Tester les hypothèses

Bienvenue dans le chapitre consacré à la phase *"Tester les hypothèses"* !

Cette phase examine si l'intervention retenue contribue à des changements mesurables du comportement ciblé, en s'attaquant aux principales barrières comportementales (identifiés lors des phases *Définir* et *Explorer*).

Les enseignements générés à ce stade permettent de déterminer si une intervention doit être ajustée, mise à l'échelle ou abandonnée, et constituent le fondement de la phase Scale, au cours de laquelle les données probantes sont traduites en actions élargies.

Contrairement aux chapitres précédents, qui se concentrent sur des étapes et des outils spécifiques pour appliquer les sciences comportementales, ce chapitre est organisé légèrement différemment.

Réaliser des évaluations rigoureuses visant à mesurer l'efficacité des interventions comportementales est une tâche complexe qui dépasse la portée de ce guide. Toutefois, ce chapitre propose des orientations essentielles pour comprendre la valeur des évaluations d'impact, les difficultés liées à l'établissement de la causalité, ainsi que les principaux éléments et décisions impliqués dans la conception et la coordination d'expérimentations. L'objectif est de fournir une compréhension de base du test d'hypothèses d'intervention afin de faciliter une collaboration plus efficace avec les spécialistes de l'évaluation.

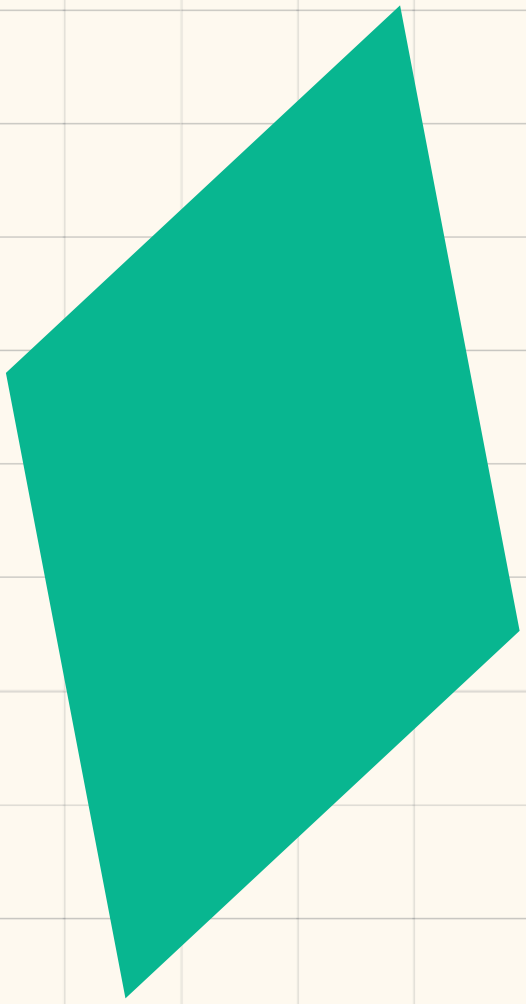
Compte tenu de la diversité des approches d'évaluation disponibles, ce chapitre présente les méthodes les plus couramment utilisées dans les sciences comportementales. Nombre de ces activités nécessitent une expertise technique avancée. Il est donc recommandé de consulter des spécialistes de l'évaluation et d'utiliser ce chapitre pour comprendre le type d'expertise requis, les questions pertinentes à poser aux spécialistes, ainsi que les défis susceptibles de se présenter. Pour celles et ceux souhaitant mener directement des évaluations, des manuels supplémentaires et ressources pratiques sont disponibles tout au long du chapitre, ainsi que dans la section « En savoir plus » à la fin de cette phase.

Pourquoi tester les hypothèses ?

Cette section explore les quatre facteurs suivants :

1. **L'importance de l'évaluation**
2. **Le défi de la causalité**
3. **Le cadre contrefactuel :** comprendre ce qui se serait passé autrement
4. **Comment la randomisation crée la norme de référence pour les contrefactuels**

Cet ensemble de notions constitue la base permettant de sélectionner la méthode d'évaluation appropriée et d'utiliser les données probantes pour renforcer les interventions.



