données à plusieurs points avant le début de l'intervention. Pour améliorer la crédibilité des estimations de l'impact, il est préférable d'avoir au moins trois tours de collectes de données, soit deux avant le traitement et au moins un à la fin de l'intervention (voir ci-dessus). Ceci signifie que, à moins que des données sur les parti-cipants et non-participants soient disponibles par d'autres canaux, tels qu'une enquête de ménages existante, les coûts d'une telle évaluation peuvent être beaucoup plus élevés que ceux d'autres techniques d'évaluation d'impact.

## AVANTAGES

- Cette méthode fournit un moyen de tenir compte des différences observables et non-observables entre participants et non-participants. Plus précisément, elle contrôle tous les effets individuels qui restent constants dans le temps, ou qui suivent la même ligne de changement dans le temps (quand par exemple les groupes traités et de comparaison affichent des tendances similaires dans les résultats d'intérêt).
- Même si la méthode n'est pas expérimentale, elle permet une vérification (partielle) de l'hypothèse qui la rend valide à l'interne. Ceci signifie que nous pouvons avoir une idée de la validité ou non de nos impacts estimés. Si de bonnes données administratives sont disponibles, la méthode peut être appliquée assez facilement, et même ex-post, sur la base de données avant-et-après une programme.

## LIMITES

- Cette méthode produit des résultats moins fiables que les méthodes de sélection randomisées.
- Pour tester l'hypothèse-clé de « *tendance commune* », au moins trois tours de collectes de données sont requis, ce qui rend la mise en œuvre coûteuse s'il n'y a pas de données initialement disponibles.

## Encadré 5.5 : Évaluation d'une composante d'activation d'un marché du travail pour participants d'un programme de transfert d'espèces conditionnel en Argentine

Le programme Plan Jefes est un programme de transfert d'espèces conditionnel introduit pendant la crise économique de l'Argentine de 2001–2002. Les réformes de ce programme, suite au relèvement après la crise, ont inclus la mise en œuvre d'une Assurance Formation et Emploi (Seguro de Capacitación y Empleo, SCE) en 2006, pour fournir un appui au relèvement de compétences, à la formation professionnelle, à la recherche d'un emploi et au placement professionnel pour les participants éligibles au Plan Jefes.

Les participants au SCE reçoivent une dotation mensuelle et les mesures d'activation suivantes :

- Assistance pour la réussite d'une éducation primaire et secondaire
- Formation professionnelle et stages d'apprentissage
- Services d'intermédiation à l'emploi
- Mesures indirectes de création d'emploi (ex : subvention d'emploi)
- Promotion du travail autonome et de la création de microentreprises.

L'OIT a étudié à l'aide d'estimateurs DID l'effet de la mise en œuvre de ces outils du marché du travail actif sur le statut et la qualité de l'emploi des bénéficiaires du programme Plan Jefes. Afin d'isoler l'effet de ces outils, un groupe de comparaison avec des caractéristiques similaires de celles des participants SCE devait être identifié.

Comme le transfert du Plan Jefes au nouveau programme était graduel, les chercheurs ont pu sélectionner des participants du Plan Jefes qui étaient conformes aux exigences des bénéficiaires du SCE mais qui n'avaient pas encore été transférés au nouveau programme. Une hypothèse-clé importante de cette stratégie d'identification était que la transition entre les programmes n'était influencée par aucun des facteurs pouvant mener à des différences sur les résultats d'intérêt. Un total de 1.149 non-participants ont été sélectionnés à partir des données de l'Enquête Permanente auprès des Ménages d'Argentine – une enquête trimestrielle effectuée par l'Institut National des Statistiques d'Argentine (INDEC) qui inclut des questions sur les caractéristiques individuelles, l'éducation et la performance professionnelle des personnes. Les participants et non-participants sélectionnés étaient similaires en genre, âge et niveau d'éducation.[5]

Les évaluateurs ont comparé une gamme de résultats entre les deux groupes à deux moments différents (référence et suivi). Cette approche a permis d'identifier les effets de causalité du programme SCE tout en contrôlant le biais de sélection dû aux caractéristiques observables et non-observables des participants.

La structure de groupe de l'enquête a permis aux chercheurs de collecter des données sur les participants et les non-participants à plusieurs moments dans le temps, aussi bien avant qu'après la participation au programme. Ceci a permis d'émettre l'hypothèse que, sans le programme, les résultats des participants et des non-participants auraient changé de la même façon (l'hypothèse de «la tendance commune» à tester et confirmer).

Les résultats de l'étude ont montré que le programme a eu un effet positif sur la qualité d'emploi des participants, comme par exemple la probabilité d'avoir un emploi formel et un salaire horaire plus élevé, ainsi qu'une plus basse probabilité d'avoir un emploi moins bien payé et de travailler un nombre excessif d'heures. Il n'a pas affecté leur statut d'emploi (par exemple la probabilité d'être employé). L'évaluation a aussi démontré les effets hétérogènes, révélant que le programme a eu un impact plus élevé sur les jeunes bénéficiaires, et aucun effet sur les femmes.

[5] Pour corriger ces différences observées, les chercheurs ont aussi appliqué la méthode d'appariement du score de propension (PSM). Voir la section suivante pour une description plus détaillée de cette méthodologie.

# Appariement par score de propension (PSM, Propensity score matching)

L'appariement par score (ou coefficients) de propension (PSM, *propensity score matching*) est une approche très utilisée parmi les méthodes d'évaluation quasi-expérimentales. Son principe de base est de construire un groupe de comparaison en appariant les participants à des non-participants similaires, sur la base de la prédiction de leur probabilité à participer à l'intervention. C'est ce qu'on appelle le score de propension, qui est calculé à l'aide de plusieurs caractéristiques observées.

## COMMENT ÇA MARCHE ?

Une gamme de covariables potentiellement pertinentes doivent être sélectionnées afin de calculer le score de propension des non-participants, fondé sur leur probabilité à être traité. Le but est d'inclure dans le calcul du score de propension toutes les covariables qui affectent aussi bien la participation au programme que les résultats. Les non-participants sont ensuite appariés avec les participants sur la base de leurs scores respectifs. Il y a plusieurs types de procédures d'appariement, dont l'approche la plus commune est celle du voisin le plus semblable par laquelle chaque participant est apparié au non-participant qui a le score de propension le plus proche. Plus proche est le score, meilleure est la qualité de l'appariement. Des tests d'équilibre peuvent être effectués pour évaluer dans quelle mesure l'appariement a marché. En conséquence, la différence moyenne des deux groupes de résultats d'intérêt pertinents est équivalente à l'impact de l'intervention.

Par exemple, prenons un programme de formation professionnelle ciblant la jeunesse rurale qui a 1.000 participants. Les données d'avant-programme sur les caractéristiques clés des participants sont disponibles, telles que le sexe, l'âge, l'éducation et les aspects clés de leur historique sur le marché du travail. Des données d'enquête secondaires peuvent être utilisées pour construire un groupe de comparaison sur la base de leur score de propension, estimant la probabilité de traitement pour un grand nombre de personnes, fondé sur les caractéristiques susmentionnées. Un total de 1.000 personnes ayant les meilleurs appariements de scores de propension seront sélectionnées comme groupe de comparaison pour l'intervention. Les données post-intervention de la comparaison pourraient être collectées par le biais de la même source de données secondaires (s'il s'agit d'un jeu de données de panel, plusieurs vagues peuvent être collectées.).

## QUAND UTILISER LA MÉTHODE PSM ?

La méthode PSM est une méthode particulièrement utile lorsque nous disposons de données secondaires extensives et riches en contenu, vu qu'elles sont nécessaires pour définir un bon score de propension et pour apparier un nombre suffisant de participants

## Encadré 5.6: Comparer les participants aux non-participants

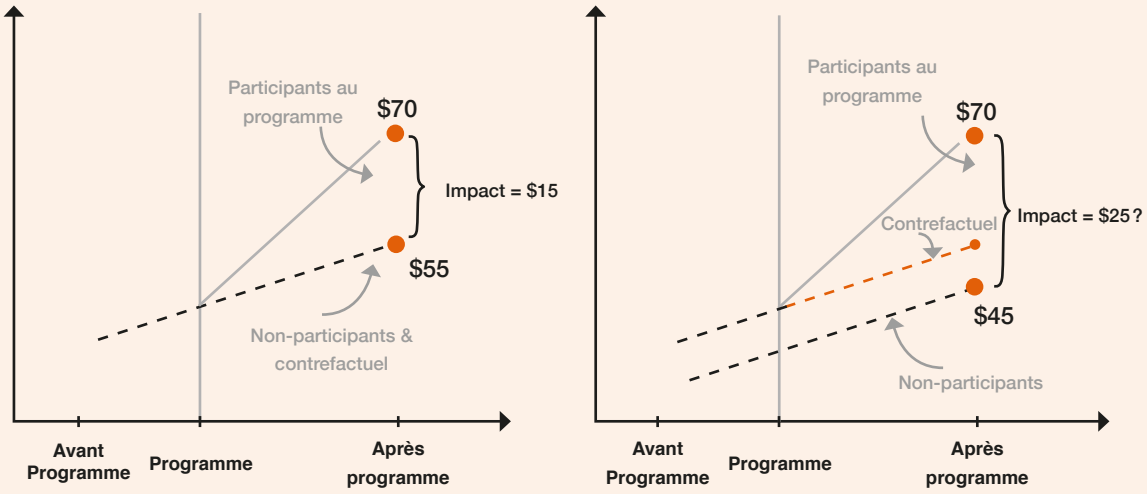
Parfois, bien que nous soyons capables d'identifier un groupe de comparaison, il se pourrait que nous n'ayons que des données disponibles sur quelques variables de résultats clés et aucune information sur les covariables, telles que le contexte socio-économique, les connaissances, les compétences, etc. Dans ces cas, on peut utiliser une simple méthodologie d'évaluation d'impact et comparer les résultats des participants et non-participants. Le contrefactuel est ainsi estimé par le résultat des personnes n'ayant pas participé au programme. Cependant, cette méthode n'est pas susceptible de produire des résultats crédibles ou des informations utiles sur l'effet réel de notre programme.