1. L'importance de l'évaluation

Imaginons une situation dans laquelle un patient est malade et un médecin lui propose un nouveau médicament. Lorsque le patient demande si ce médicament est efficace, le médecin répond : « Nous ne l'avons pas testé formellement, mais il fonctionne probablement. Plusieurs patients qui l'ont pris semblaient aller mieux, et notre équipe en est assez convaincue. » Il faut alors se demander si le patient accepterait de prendre ce médicament. La plupart des gens refuseraient, et à juste titre.

Pourtant, dans le domaine des programmes sociaux et des interventions comportementales, l'inverse est fréquemment observé. Des programmes sont mis en œuvre sur la base de bonnes intentions, de cadres théoriques prometteurs et d'histoires anecdotiques de réussite, mais sans preuves rigoureuses de leur impact. Pourquoi, dès lors, les interventions sociales sont-elles soumises à des exigences moins strictes que celles de la médecine, alors que toutes deux visent à améliorer le bien-être humain ?

Quand les histoires de réussite sont trompeuses

Le développement international est marqué par une longue histoire de programmes bien accueillis, attirant l'attention, mobilisant des financements et semblant promettre des résultats révolutionnaires — jusqu'à ce que leur impact soit évalué de manière rigoureuse. L'exemple le plus célèbre est celui de la microfinance. Apparue dans les années 1980, la microfinance a été saluée comme un outil transformateur de réduction de la pauvreté. En offrant de petits prêts aux personnes vivant dans des contextes à faibles revenus, elle visait à encourager l'entrepreneuriat et la croissance économique. Le modèle s'est diffusé rapidement et a suscité un enthousiasme mondial.

Au fil du temps, toutefois, les évaluations rigoureuses ont révélé un tableau plus nuancé. Si la microfinance a effectivement amélioré l'accès au crédit, son impact sur la réduction de la pauvreté, la mobilité économique et le bien-être à long terme est resté moins évident. Des études de recherche approfondies ont mis en évidence l'alourdissement de la dette chez certains emprunteurs, une mise à l'échelle limitée, des performances économiques modestes et une faible contribution à la lutte contre la pauvreté structurelle [1] [2] [3] . Des articles et des livres, tels que « Big Money Backs Tiny Loans That Lead to Debt, Despair and Even Suicide » et « More Than Good Intentions » (Plus que de bonnes intentions), illustrent la désillusion qui a suivi, ainsi que le rôle crucial des évaluations pour révéler ce que les anecdotes ne pouvaient pas montrer.

La microfinance n'est pas un cas isolé. D'autres programmes très médiatisés, tels que PlayPump[4] , Un ordinateur portable par enfant,[5] et le projet Millennium Villages[6] , ont suscité un enthousiasme initial, mais n'ont pas confirmé leur impact lorsqu'ils ont été évalués de manière rigoureuse.

1 *John, B. (14 novembre 2024). Défis et limites de la microfinance dans la réduction à grande échelle de la pauvreté et la création d'emplois [Document de travail].*

2 *Akbari, M., Nikijoo, I., Khodapanah, B., Foroudi, P., & Padash, H. (2025). Quarante ans de recherche sur la microfinance et son impact sur les consommateurs : examen et programme de recherche à l'aide du cadre ADO-TCM. International Journal of Consumer Studies, 49(4), e70101.*

3 *Blanc, J. (2014). Microfinance, dette et surendettement : jongler avec l'argent, Isabelle Guérin, Solène Morvant-Roux et Magdalena Villarreal (dir.). Éditions Routledge, Londres, Royaume-Uni, 2014, 316 pages. Revue internationale de l'économie sociale : recma, (334), 122-124.*

4 UNICEF. (2007). *Évaluation du système d'approvisionnement en eau PlayPump® en tant que technologie appropriée pour les programmes d'eau, d'assainissement et d'hygiène* https://www-tc.pbs.org/frontlineworld/stories/southernafrica904/flash/pdf/unicef_pp_report.pdf

5 Cristia, Julian, Ibarraran, Pablo, Cueto, Santiago, Santiago, Ana et Severin, Eugenio, Technologie et développement de l'enfant : données issues du programme « Un ordinateur portable par enfant » (février 2012). Document de travail de la BID n° IDB-WP-304, disponible sur SSRN : https://ssrn.com/abstract=2032444.

6 Mitchell, S., Gelman, A., Ross, R., Chen, J., Bari, S., Huynh, U. K., ... Sachs, J. D. (2018). *Le projet Millennium Villages : une évaluation rétrospective, observationnelle et finale. The Lancet Global Health, 6*(5), e500–e513. https://doi.org/10.1016/S2214-109X(18)30065-2

Le piège des suppositions : pourquoi les bonnes intentions ne suffisent pas

Même avec les meilleures intentions, nous avons tendance à formuler des suppositions qui peuvent orienter les interventions dans la mauvaise direction. Ce « piège des suppositions » opère à plusieurs niveaux :

- **Supposer que l'on comprend le problème**. Souvent, les enjeux sont diagnostiqués à partir de perspectives individuelles plutôt qu'à partir d'une compréhension approfondie de l'expérience vécue par les communautés. Ce qui semble évident de l'extérieur peut ignorer ou manquer des éléments essentiels de contexte et de complexité.
- **Supposer que l'on sait ce qui fonctionnera**. Sur la base d'une expertise personnelle ou d'expériences passées, certaines personnes peuvent être convaincues que certaines approches réussiront, sans disposer de preuves suffisantes. Pourtant, il est fréquent que les facteurs susceptibles d'influencer les résultats n'aient pas encore été observés.
- **Supposer que la mise en œuvre suivra le plan prévu**. Il est courant de sous-estimer les défis pratiques et de surestimer la capacité des interventions à suivre fidèlement leur conception lorsqu'elles sont appliquées dans des contextes réels.
- **Supposer que quelques témoignages positifs prouvent le succès**. Lorsque des retours favorables sont recueillis ou que l'on observe des moments positifs, il est fréquent de généraliser ces expériences, en leur accordant un poids excessif dans l'évaluation du résultat global.
- **Supposer que la corrélation signifie la causalité**. Lorsque la situation s'améliore après une intervention, il est naturel d'attribuer ce changement à notre action, même lorsque d'autres facteurs peuvent en être responsables.

Ces suppositions ne proviennent ni de la négligence ni de l'incompétence : elles découlent du fonctionnement même de notre cognition. Comme mentionné dans les modules précédents, l'esprit humain cherche des modèles, préfère les informations qui confirment ses croyances, et construit des récits cohérents — même lorsque la réalité est plus complexe. Si ces tendances nous sont souvent utiles, elles peuvent nous induire en erreur lorsqu'il s'agit d'évaluer des interventions sociales complexes.

Sans évaluation systématique, ces suppositions ne sont jamais remises en question. Cela peut conduire à investir dans des programmes qui paraissent efficaces, mais qui ne produisent en réalité aucun changement significatif, ou pire, qui peuvent générer des effets indésirables. L'évaluation offre le processus structuré nécessaire pour aller au-delà des suppositions et comprendre l'impact réel du travail accompli.

La proposition de valeur : pourquoi l'évaluation vaut l'investissement

Les évaluations ne sont pas de simples exercices académiques : elles génèrent une valeur concrète :

- **Optimisation des ressources :** Dans des contextes où les ressources sont limitées, l'évaluation permet d'orienter les financements vers des interventions dont l'efficacité est démontrée.
- **Correction de trajectoire :** Une évaluation réalisée en temps opportun aide à identifier et à résoudre les problèmes de mise en œuvre avant le passage à l'échelle, évitant ainsi la diffusion de stratégies inefficaces.
- **Renforcement de la confiance des parties prenantes :** Une évaluation rigoureuse renforce la crédibilité auprès des bailleurs, des gouvernements et des communautés, facilitant les partenariats et le soutien à long terme.
- **Mise à l'échelle et réplication :** Les programmes bien évalués offrent un modèle reproductible, permettant à des approches efficaces de bénéficier à un plus grand nombre de communautés.
- **Prévention des préjudices :** L'évaluation peut détecter des conséquences négatives imprévues d'interventions bien intentionnées, avant qu'elles n'affectent des populations plus larges.