En particulier, si les non-participants (groupe de comparaison) diffèrent des participants (groupe de traitement) de façon pertinente aux résultats, ce type de comparaison ne sera pas valide car un biais de sélection sera introduit. Plus précisément, cette méthode repose sur deux fortes hypothèses.

Premièrement, il faut assumer que les participants et non-participants au programme ont eu, en moyenne, des résultats similaires au début du programme.

La partie de droite de la figure 5.8 illustre une situation où les participants avaient déjà un revenu plus élevé que les non-participants au début de l'intervention. Ce cas cause une surestimation du vrai impact de l'intervention

### FIGURE 5.8: COMPARER LES PARTICIPANTS AUX NON-PARTICIPANTS




Deuxièmement, on doit assumer qu'en l'absence d'intervention, les deux groupes se seraient développés de façon similaire dans le temps. Ceci nécessite l'hypothèse selon laquelle, en moyenne, les participants auraient réagi de la même façon que les non-participants à tous les facteurs externes. Notez que dans la situation illustrée sur la droite de la figure 5.8, cette hypothèse est vraie. La ligne en pointillés noirs, qui décrit comment les non-participants se sont développés dans le temps, et la ligne en pointillés rouges, qui décrit comment les participants auraient évolué en l'absence de l'intervention, restent parallèles dans le temps.

Pour obtenir des estimations d'impact correctes par cette méthode, les deux hypothèses doivent être maintenues. Un inconvénient majeur de cette méthode est que si l'on observe seulement

les résultats après l'intervention on ne peut tester aucune des deux hypothèses directement et, dans beaucoup de cas, elles pourraient ne pas être vraies. Si nous considérons le critère de sélection des jeunes pour l'intervention : il pourrait être sur la base du premier arrivé, premier servi. Dans ce cas, ceux qui ont un meilleur accès à l'information sur l'existence du programme, ceux qui vivent à proximité, ceux qui sont encouragés par leurs parents ou tout simplement ceux qui sont les plus motivés pour y participer sont plus susceptibles de finir par prendre part à l'intervention.

En résumé, comparer des participants et des non-participants à la fin d'une intervention sans une connaissance approfondie des variables du contexte qui permettraient d'employer des techniques plus sophistiquées, telles que la méthode PSM, n'est pas recommandé pour une évaluation d'impact.

et de non-participants avec des scores similaires, par exemple pour trouver une assez grande zone de soutien commun. De plus, la méthode PSM s'appuie sur l'hypothèse selon laquelle seuls les facteurs observables influencent la participation et les résultats (*hypothèse d'indépendance conditionnelle*). Ainsi, la méthode PSM devrait seulement être employée si l'on dispose d'une bonne connaissance des facteurs motivant la participation au programme et des résultats d'intérêt, et devrait être évitée si l'on peut s'attendre à ce que des caractéristiques non-observables affectent ces variables. Dans tous les cas, il faut être très prudent avant de décider combien de variables spécifiques et lesquelles doivent être sélectionnées pour l'estimation du score de propension.

## AVANTAGES

La méthode PSM est une méthodologie d'évaluation d'impact robuste qui, si ses hypothèses sont confirmées, peut aider à éliminer le biais de sélection et à produire des résultats valides à l'interne. Comme pour d'autres méthodes quasi-expérimentales, elle peut être appliquée sur la base de sources de données existantes et aucune affectation aléatoire à l'intervention est nécessaire.

En appariant sur la propension à recevoir un traitement, la méthode PSM réduit le nombre de dimensions sur lesquelles apparier les participants et les unités de comparaison à une seule et, en conséquence, rend l'appariement relativement facile et simple.

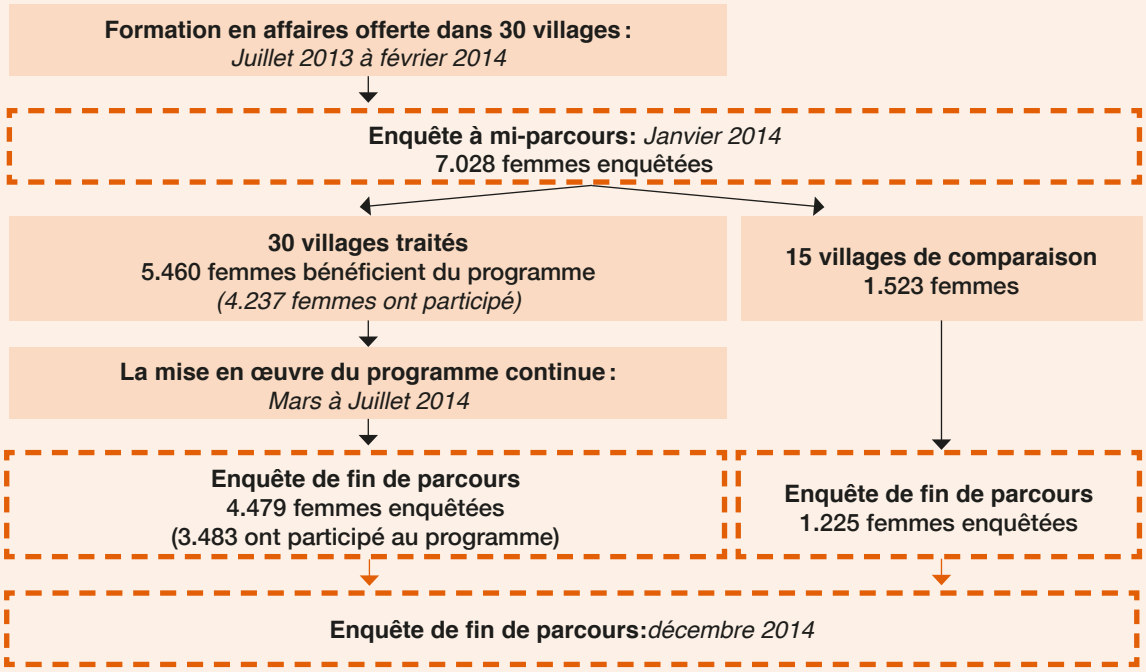
## LIMITES

- Appliquer la méthode PSM nécessite généralement de larges ensembles de données.
- L'appariement ne peut être effectué que sur des caractéristiques observables. Il y a donc toujours un risque que le biais de sélection des caractéristiques non-observables motivant la participation au programme affecte les résultats de l'évaluation.
- Appliquer la méthode PSM est statistiquement complexe et requiert un certain degré d'expertise.

## Encadré 5.7 : Autonomiser les jeunes femmes par la formation en affaires et professionnelle dans la Haute Égypte rurale – le programme Neqdar Nesharek

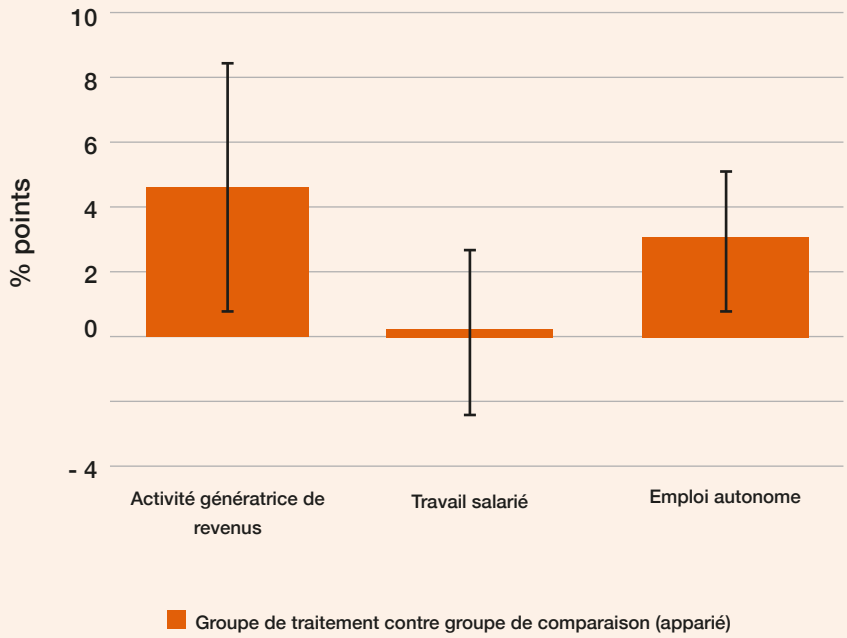
De 2013 à 2014 le Conseil de la Population (*Population Council*) a mis en œuvre le programme Neqdar Nesharek (qui signifie "Nous pouvons participer") dans la zone rurale de la Haute Égypte. Le programme a ciblé 4.500 jeunes femmes âgées de 16 à 29 ans, adoptant l'approche « espaces sûrs » des moyens de subsistance en adressant les besoins communautaires spécifiques des femmes vulnérables. L'intervention visait à autonomiser les jeunes femmes en leur apportant des compétences en affaires, des formations professionnelles et en les aidant à démarrer un commerce ou en les appuyant dans leur recherche d'emploi. Le programme de formation consistait en trois composantes principales : (1) formation en compétences de commerce, (2) formation professionnelle et (3) formation en compétences essentielles, droits et éducation civique.

### FIGURE 5.9 : CONCEPTION DE L'ÉVALUATION D'IMPACT (SIMPLIFIÉ)




L'intervention était accompagnée d'une évaluation d'impact pour évaluer l'effet du Neqdar Nesharek sur les résultats du marché du travail et les mesures d'autonomisation sociales des jeunes femmes. L'évaluation a utilisé un modèle PSM. Les impacts ont été calculés en appariant les femmes qui ont participé au programme avec des femmes des villages du groupe de comparaison ayant des caractéristiques socio-économiques similaires et en comparant les résultats clés du programme entre les deux groupes (voir figure 5.9).

## FIGURE 5.10 : IMPACTS SUR LES RÉSULTATS DU MARCHÉ DU TRAVAIL




L'évaluation a déterminé que le programme a eu un impact considérable sur l'autonomisation économique des participantes au programme, telle que mesurée par leur engagement dans des activités génératrices de revenus. Les participantes au programme étaient à 4,5 points de pourcentage plus susceptibles d'être engagées dans des activités génératrices de revenus que les femmes du groupe de comparaison. Comme illustré sur la figure 5.10, la plupart de l'impact positif était causé par une augmentation de l'engagement des participantes dans des activités d'emploi autonome. Par contre, le niveau de participation au travail salarié n'a pas changé de manière significative pour les femmes du groupe de traitement.

Source: OIT, 2017

# Modèle de régression par discontinuité (RDD, *Regression discontinuity design*)

Les modèles de régression par discontinuité (RDD) sont souvent utilisés lorsque l'éligibilité à une intervention sur le marché du travail est fondé sur un type de classement continu des bénéficiaires potentiels ou des candidats, par exemple, un âge limite.

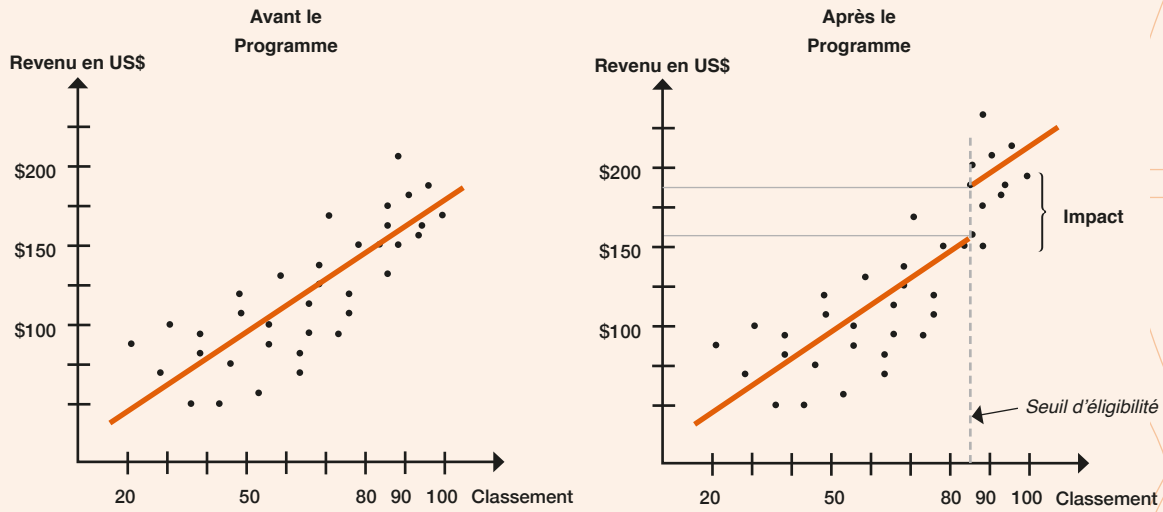
## COMMENT ÇA MARCHE ?

Le principe des modèles d'évaluation de la discontinuité (ou indice d'éligibilité) est que les personnes qui se classent juste au-dessus et juste en-dessous du seuil défini ne sont pas vraiment différentes les unes des autres ou, au moins, la différence peut être persistante dans les autres classements. Par exemple, les jeunes de 25 ans qui pourraient être éligibles à un programme de formation professionnelle pour jeunes ne sont pas susceptibles d'être très différents de leurs pairs de 26 ans qui peuvent ne plus être éligibles. Si nous nous trouvons dans une situation où quelques jeunes qui reçoivent le programme (ceux qui sont juste au-dessus du seuil) et d'autres qui ne le reçoivent pas (ceux juste en-dessous du seuil) ne sont pas fondamentalement différents les uns des autres, alors une comparaison des résultats de ces deux groupes permettrait d'analyser l'impact du programme.

La figure 5.11 illustre ce que nous pourrions découvrir en analysant l'impact d'une initiative de microcrédit pour jeunes. Le graphique à gauche indique qu'au moment de postuler au programme, ceux qui ont eu de meilleures notes avaient déjà tendance à avoir des revenus plus élevés.

Il pourrait y avoir plusieurs raisons pour cela, par exemple, que ceux qui ont un niveau d'éducation plus élevé gagnent déjà plus et leur éducation leur garantit aussi de meilleures notes. Ou ceux qui sont plus motivés à se lancer dans une affaire ont déjà plus un esprit d'entrepreneur, ce qui serait reflété dans leur revenus plus élevés,

**FIGURE 5.11 : EXEMPLE DE GRAPHE DE DISCONTINUITÉ**



et que leur motivation les a aussi aidés à convaincre le jury de les soutenir. Il y a plusieurs autres explications possibles qu'il n'est pas nécessaire de comprendre pour appliquer cette méthode.

Au lancement du programme, la banque de microfinance locale a décidé que le seuil pour l'obtention d'un prêt était le score de 85 et tous les candidats ont été admis ou rejetés en fonction de leur classement par rapport à ce seuil. Maintenant, il nous faut déterminer si le programme de microcrédit a eu un impact sur les revenus. Comme illustré sur la figure 5.11 (graphique de droite), nous assumons que ceux qui ont eu un score en-dessous de 85 ont les mêmes résultats qu'avant, alors que ceux qui ont eu 85 ou plus ont tous des résultats augmentés. Sur la base de cette information, il est possible d'identifier l'impact du programme, qui sera représenté par la différence des résultats (c'est-à-dire, la discontinuité de la relation linéaire) proche du seuil limite.

## QUAND UTILISER UNE RDD ?

Dans beaucoup de cas nous ne pouvons pas planifier l'évaluation pendant la conception du programme. Toutefois, il arrive parfois que nous puissions utiliser les règles de ciblage du programme pour obtenir un bon groupe de comparaison ex-post. Certains programmes utilisent un classement continu des bénéficiaires potentiels, tel que le score du test, la cote de crédit ou l'indice de pauvreté, et établissent un seuil limite pour l'admission au programme. Dans le cas des interventions sur le marché du travail pour jeunes, il y a souvent un âge-limite. Seuls les jeunes en-dessous d'un âge spécifié sont éligibles pour le programme. Cette règle d'éligibilité peut être utilisée pour effectuer une évaluation d'impact fondée sur un modèle RDD.

## AVANTAGES

- Le modèle RDD peut être appliqué ex-post, s'il y a suffisamment de données administratives.
- Il peut utiliser une règle existante d'affectation pour construire un groupe de comparaison valide et n'a donc pas besoin d'exclure un groupe éligible de l'intervention.