2. Le défi de la causalité

Au-delà du « avant et après »

Lorsqu'un programme est mis en œuvre et que l'on observe des améliorations, il semble naturel de supposer que l'intervention a fait la différence. Par exemple : une campagne de vaccination est lancée et les taux de maladie diminuent ; un programme d'éducation parentale commence et la fréquentation scolaire augmente. Ces liens paraissent évidents, mais ils peuvent être trompeurs.

Le défi fondamental de l'évaluation consiste à déterminer si l'intervention a réellement causé les changements observés, ou si ces changements se seraient produits pour d'autres raisons. Cette question est plus complexe qu'il n'y paraît à première vue.

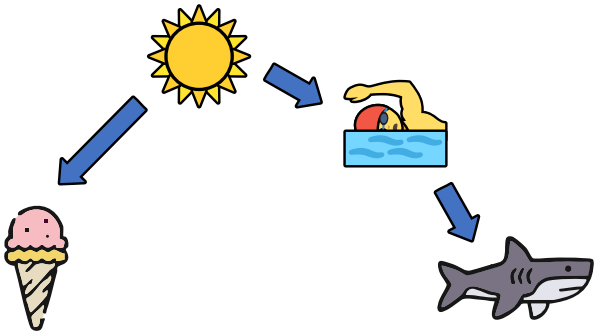
Corrélation et causalité

La corrélation signifie que deux événements se produisent en même temps, tandis que la causalité signifie que l'un provoque l'autre. Cette distinction est essentielle pour évaluer l'impact des programmes ou des interventions.

Prenons un exemple classique : les glaces et les attaques de requins. Les données montrent que lorsque les ventes de glaces augmentent, les attaques de requins augmentent également. Les ventes de glaces provoquent-elles les attaques de requins ? Bien sûr que non. Il s'agit de ce que les statisticiens appellent le « problème de la troisième variable » ou « confusion par cause commune », lorsqu'un facteur caché influence simultanément deux variables. Dans cet exemple, la température est la variable cachée qui influence indépendamment les deux phénomènes. Pendant les mois d'été :

- Des températures plus élevées augmentent la consommation de glaces
- Les mêmes températures plus élevées amènent davantage de personnes à se baigner dans l'océan
- Plus de nageurs dans l'eau accroît la probabilité de rencontres avec des requins

Ce phénomène peut être illustré par un diagramme causal simple ou graphe acyclique dirigé. Les flèches bleues représentent les influences causales. Il n'y a pas de flèche reliant les ventes de glaces aux attaques de requins, car il n'existe aucun lien causal direct entre elles. Elles sont corrélées (elles se produisent en même temps), mais non causales (l'une ne provoque pas l'autre).



Pourquoi cela est essentiel pour les programmes:
Sans compréhension claire de la différence entre la corrélation et la causalité, on risque de tirer de mauvaises conclusions, ainsi que de concevoir ou de mettre à l'échelle des interventions qui ne sont pas réellement responsables du changement observé.

Prenons un exemple concret de développement. Imaginons qu'un programme de nutrition de l'UNICEF soit mis en œuvre dans plusieurs communautés. Peu de temps après,

les indicateurs de croissance des enfants commencent à s'améliorer. Il peut sembler intuitif d'attribuer cette amélioration au programme, mais que pourrait-il se passer d'autre ?
Voici quelques explications possibles :

- Peut-être est-ce la saison des récoltes, ce qui augmente naturellement la disponibilité alimentaire
- Peut-être qu'une autre organisation a commencé à fournir de l'eau potable, réduisant les maladies diarrhéiques
- Ou peut-être que le gouvernement a mis en place une politique économique qui a augmenté les revenus des ménages au même moment

Chacun de ces facteurs pourrait expliquer ou contribuer à l'amélioration observée. En supposant que le programme soit la cause du changement — alors qu'il découle en réalité d'autres facteurs — on risque d'investir dans ou de mettre à l'échelle des interventions qui ne fonctionnent pas réellement. Pire encore, on peut passer à côté de ce qui a véritablement généré le changement, et perdre l'occasion de reproduire ou de renforcer les solutions les plus efficaces.