## LIMITES

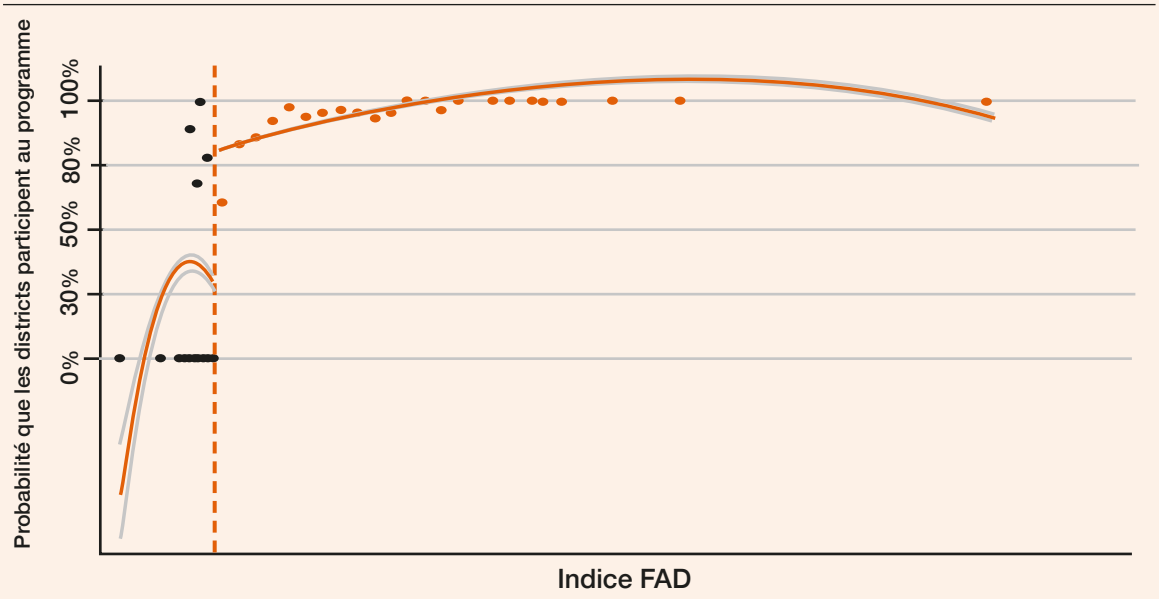
- L'exigence principale d'utilisation des modèles de discontinuité est que la participation au programme soit déterminée par une règle de ciblage explicitement spécifiée ; en d'autres termes, par une échelle ou un score continu. Pour que cette méthode marche, il nous faut plusieurs observations dans la zone immédiatement au-dessus et en-dessous du seuil limite pour avoir assez de jeunes à comparer les uns aux autres. À moins que l'évaluation soit faite sans données de référence ou qu'elle puisse utiliser des données du programme existantes, un modèle de discontinuité requiert les mêmes outils de collecte de données qu'un modèle de loterie et, donc, a un coût similaire.
- La valeur informative des résultats est limitée à l'échantillon autour du seuil limite, ce qui serait pertinent, par exemple, pour savoir si un programme devrait être intensifié pour inclure d'autres groupes d'âge ou régions.

## Box 5.8: Efficacité d'un programme d'activation des allocations par l'emploi sur les résultats du marché de l'emploi: Construyendo Perú

Les programmes de travaux publics constituent un outil de politique de plus en plus prisé dans les pays en voie de développement. De 2007 à 2011, le gouvernement du Pérou a mis en œuvre le programme Construyendo Perú dont l'objectif essentiel est le soutien des sans-emplois en situations de pauvreté. Le programme leur fournit des emplois temporaires et des formations en développement de compétences par le financement de projets d'investissements publics à forte utilisation de main-d'œuvre non qualifiée.

L'OIT a évalué les effets à moyen et long terme du programme en utilisant une approche de la régression par discontinuité. L'évaluation exploite une règle d'affectation intéressante du programme au niveau du district qui consiste à sélectionner les districts bénéficiaires en les classant selon leur indice FAD (Factor de Asignación Distrital). Le FAD est un indice composite qui combine l'information démographique à un indice de carence en développement humain et un indice de sévérité de la pauvreté. Ainsi, les districts avec un indice FAD au-dessus d'un certain seuil (c'est à dire ceux à forte pauvreté et carences de développement) ont été admis au programme et les districts en-dessous de ce seuil n'ont pas participé au programme. Ceci est un exemple d'un modèle flou de la régression par discontinuité. Comme on le voit dans la figure 5.12, les districts juste au-dessus du seuil limite étaient beaucoup plus susceptibles de participer au programme que ceux en-dessous de ce seuil.

FIGURE 5.12: DISCONTINUITÉ DE LA PROBABILITÉ DES DISTRICTS PARTICIPENT AU PROGRAMME




L'évaluation a ainsi tenté d'estimer l'impact causal du programme en comparant les résultats des personnes se situant autour du seuil limite de l'indice FAD. L'évaluation a trouvé qu'à moyen terme (trois à cinq ans) l'intervention a permis une augmentation de l'emploi et a réduit l'inactivité des femmes et des participants au programme les moins bien éduqués. Cependant, le programme n'a pas pu améliorer les opportunités des participants les moins-éduqués en termes de qualité de l'emploi et, en fait, a eu un impact défavorable sur les perspectives de qualité d'emploi des femmes et des personnes les plus éduquées (par exemple en augmentant la probabilité de l'emploi informel).

Source: Escudero, 2016

# Comparaisons simples : Avant et après

Parfois la randomisation n'est pas possible et, de plus, les conditions pour une évaluation quasi-expérimentale valide ne sont pas réunies, par exemple, si on ne peut pas trouver un groupe de comparaison adéquat avec des informations de référence disponibles et/ou si l'hypothèse de tendance commune ne peut être confirmée. Dans ces cas-là, il est conseillé de reconsidérer si cela vaut vraiment la peine d'effectuer une évaluation d'impact quantitative.

S'il n'est pas possible d'inclure un groupe de comparaison dans une évaluation d'impact, l'approche la plus élémentaire est de simplement comparer les résultats des participants du programme avant et après l'intervention. Cette simple approche peut donner une idée du changement qui s'est produit au cours d'une intervention, mais devrait être considérée plus comme un système d'observation que comme un moyen de fournir la preuve de l'impact causal d'une intervention, vu qu'il n'y a aucun moyen de savoir si le changement observé devrait être attribué à l'intervention en question ou à d'autres circonstances.

Prenant l'exemple d'un programme de formation, nous pourrions observer que le revenu mensuel des participants a augmenté de $50 avant l'intervention à $60 après et, ainsi, conclure que l'impact de l'intervention était de $10 par mois par personne (voir figure 5.13, graphique de gauche). Cependant, en l'absence de l'intervention, le niveau de revenu aurait quand même pu augmenter à cause d'un changement de circonstances (ex : la situation correspondant à un des scénarios illustré sur le graphique de droite de la figure 5.13), nous ne pourrions alors pas obtenir une estimation exacte de l'intervention.

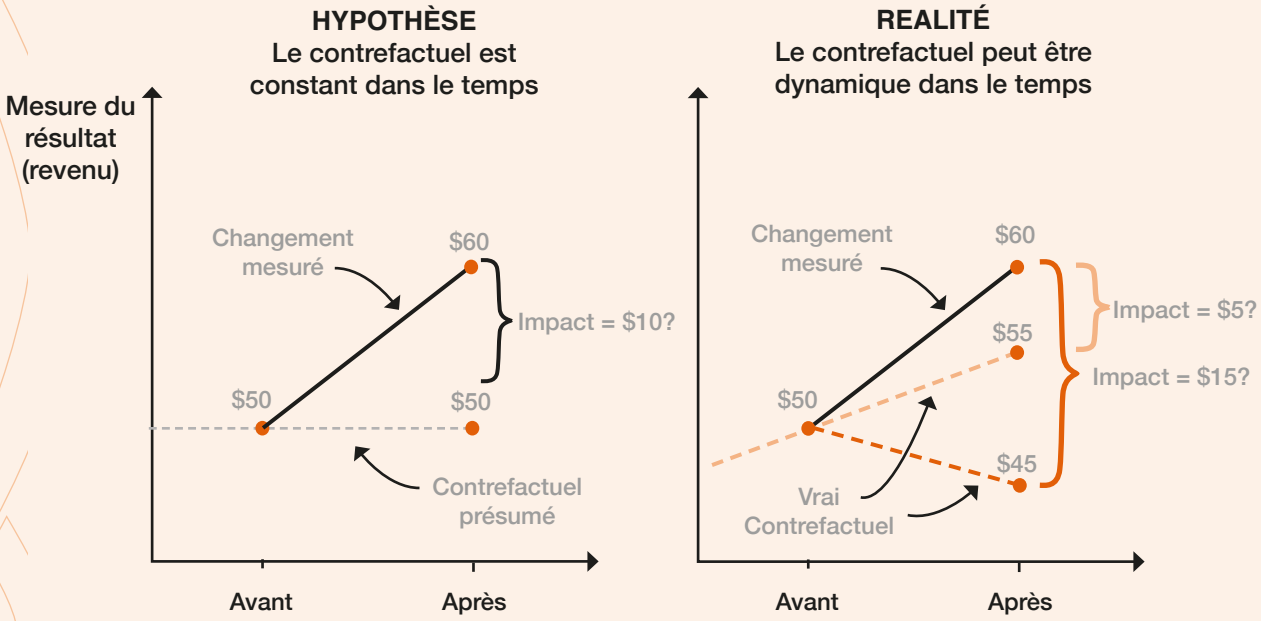
Vu que les vrais scénarios contrefactuels (les lignes pointillées sur la figure) ne peuvent pas être observés, nous n'avons aucun moyen de savoir si le cas applicable à une évaluation particulière est celui illustré sur la gauche de la figure 5.13 ou celui qui figure sur la droite. Il est ainsi impossible de se faire une idée pour savoir si l'impact estimé avec cette méthode est le vrai impact de notre intervention ou un impact « contaminé ».

**TIP**

**Les comparaisons avant et après sont parfois appelées *études par traceurs*, surtout dans le contexte des enquêtes normalisées administrées aux diplômés de l'éducation secondaire ou tertiaire ou des programmes de Formation (Technique et) Professionnelle. Les études par traceurs ont des moyens limités pour évaluer l'impact mais sont de puissants outils pour mesurer l'employabilité des diplômés et pour recueillir des feedbacks aux fins d'améliorer le programme d'étude. Pour un guide détaillé sur les études par traceurs, voir Schomburg (2016).**

Effectuer des comparaisons avant-et-après est pertient s'il y a des raisons de croire qu'en l'absence du traitement, les revenus resteraient en moyenne inchangés. Ceci pourrait s'appliquer aux interventions qui (a) sont administrées sur une *courte durée* (par exemple, des interventions de formation professionnelle courte, des services de conseil sur l'emploi ou des évènements visant à changer les attitudes des participants) et (b) sont conçues pour produire des *effets d'intérêt à court terme*. Cependant, les limites mentionnées ci-dessus persistent et, contrairement aux méthodes expérimentales et quasi-expérimentales bien mises en œuvre, de simples comparaisons de type avant-et-après ne peuvent pas être considérées comme des évaluations d'impact robustes. Leur niveau de consistance peut être amélioré, d'abord, en contrôlant les facteurs potentiels confusionnels dans un modèle de discontinuité (au lieu de simplement comparer les résultats) et/ou, ensuite, par l'application complémentaire de méthodes qualitatives afin de résoudre les mécanismes de causalité sous-tendant les changements observés sur les résultats.

## FIGURE 5.13 : COMPARER LES RÉSULTATS AVANT/APRÈS



# Améliorer la pertinence des évaluations d'impact quantitatives

Comme indiqué dans la section précédente, il y a un éventail de bonnes méthodes quantitatives pour fournir une réponse valide en interne à la question de base de l'évaluation : « Le projet a-t-il marché ? », c'est-à-dire, « A-t-il affecté les résultats d'intérêt tels qu'ils ont été définis dans notre intervention et nos objectifs de leçons à tirer ? ». La question est de savoir si l'intervention dans son ensemble a eu un impact important, mais elle n'est sûrement pas la seule question à se poser.

Pour une compréhension détaillée et holistique du comment et pourquoi les effets d'un programme d'emploi pour jeunes se sont déroulés, nous devons "creuser plus en profondeur". Bien comprendre l'hétérogénéité et les impacts des mécanismes de causalité qui provoquent les effets observés nous aide à tirer de précieuses leçons de l'évaluation et permet de mieux comprendre si un programme pourrait marcher dans d'autres contextes.

Ceci nous ramène à la question centrale de la validité interne et externe (voir le début de cette Note). Pour avoir une validité interne, les méthodes doivent être robustes et proprement mises en œuvre. Pour avoir une validité externe, il faut comprendre les facteurs contextuels pertinents du programme et leurs effets potentiels sur les résultats de l'évaluation. Il est difficile d'atteindre les objectifs en utilisant une seule méthode.

Par exemple, les évaluations expérimentales correctement appliquées sont en mesure de fournir des informations crédibles sur les impacts qui peuvent être uniquement attribués au projet, mais renseignent très peu sur leur réplicabilité dans d'autres environnements. Il est crucial de réaliser que les évaluations d'impact quantitatives nous disent « ce » qui est arrivé – l'effet moyen du traitement – mais ne nous disent pas « pourquoi ». Pour cela, l'application complémentaire de méthodes qualitatives est nécessaire.

## MESURER UNE VARIETÉ D'IMPACTS

D'abord, il peut s'avérer utile d'avoir une vue plus nuancée de l'impact actuel du programme. Ceci peut être partiellement accompli par les modèles quantitatifs décrits ci-dessus. Les questions pertinentes à se poser seraient :

- Les résultats varient-ils selon différents groupes de bénéficiaires (par exemple si les jeunes hommes en bénéficient, mais pas les jeunes femmes) ?
- Quel est l'impact de l'intervention à court terme et celui à long terme ?
- L'intervention a-t-elle des retombées positives ou négatives ? Il y a-t-il des résultats escomptés ou involontaires en-dehors du groupe cible actuel ?

Ensuite, on pourrait aussi être intéressé de tester les modèles transversaux, comment l'efficacité de l'intervention change selon les modifications que l'on apporte au modèle. Ces modèles permettent d'examiner les questions suivantes :

- Un modèle d'intervention est-il plus efficace qu'un autre ? On pourrait comparer des interventions alternatives (faire des dons aux start-ups de jeunes entrepreneurs au lieu de prêts, par exemple), ou tester la combinaison la plus efficace des composantes du programme (formation seule, formation plus stage, ou formation plus stage plus parrainage).
- Quel est le dosage le plus efficace de l'intervention ? Par exemple, doit-on fournir

20, 50 ou 100 heures de formation (voir tableau 5.3 pour plus de questions d'évaluation d'impact) ?

Les modèles transversaux aident à identifier plus que seulement l'impact global d'un projet : ils évaluent aussi les caractéristiques spécifiques de l'intervention et pourquoi elles marchent ou ne marchent pas. Par exemple, un programme peut fournir des aptitudes professionnelles et entrepreneuriales, telles que la menuiserie ou la couture, avec en plus un petit capital de départ pour se lancer en affaires. La provision de dotations en espèces pourrait revenir cher ou être difficile sur le plan politique, et le directeur de programme peut ainsi se demander si le capital de départ est nécessaire, ou si les participants sont mieux capables de mettre en œuvre leur formation sans le capital. Un modèle transversal peut aider à déterminer une conception de projet optimale dans ce cas. En pratique, cela nécessite une comparaison des résultats de différents groupes de traitement à un groupe de comparaison et à l'un par rapport à l'autre.

## COMBINER LES APPROCHES QUANTITATIVES ET QUALITATIVES

En outre, nous pourrions être intéressés à découvrir les canaux par lesquels l'impact se manifeste – c'est-à-dire, comprendre pourquoi et comment un impact se déroule. Par exemple, on pourrait vouloir répondre aux questions suivantes :

- Comment et pourquoi les choses se sont-elles déroulées telles qu'elles se sont déroulées ?
- Pourquoi un projet (ou une de ses parties) n'a pas marché comme prévu ?
- Que peut-on apprendre de l'échec ?

Si de telles pistes potentielles de recherche sont envisagées au niveau de la conception, les théories peuvent être partiellement testées par le biais des méthodes quantitatives énoncées ci-dessus. Pour atteindre cet objectif, les enquêtes doivent inclure des questions conçues pour capturer divers facteurs (résultats intermédiaires) au travers desquels l'impact est supposé opérer afin de vérifier si l'intervention affecte ces résultats intermédiaires. Cependant, plusieurs résultats d'interventions pour l'emploi de jeunes (telles que la santé mentale, autonomisation ou les relations au foyer) sont complexes et multidimensionnels et peuvent ne pas pouvoir être capturés avec des méthodes quantitatives. Les méthodes mixtes permettent de tracer les indicateurs qualitatifs et fournissent une analyse d'étude de cas sélectionnés pour faciliter une meilleure compréhension de la dynamique et des résultats de l'intervention. Par exemple, les interviews qualitatives structurées et semi-structurées, dans lesquelles les participants sont libres de relater des récits de leurs expériences de vie qui sont hors-catégorie des informations quantifiables, peuvent aider à cerner la compréhension de l'impact d'un programme (Bamberger et al., pp. 6–7; Leeuw et Vaessen, 2009).

Les méthodes de collecte de données qualitatives pourraient être particulièrement utiles pour collecter des informations sur la bonne mise en œuvre, ou pas, de l'intervention (voir Note 4 sur les évaluations de performance). Comprendre le processus de mise en œuvre est crucial pour la découverte de la façon dont la mise en œuvre de l'intervention affecte les résultats ainsi que pour l'interprétation correcte des constats pour pouvoir déterminer si les résultats décevants sont dus aux insuffisances du modèle d'intervention ou à celles de la mise en œuvre. De plus, les techniques qualitatives peuvent éclairer sur la raison de certains constats spécifiques et, particulièrement, sur la question de savoir pourquoi les effets ont été différents au sein de la population ciblée (par exemple, entre les jeunes des campagnes et ceux des villes ou entre jeunes femmes et jeunes hommes).