Comprendre la différence entre la corrélation et la causalité permet d'éviter ces écueils, en encourageant plutôt de meilleures questions, des évaluations plus intelligentes et des décisions mieux éclairées.

Facteurs de confusion : pourquoi les effets sont difficiles à isoler

Lorsqu'une intervention est mise en œuvre et que ses résultats sont mesurés, de nombreux facteurs autres que le programme peuvent influencer les changements observés. Ces facteurs indépendants sont appelés variables de confusion. Si l'on ne tient pas compte de ces variables, on risque d'attribuer à l'intervention des effets qui sont en réalité causés par autre chose.

Ce défi est appelé endogénéité : une situation où la relation entre une intervention et ses résultats est faussée parce que d'autres variables interviennent. Reconnaître ce phénomène permet de comprendre pourquoi de simples comparaisons « avant/après » peuvent être trompeuses.

Ci-dessous figurent quelques-uns des facteurs de confusion les plus courants, illustrés par des exemples issus de contextes typiques des programmes de l'UNICEF :

Facteurs de confusion liés au temps :

Changements qui se seraient produits indépendamment de notre intervention.

- **Variations saisonnières.** Les indicateurs nutritionnels s'améliorent après le lancement d'un programme d'alimentation, mais celui-ci a débuté juste avant la saison des récoltes, lorsque la nourriture est naturellement plus disponible.
- **Tendances préexistantes.** La scolarisation augmente après une campagne éducative, mais les données montrent que les taux étaient déjà en hausse grâce au développement économique à long terme.

Facteurs de confusion liés à la sélection :

Différences entre les personnes qui participent à un programme et celles qui n'y participent pas.

- **Biais d'auto-sélection.** Les familles qui rejoignent un programme parental peuvent être déjà plus investies dans l'éducation de leurs enfants, donnant l'impression que le programme est plus efficace qu'il ne l'est réellement.
- **Biais de ciblage.** Un programme WASH cible des communautés présentant des taux élevés de diarrhée. Même sans intervention, ces taux extrêmes pourraient diminuer avec le temps, simplement en raison d'une variation naturelle.

Facteurs de confusion liés à l'environnement :

Des événements externes survenant au même moment.

- **Programmes simultanés.** Une campagne de protection de l'enfance est lancée au moment où le gouvernement renforce l'application des lois sur le travail des enfants. Il devient alors difficile de déterminer quelle initiative est à l'origine des changements observés.
- **Changements de politique publique.** Un programme de nutrition préscolaire est déployé au moment où une subvention nationale sur les denrées alimentaires est introduite. Les deux peuvent contribuer à l'amélioration des indicateurs nutritionnels.

Facteurs de confusion liés à la mesure :

Changements dans la manière de suivre ou de détecter les résultats

- **Amélioration du suivi.** Après l'introduction d'un nouveau système de signalement, parallèlement à une initiative de lutte contre la traite, le nombre de cas augmente. Cette hausse ne traduit pas une augmentation de la traite, mais une amélioration de la détection.

Ces exemples montrent qu'il est difficile de déterminer si une intervention est réellement à l'origine des changements observés. Lorsque plusieurs facteurs influencent les résultats en même temps, comment isoler le véritable impact d'un programme ?

3. Le cadre contrefactuel : comprendre ce qui se serait produit autrement

Au cœur de l'inférence causale se trouve une question en apparence simple : Que se serait-il passé si l'intervention n'avait pas eu lieu ?

Ce scénario alternatif — où le programme n'existe pas — est appelé le contrefactuel. Il constitue le point de référence auquel on compare le résultat observé afin de déterminer si le programme a réellement fait une différence.

Prenons l'exemple d'un enfant qui reçoit un vaccin et ne contracte pas la maladie. Le vaccin a-t-il réellement empêché l'infection, ou l'enfant serait-il resté en bonne santé même sans vaccination ? Les deux situations — l'enfant vacciné et non vacciné — ne peuvent pas être observées simultanément. Ce dilemme est ce que les chercheurs appellent « le problème fondamental de l'inférence causale ». Il est tout simplement impossible d'observer à la fois le réel et le contrefactuel pour un même individu.