**Tableau 5.3 : Catégories de questions sur l'évaluation d'impact**

| Question | Description | Données additionnelles nécessaires | Exemples de résultats d'évaluation et interprétation |
|---|---|---|---|
| **Quel est l'impact global de l'intervention pour les résultats A, B, C du groupe X? … dans le contexte Y?** | Cette question est la question d'impact d'évaluation standard | n/a (collecte de données standard basée sur la méthode choisie) | L'impact moyen de l'intervention de formation sur le revenu des jeunes est de +$20 par mois. L'intervention a un impact positif sur le revenu des participants. |
| **Les résultats varient-ils selon les groupes de populations?** | Les interventions affectent souvent les groupes différemment (hétérogénéité des impacts). Ne mesurer que les impacts moyens peut cacher ces différences, donc il faut ventiler les impacts par groupe de population | • Information sociodémographique des participants et groupe de comparaison (âge, sexe, revenu, etc.)<br>• Pour pouvoir ventiler des résultats, le nombre de personnes à évaluer (l'échantillonnage) doit augmenter avec chaque catégorie d'information à analyser | L'augmentation moyenne des revenus est de $40 pour les garçons et $0 pour les filles. Les jeunes plus âgés bénéficient plus que les plus jeunes ($30 contre $10, en moyenne). Donc, l'intervention n'est pas efficace de façon égale pour tous les participants. Il nous faut savoir pourquoi les groupes bénéficient différemment et si possible adapter les ciblage et la structure du programme pour accommoder certains groupes |
| **Quel sont les impacts à court terme et à long terme du programme?** | Le changement de résultats peut ne pas être constant dans le temps. Les effets à court terme pourraient s'estomper, tandis que ceux à long terme pourraient ne pas se manifester avant plusieurs années après la fin de l'intervention | Données sur une période de temps prolongée (en pratique, ceci revient souvent à suivre les groupes de traitement et de comparaison pendant plusieurs années) | À la fin du programme, on observe un revenu moyen par participant de –$5 (perte) comparé aux contrôles. Deux ans après le programme, l'augmentation moyenne du groupe de traitement est de $20. Ceux ayant participé à la formation n'ont pas pu travailler autant que leurs pairs pendant la formation, donc ils ont perdu des revenus. Avec le temps, toutefois, la formation a porté ses fruits et les participants ont pu avoir des revenus plus élevés que leur contrepartie qui n'y a pas participé. Ne regarder que les résultats à court terme peut induire en erreur |
| **L'intervention a-t-elle des effets de retombées?** | L'intervention peut avoir des effets indirects sur les non-participants (positifs et négatifs) | • Données hors groupes de traitement et comparaison, incluant famille ou membres de la communauté<br>• Plusieurs groupes de traitement (un reçoit le modèle A, l'autre le modèle B, etc.)<br>• Le nombre de personnes évaluées doit être assez large pour créer plus d'un groupe de traitement ainsi qu'un groupe de comparaison | Non seulement les participants ont un revenu moyen accru de $20, mais leurs pairs ont aussi affiché une augmentation de $5. Les participants ont apparemment transféré leurs compétences à d'autres. |
| **Quel est le modèle d'intervention le plus efficace : le modèle A ou le modèle B?** | Il y a souvent une ambiguïté sur le meilleur modèle d'intervention possible. Les questions pourraient être axées sur la comparaison des interventions ou sur des combinaisons de composantes du programme | • Plusieurs groupes de traitement (un reçoit le modèle A, l'autre le modèle B, etc.)<br>• Le nombre de personnes évaluées doit être assez large pour créer plus d'un groupe de traitement ainsi qu'un groupe de comparaison | L'augmentation moyenne des revenus est de $5 pour ceux ayant reçu la formation et un stage. Ainsi, fournir une expérience de travail pratique en plus d'une formation semble considérablement améliorer l'impact. |
| **Quel est le dosage d'intervention le plus efficace?** | Le plus n'est pas toujours le mieux Trouver un équilibre de la quantité de services à rendre est important pour maximiser l'impact d'un côté et minimiser les coûts de l'autre | • Plusieurs groupes de traitement (un reçoit le modèle A, l'autre le modèle B, etc.)<br>• Le nombre de personnes évaluées doit être assez large pour créer plus d'un groupe de traitement ainsi qu'un groupe de comparaison | L'augmentation du revenu moyen est $0 pour ceux qui ont reçu 1 mois de formation, $20 pour ceux qui en ont reçu 3 mois, et $20 pour ceux qui en ont reçu 6 mois. Bien que 1 mois de formation soit insuffisant, 6 mois de formation n'a apporté aucun bénéfice par rapport à 3 mois de formation. La durée optimale de la formation semble être aux environs de 3 mois |
| **Pourquoi l'intervention (n') a-t-elle (pas) marché? Pourquoi n'a-t-elle marché que pour une partie de la population cible/que sur une certaine durée?** | En plus d'évaluer l'impact même, il est crucial de comprendre comment et pourquoi il s'est manifesté tel qu'observé | Les données quantitatives et qualitatives, sont idéalement triangulées pour établir des connexions causales raisonnables. Par exemple, les interviews détaillées de participants à la formation, de formateurs et d'employeurs | Les employeurs ont valorisé des compétences spécifiques pouvant être réalistiquement acquises par les participants en 3 mois. Prolonger la formation n'a apporté aucune valeur ajoutée aux compétences des employés. Les employeurs ne voulaient pas prolonger la formation car ils perdraient en temps de travail de ces employés.<br>La formation n'a entraîné une augmentation de revenu que pour les garçons, vu que les employeurs ont tendance à affecter les filles à des tâches différentes pour lesquelles les compétences acquises par la formation sont moins utilisées |

## Encadré 5.9: La reconstitution du processus

La reconstitution du processus requiert une analyse approfondie des divers évènements qui lient une intervention à un ou plusieurs résultats provisoires ou finaux, ainsi que ses relations causales. Souvent (mais pas toujours) les méthodes de reconstitution de processus cherchent à développer et tester les mécanismes théoriques pouvant être généralisés et élargis à d'autres interventions et contextes. En bref, la reconstitution de processus est appliquée comme suit :

### 1. Développer un mécanisme hypothétique causal sur la façon dont le changement se produit

La première étape est de faire une description narrative du processus qui va être évalué. Cela peut être une théorie du changement d'un projet, y compris les personnes et activités qui y sont impliquées. Il est important que le processus soit défini dans ses plus petits éléments individuels, lesquels doivent être à la fois essentiels au fonctionnement du processus et mesurables.

Par exemple : "Les enseignants donnent des formations professionnelles pour les jeunes sans emploi" ; "Les étudiants suivent les formations professionnelles" ; "Les étudiants acquièrent de nouvelles connaissances sur comment et où trouver des emplois" ; "Les étudiants recherchent des emplois de façon de plus en plus efficace" ; "Les étudiants ont une plus grande probabilité de trouver un emploi ".

Pour effectuer des liens de causalité plausibles entre les diverses parties de ce mécanisme, il est nécessaire d'identifier des explications alternatives possibles pour l'occurrence de chaque élément individuel et chercher la preuve qui confirme ou élimine ces explications. Par exemple : "Les étudiants ont acquis des connaissances sur comment et où chercher des emplois de manière indépendante" ; "Les étudiants ont trouvé des emplois à cause d'une amélioration du marché du travail local".

### 2. Définir et collecter les preuves requises

Après avoir défini le mécanisme, ou notre théorie, il nous faut définir la preuve empirique nécessaire pour l'analyse de chaque lien de la chaine causale. Ceci s'applique aussi bien à "notre" mécanisme qu'aux hypothèses alternatives. En conséquence, la preuve précédemment identifiée sera établie par la collecte de données primaires ou secondaires. Les sources de telles preuves peuvent, entres autres, être les interviews de parties prenantes, des documents du programme, des données d'enquête, des comptes-rendus de réunions, et des statistiques. Les preuves doivent être collectées de façon à pouvoir confirmer ou réfuter les diverses hypothèses. Il est toujours recommandé de trianguler les méthodes, c'est à dire utiliser différentes méthodes pour évaluer le même élément sous différents angles.

### 3. Évaluer les données fondées sur les preuves et tirer des conclusions

Les données fondées sur les preuves recueillies sont ensuite examinées dans une procédure similaire à celle utilisée pour le procès d'un crime. Dans la reconstitution du processus, nous cherchons à mettre en place une proposition qui offre assez de preuves pour raisonnablement assumer que chaque élément du mécanisme s'est déroulé à cause d'un autre élément avec lequel certains résultats ont été produits.

Il existe plusieurs tests pour évaluer la solidité des preuves de chaque hypothèse. Par exemple, le test de preuve probante ("*smoking gun*" test) est une preuve convaincante se référant directement au mécanisme en question. Ainsi, par exemple, une déclaration d'un participant à une formation professionnelle telle que : "Grâce à ce que j'ai appris dans la formation, je me sens bien plus confiant dans ma recherche d'emploi et j'envoie plus de demandes qu'avant" peut nous permettre d'établir que ce participant n'a pas amélioré son attitude de recherche d'emploi – qui pourrait être une variable de résultat intermédiaire mesurée quantitativement – à cause d'autres raisons.[4]

Lorsque l'on évalue les preuves d'hypothèses alternatives concourantes, il est important de garder à l'esprit que la solidité de la preuve globale d'un certain mécanisme n'est jamais plus solide que la preuve la plus faible d'un maillon de la chaîne. Finalement, basé sur les conclusions de cet exercice, le mécanisme hypothétique et les hypothèses alternatives seront confirmées ou infirmées.

[4] Pour plus de tests et de détails sur leur application, voir Bennett, 2010.

Au lieu de se substituer à une évaluation d'impact quantitatif, plusieurs des stratégies d'évaluation mentionnées ci-dessus peuvent contribuer à l'évaluation d'une intervention spécifique. Ainsi, employer une approche de méthodes mixtes nous permet de combiner les points forts et neutraliser les faiblesses des outils d'évaluation qualitatifs comme quantitatifs, permettant de créer de la sorte un modèle général d'évaluation plus robuste.

En pratique, employer un modèle de méthodes mixtes pour une évaluation d'impact requiert une collecte de données qualitatives, par exemple, par les interviews d'informateurs clés ou de groupes focus lors des visites de terrain, et de données quantitatives, telles que s'appuyer sur des données administratives, des enquêtes ou sources de données secondaires, comme des enquêtes ménages (voir aussi le tableau 3.5 de la Note 3).