À défaut, le contrefactuel est approximé en identifiant ou en créant un groupe de comparaison valide. Ce groupe doit être aussi similaire que possible au groupe bénéficiant de l'intervention et soumis aux mêmes conditions externes — par exemple, les variations saisonnières, les changements économiques ou les réformes politiques — mais sans recevoir l'intervention.

Si les deux groupes font face au même contexte, tout écart significatif entre leurs résultats peut être attribué au programme lui-même. Il s'agit là de la base de toute évaluation crédible. Un contrefactuel construit avec rigueur permet de dépasser les suppositions et de répondre avec confiance à une question essentielle : L'intervention a-t-elle fait la différence, ou le changement se serait-il produit de toute façon ?

Résultats potentiels : une manière formelle de penser les contrefactuels

Pour raisonner sur l'impact causal, les statisticiens utilisent ce que l'on appelle le cadre des résultats potentiels (potential outcomes framework). Ce cadre offre une structure formelle permettant d'évaluer la différence qu'une intervention produit en imaginant plusieurs réalités possibles pour chaque unité d'observation. Une unité peut être une personne, un ménage, une école ou une communauté, et pour chacune, il existe deux résultats potentiels :

- $\mathbf{Y}_1$: le résultat si l'unité **reçoit** le traitement / l'intervention
- $\mathbf{Y}_0$: le résultat si l'unité **ne reçoit pas** le traitement / l'intervention

L'effet causal est la différence entre ces deux résultats potentiels : $\mathbf{Y}_1 - \mathbf{Y}_0$.

Voici le réel défi : il n'est possible d'observer qu'un seul de ces deux résultats pour un individu donné. Lorsqu'un enfant reçoit un vaccin, nous pouvons observer ce qui se passe avec vaccination (Y_1). En revanche, il est impossible d'observer ce qu'il serait advenu du même enfant sans vaccination (Y_0). Cette alternative non observée — le contrefactuel — reste à jamais inconnue.

C'est le problème fondamental de l'inférence causale : pour établir la causalité, il faudrait connaître à la fois ce qui s'est produit et ce qui se serait produit autrement, alors que seule une réalité peut être observée.

Enfant A (Groupe de traitement) → A reçu l'intervention → Résultat observé : Y_1

Enfant B (Groupe de comparaison) → N'a PAS reçu l'intervention → Résultat observé : Y_0

Effet causal = $Y_1 - Y_0$

La solution aux contrefactuels : passer d'un contrefactuel individuel à un contrefactuel de groupe

Puisqu'il est impossible d'observer simultanément les deux résultats pour une même personne (nous ne pouvons pas encore cloner les individus ni voyager entre des réalités parallèles !), il faut déplacer l'analyse du niveau individuel au niveau des groupes. En construisant avec soin des groupes de comparaison très similaires, il devient possible d'estimer ce qui se serait produit chez les personnes ayant reçu l'intervention, et chez celles qui ne l'ont pas reçue.

Au lieu de demander : « Cet enfant a-t-il pris du poids grâce au programme ? », la question devient plutôt : « En moyenne, combien de poids supplémentaire les enfants prennent-ils lorsqu'ils participent au programme de nutrition, par rapport à ceux qui n'y participent pas ? ».

Désormais, l'attention étant portée sur des groupes plutôt que sur des individus, il est nécessaire d'examiner comment créer des contrefactuels valides au niveau collectif. Autrement dit, comment constituer un groupe de comparaison qui soit aussi similaire que possible sur un grand nombre de facteurs/variables observables et non observables.

Le défi de la construction de contrefactuels valides

Maintenant que l'on comprend la nécessité de comparer des groupes plutôt que des individus, la question suivante est de savoir comment créer un groupe qui représente fidèlement ce qui se serait produit en l'absence de l'intervention.