Les modèles d'évaluation utilisant des méthodes mixtes sont très proches des évaluations d'impact fondées sur la théorie, et les informent aussi. Comme le remarquent White and Phillips (2012), les évaluations d'impact fondées sur la théorie visent à établir des liens de causalité « en collectant des preuves pour valider, invalider ou réviser les explications hypothétiques, avec le but de documenter les liens de la chaîne causale en question ». Elles cherchent souvent à combiner toutes les preuves quantitatives et qualitatives disponibles pour établir hors de tout doute raisonnable qu'une interprétation a impacté ses participants. L'encadré 5.9 présente la méthode de reconstitution du processus comme méthodologie d'évaluation d'impact fondée sur la théorie.

## POINTS CLÉS

1. **Les évaluations d'impact adressent les questions de cause-à-effet pour déterminer si, et dans quelle mesure, une intervention a causé un changement observable.** Comprendre l'impact impose d'isoler les effets de l'intervention des autres facteurs qui influencent les résultats des bénéficiaires.

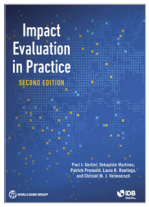
2. **Quantifier les impacts des interventions nécessite l'estimation du contrefactuel**; c'est-à-dire, ce qui serait arrivé aux bénéficiaires en l'absence de l'intervention. Pour ce faire, la plupart des modèles d'évaluation quantitative de l'impact ont recours à un groupe de comparaison qui a autant de caractéristiques similaires avec les bénéficiaires que possible.

3. **Les modèles d'évaluation d'impact observables incluent les méthodes de différences et d'appariement.** Ils peuvent être appliqués à un large éventail de contextes et fondés sur des sources de données secondaires, mais pour certaines interventions, ces méthodes ne sont pas suffisantes pour estimer les impacts de façon crédible. Les modèles expérimentaux qui se fondent sur un degré de randomisation peuvent produire des estimations d'impact très crédibles mais peuvent être onéreuses et difficiles à mettre en œuvre pour certaines interventions.

4. **Pour maximiser la compréhension du «pourquoi» les interventions ont-elles marché, ou n'ont pas marché, utilisez des approches de méthodes mixtes** qui emploient des données qualitatives et quantitatives et utilisent plusieurs méthodologies d'analyse.

## RESSOURCES CLÉS

- Card, D.; Ibarrarán, P.; Villa, J.M. 2011. *Building in an evaluation component for active labor market programs: A practitioner's guide*, IZA Discussion Paper No. 6085 (Bonn, Institut des Études sur le Travail (IZA)).




- Gertler, P.J.; Martinez, S.; Premand, P.; Rawlings, L.B.; Vermeersch, C.M., 2016. *Impact evaluation in practice*, Second Edition (Washington DC, Banque interaméricaine de développement et Banque mondiale).


- Duflo, E.; Glennerster, R.; Kremer, M. 2006. *Using randomization in development economics research: A toolkit*, BREAD Working Paper No. 136 (Bureau de Recherche et d'Analyse économique du développement).


- Khandker, S.R.; Koolwal, G.B.; Samad, H.A. 2010. *Handbook on impact evaluation: Quantitative methods and practices* (Washington, DC, Banque internationale pour la reconstruction et le développement (BIRD) et Banque mondiale).

# RÉFÉRENCES


- Bamberger, M.; Rao, V.; Woolcock, M. 2010. *Using mixed methods in monitoring and evaluation: Experiences from international development*, Policy Research Working Paper No. 5245 (Washington, DC, Banque mondiale).
- Bennett, A. 2010. "Process tracing and causal inference", in H. Brady and D. Collier (eds.) *Rethinking social inquiry: Diverse tools, shared standards* (2nd edn) (Lanham, MD, Rowman & Littlefield), pp. 207–220.
- Card, D.; Kluve, J.; Weber, A. 2009. *Active labour market policy evaluations: A meta-analysis*, IZA Discussion Paper No. 4002 (Bonn, Institute for the Study of Labour (IZA)).
- Card, D.; Ibarrarán, P.; Villa, J.M. 2011. *Building in an evaluation component for active labor market programs: A practitioner's guide*, IZA Discussion Paper No. 6085 (Bonn, Institute for the Study of Labour (IZA)).
- Duflo, E.; Glennerster, R.; Kremer, M. 2006. *Using randomization in development economics research: A toolkit*, BREAD Working Paper No. 136 (Bureau for Research and Economic Analysis of Development).
- Escudero, V. 2016. *Workfare programmes and their impact on the labour market: Effectiveness of Construyendo Perú*, Research Department Working Paper No. 12 (Genève, OIT).
- Fiala, N. 2015. *Access to finance and enterprise growth: Evidence from an experiment in Uganda*, Employment Working Paper No. 190 (Genève, OIT).
- Gertler, P.J.; Martinez, S.; Premand, P.; Rawlings, L.B.; Vermeersch, C.M., 2016. *Impact evaluation in practice*, Second Edition (Washington DC, Inter-American Development Bank and World Bank).
- International Labour Organization (ILO). 2017. *Empowering young women through business and vocational training: Evidence from rural Upper Egypt*, Taqeem Impact Brief Series, Issue 10 (Genève).
- Khandker, S.R.; Koolwal, G.B.; Samad, H.A. 2010. *Handbook on impact evaluation: Quantitative methods and practices* (Washington, DC, Banque internationale pour la reconstruction et le développement (BIRD) et Banque mondiale).
- Leeuw, F.; Vaessen, J. 2009. *Impact evaluations and development: NONIE guidance on impact evaluation* (Washington, DC, Network of Networks on Impact Evaluation).
- Schomburg, H. 2016. *Carrying out tracer studies: Guide to anticipating and matching skills and jobs*, Vol. 6 (Luxembourg, ETF, CEDEFOP, ILO).
- White, H.; Phillips, D. 2012. *Addressing attribution of cause and effect in small n impact evaluations: Towards an integrated framework*, Working Paper 15 (New Delhi, International Initiative for Impact Evaluation).

## Étude de cas

# ÉVALUER LA CROISSANCE DES MICRO-ENTREPRISES RURALES AU MOYEN DE DIFFÉRENTES MÉTHODES D'ÉVALUATION

Mention : Ce qui suit est une étude de cas fictive.
Toute information contenue dans cette étude a été inventée pour les besoins d'apprentissage.

## Objectifs d'apprentissage

À la fin de cette étude de cas, les lecteurs pourront démontrer les acquis d'apprentissage suivants :

- Identifier les méthodes d'évaluation d'impact sans que leur soit indiqué la méthode qui a été employée
- Explorer le défi de produire des estimations de l'impact causal d'un programme de développement, et les diverses façons d'estimer les impacts en utilisant des modèles de groupes de comparaison.
- Développer une connaissance intuitive de quand et comment les méthodes d'évaluation d'impact produiront des résultats biaisés, en apprenant le concept de biais de sélection et en comprenant que la valeur des modèles de groupes de comparaison dépend de leur capacité à éliminer le biais de sélection.

## Introduction et contexte de l'étude de cas

Les microentreprises sont essentielles dans les zones rurales où les options d'emploi formel sont limitées, aussi bien pour leur capacité à fournir de l'emploi informel que pour assurer la sécurité économique du foyer des petits entrepreneurs. Cependant, une fois qu'une affaire est lancée, sa croissance pose plusieurs défis.

Que peut-on faire pour aider le développement des commerces en milieu rural ? Le programme de formation pour l'autonomisation économique rurale (TREE, *Training for Rural Economic Empowerment*) teste certaines de ces contraintes pour comprendre quels types de services financiers et de formations impactent la croissance des entreprises, pour qui et pourquoi.

L'Organisation Internationale du Travail (OIT) a effectué la formation testée ici en utilisant sa méthodologie TREE, une approche de développement dont l'objectif est que les femmes et les hommes vivant dans la pauvreté puissent acquérir les compétences et la connaissance nécessaires pour améliorer leurs revenus et assumer un rôle plus actif dans le développement de leurs communautés. De plus, une organisation locale de microfinance a effectué des prêts individuels de 200 USD à un taux d'intérêt annuel réduit de 20 pour cent (le taux standard étant 25 pour cent).

Cette étude de cas vise 400 propriétaires de microentreprises rurales à qui il a été donné l'opportunité de participer à un programme de formation professionnelle et de recevoir des prêts. Au total, 144 des 400 entrepreneurs ont reçu la formation et le prêt.

## Comparer différentes méthodes d'évaluation d'impact : TREE a-t-il marché ?

Le programme TREE a-t-il marché ? Le programme a-t-il augmenté les profits des entrepreneurs ? De quoi a-t-on besoin pour mesurer si un programme a marché, ou s'il a eu un impact ?

En général, se demander si un programme marche revient à se demander si le programme a atteint ses objectifs de changer certains résultats pour ses participants, et s'assurer que ces changements ne sont pas causés par d'autres facteurs. Il nous faut simultanément démontrer que, si le programme n'avait pas été mis en œuvre, les changements observés ne se seraient pas produits (ou seraient différents). Mais comment peut-on savoir ce qui se serait passé ? Mesurer ce qui se serait passé en l'absence du programme nous demande d'entrer dans un monde imaginaire dans lequel le programme n'a jamais été offert à ces participants. Les résultats de ces mêmes participants dans ce monde imaginaire sont ce que nous appelons le contrefactuel. Vu que l'on ne peut pas observer le vrai contrefactuel, le mieux que l'on puisse faire est de l'estimer en l'imitant.

Le défi principal de l'évaluation d'impact d'un programme est de construire ou d'imiter le contrefactuel. Nous accomplissons typiquement cela en sélectionnant un groupe de personnes qui ressemble autant que possible aux participants du programme mais qui n'a pas participé au programme. Ce groupe s'appelle le groupe de comparaison, et ce dernier, idéalement, diffère du groupe de bénéficiaires seulement dans le fait que ses membres n'ont pas participé au programme.

Nous estimons alors « l'impact » comme étant la différence observée à la fin du programme entre les résultats du groupe de comparaison et ceux du groupe de participants au programme.

De façon importante, la précision de l'estimation de l'impact est fonction de la mesure dans laquelle le groupe de comparaison est arrivé à imiter le contrefactuel. Ainsi, la méthode utilisée pour sélectionner le groupe de comparaison est une décision essentielle dans la conception d'une évaluation d'impact.

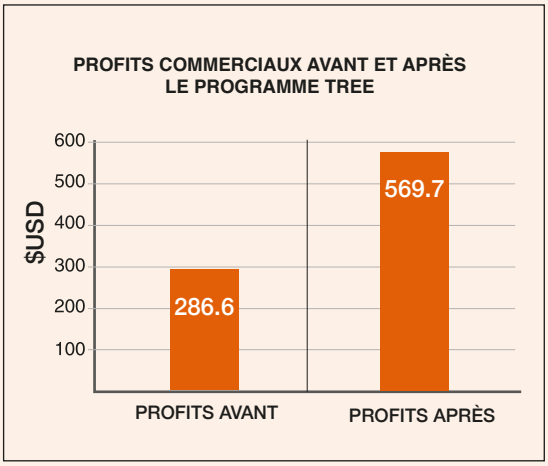
Ceci nous ramène à nos questions : Le projet a-t-il marché ? Quel était son impact sur les résultats soumis à l'évaluation, soit les profits des commerces ?

Dans notre cas, l'intention du programme est principalement d' « améliorer la croissance de l'entreprise », et les profits (mesurés en USD) sont l'indicateur-clé du résultat. Donc, lorsque l'on se demande si le projet a marché, on demande en fait s'il a augmenté les profits des commerces. L'impact est la différence entre les profits réalisés après que les entreprises aient été exposées à l'intervention et ce que leurs profits auraient été si l'intervention n'avait jamais existé.

Quels groupes de comparaison et méthodes d'évaluation d'impact peut-on utiliser ? Les experts (fictifs) suivants illustrent diverses méthodes d'évaluation d'impact. Pour référence, le Tableau 5.2 au début de cette note présente une vue d'ensemble des différentes méthodes d'évaluation.

# 1ère partie – Communiqué de presse : le programme "Formation pour l'Autonomisation Économique Rurale" (TREE) aide les commerces à se développer

TREE célèbre le succès de son programme. Il a considérablement progressé dans son objectif d'aider la croissance des entreprises par la provision de prêts et la formation professionnelle. L'accomplissement du programme TREE démontre qu'apporter aux entrepreneurs une formation professionnelle, en combinaison avec des prêts pour alléger les contraintes de capital, peut produire des gains considérables.




Juste avant que le programme soit lancé, les commerces réalisaient en moyenne des profits de $286. *Mais après juste quelques mois passés dans le programme, les profits de ces commerces ont doublé !*

## Thèmes de discussion

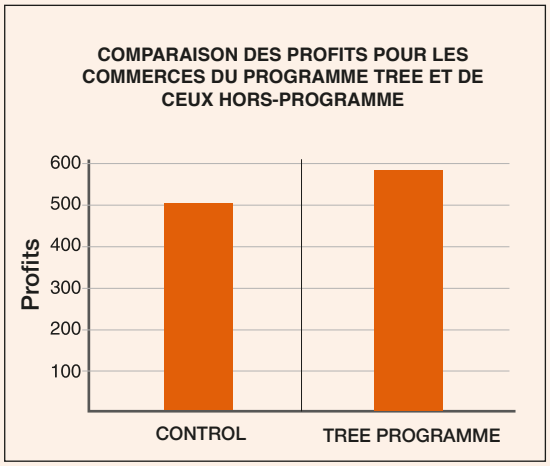
1. Quel type d'évaluation ce communiqué de presse implique-t-il ?
2. Qu'est-ce qui représente le contrefactuel ?
3. Quels sont les défis de ce type d'évaluation ?

# 2ème partie – Opinion : le projet "Formation pour l'autonomisation économique rurale" n'est pas à la hauteur

Avec un rayonnement estimé à 6 millions de stagiaires, un réseau toujours croissant de plus de 17 000 formateurs et 200 maître-formateurs dans 2 500 institutions partenaires, TREE est l'un des plus larges systèmes de formation axés sur l'appui aux micro- et petites entreprises (MPE) actuellement sur le marché. Mais les profits de ses entreprises doublent-ils réellement de volume, comme suggéré dans le premier exemple ? Des données récentes nous suggèrent le contraire.

▶ Une équipe indépendante d'évaluateurs a été engagée pour vérifier ces données.

L'équipe a comparé les profits des entreprises de TREE à ceux d'autres entreprises dans des villages alentours. Elle a estimé que les profits des commerces TREE augmentent à peine de 64 USD, et non 286 USD comme originellement estimé. Ce qui représente seulement une augmentation des profits de 12% après 6 mois du programme TREE apparié aux prêts. Il semblerait que les estimations de revenus aient été fortement surestimés et que les assurances de l'OIT sur le succès du programme étaient fausses.




## Thèmes de discussion

1. Quel type d'évaluation cet article d'opinion implique-t-il ?
2. Que représente le contrefactuel ?
3. Quel sont les défis de ce type d'évaluation ?

## 3ème partie – Lettre à la rédaction : les évaluateurs indépendants devraient évaluer équitablement et précisément

Il y a eu plusieurs rapports injustes dans la presse concernant les programmes mis en œuvre par l'OIT. Un article récent d'un évaluateur indépendant déclare que'en réalité, TREE n'aide pas la croissance des entreprises. Cependant, leur analyse utilise des métriques inadéquates pour mesurer l'impact. Elle compare les profits des entreprises de TREE à ceux d'autres entreprises dans le village – sans prendre en compte le fait que TREE cible ceux dont les profits sont particulièrement bas au départ. Si TREE engageait simplement les plus grosses entreprises dans leurs programmes, et les comparait à leurs plus petites contreparties, il pourrait déclarer sa réussite sans fournir une seule séance de formation ou un seul prêt. Mais ce n'est pas ce que fait TREE. Et, pour être réaliste, TREE ne s'attend pas à ce que ses plus petites entreprises dépassent les grosses entreprises dans le village. Il essaie simplement de causer une amélioration de l'état actuel.

En conséquence, l'indicateur devrait être l'*amélioration* des profits – et non le niveau final des profits. Lorsque nous avons refait l'analyse en utilisant la mesure de résultats la plus appropriée, les entreprises TREE ont eu une amélioration deux fois plus grande que celle des entreprises hors-TREE (augmentation de profit à 283 USD comparé à 162 USD). Si les évaluateurs avaient pensé à regarder les résultats les plus appropriés, ils auraient reconnu l'incroyable succès de TREE. Peut-être qu'ils devraient s'inscrire à quelques formations de TREE eux-mêmes.

## Thèmes de discussion

1. Quel type d'évaluation cette lettre implique-t-elle ?
2. Que représente le contrefactuel ?
3. Quels sont les défis de ce type d'évaluation ?

## 4ème partie : concevoir votre propre évaluation pour évaluer l'impact de TREE

Comme nous l'avons vu dans cette étude de cas, il y a des défis comme des réserves en ce qui concerne les trois méthodes d'évaluation détaillées ci-dessus. C'est maintenant à votre tour de concevoir une évaluation d'impact pour le programme TREE, en assumant que le programme n'a pas encore été mis en œuvre.

Pour commencer, assumez que votre équipe de recherche a mené son enquête sur plusieurs milliers de microentreprises à partir d'un recensement des commerces et a sélectionné 1.600 entrepreneurs qu'elle soumettra à l'évaluation. Tous ces entrepreneurs ont exprimé un intérêt à recevoir la formation de l'OIT et à participer au programme de prêt. Cependant, dû aux contraintes de ressources, votre chef de projet vous dit que le programme de formation et de prêt ne peut être offert qu'à un maximum de 800 entreprises.

Nous assumons aussi que le résultat d'intérêt clé demeure les « profits d'entreprise ».

1. Comment concevrez-vous l'évaluation ? En particulier, comment sélectionnerez-vous un groupe de comparaison ?
2. Quand est-ce que votre équipe de recherche collectera des données et de quelles entreprises (toutes les 1.600 ou seulement un sous-ensemble) ?
3. Pourquoi pensez-vous que ce soit une méthode d'évaluation d'impact fiable qui pourra surmonter certaines ou toutes les failles des trois méthodes discutées ci-dessus ?

International Labour Organization
Youth Employment Programme
4 route des Morillons
CH-1211 Genève 22
Switzerland

youth@ilo.org
www.ilo.org/yep
www.ilo.org/taqeem

La version française du guide a été traduite avec le soutien du l'Agence Française de Développement

AFD
AGENCE FRANÇAISE DE DÉVELOPPEMENT

978-92-2-133471-2

9 789221 334712