La construction de ce groupe de comparaison (ou contrefactuel) est l'une des étapes les plus importantes — et les plus complexes — de l'évaluation causale. Comme mentionné précédemment, les facteurs de confusion peuvent facilement fausser les conclusions. La qualité de l'inférence causale dépend entièrement de la capacité du groupe de comparaison à refléter le groupe traité sur tous les aspects, à l'exception d'un seul : celui-ci n'a pas reçu le traitement ou l'intervention.

C'est ici que le choix du dispositif d'évaluation devient essentiel. Différentes approches permettent d'estimer ce qui se serait produit pour le groupe traité s'il n'avait pas reçu l'intervention (le contrefactuel). Chaque approche comporte des arbitrages entre rigueur, faisabilité et risque de biais. Certaines conceptions permettent de formuler des conclusions causales plus solides, mais nécessitent davantage de contrôle ou de ressources ; d'autres sont plus flexibles, mais introduisent davantage d'incertitude.

1. **Comparaison avant-après – Contrefactuel faible**

   Une approche courante mais problématique consiste à comparer les résultats avant et après une intervention. Cette méthode mesure les résultats juste avant le début d'un programme, puis après sa mise en œuvre, en attribuant tout changement observé à l'intervention. Bien que simple et intuitive, cette approche est fortement vulnérable aux facteurs de confusion qui influencent les résultats dans le temps, indépendamment du programme. Parmi ces facteurs :

   - **Facteurs de confusion liés au temps :** variations saisonnières ou tendances à long terme (par exemple, un programme de formation agricole montre une augmentation des rendements, mais la période d'évaluation coïncide avec la saison de croissance naturelle).
   - **Facteurs de confusion liés à l'environnement :** programmes ou changements politiques simultanés (par exemple, un programme de

nutrition semble efficace, mais le gouvernement a introduit gratuitement des repas scolaires dans la même zone au même moment).

- **Facteurs de confusion liés à la mesure :** le fait de mesurer modifie les comportements (par exemple, des enquêtes répétées sensibilisent les ménages à l'importance du lavage des mains, entraînant des changements indépendamment du programme).

Tous ces éléments peuvent créer une illusion d'impact, alors qu'en réalité, le changement aurait pu se produire sans intervention.

Par exemple, un programme de santé communautaire lancé en avril montre des résultats encourageants en août. Or, le mois d'août correspond également au début de la saison sèche, période durant laquelle les maladies hydriques diminuent naturellement. L'amélioration observée peut donc n'avoir aucun lien avec l'intervention. La logique de l'approche « avant–après » repose sur l'idée que "le même groupe, à un autre moment" peut servir de contrefactuel à lui-même. Dans des contextes dynamiques, cette hypothèse est généralement fausse, ce qui rend ce dispositif très faible pour tirer des conclusions causales fiables.

2. **Groupe de comparaison non équivalent - une approche meilleure, mais imparfaite**

Une alternative légèrement plus robuste consiste à comparer le groupe traité à un groupe de comparaison non équivalent, c'est-à-dire un groupe qui ne reçoit pas l'intervention, mais qui est observé sur la même période. Ce dispositif permet de limiter plusieurs facteurs de confusion liés au temps, puisque les deux groupes sont exposés aux mêmes conditions externes (saisons, politiques publiques, tendances économiques, etc.).

Cependant, ce type de comparaison demeure vulnérable aux facteurs de confusion liés à la sélection, c'est-à-dire aux différences entre les groupes qui peuvent influencer les résultats indépendamment du programme. Parmi ces biais :

- **L'auto-sélection :** les ménages qui choisissent de participer peuvent déjà être plus motivés, mieux informés ou plus soucieux de leur santé.
- **La sélection administrative :** les programmes sont parfois ciblés intentionnellement sur des zones à fort besoin ou à fort potentiel de réussite, biaisant les comparaisons.
- **Différences initiales :** avant même l'intervention, les communautés de comparaison peuvent différer sur des facteurs essentiels (infrastructures, niveau de revenu, démographie, etc.).

Les chercheurs tentent souvent d'apparier les groupes sur des caractéristiques observables, mais cette stratégie présente des limites : de nombreux facteurs déterminants (motivation, aspirations, résilience, génétique, etc.) ne peuvent pas être observés et peuvent néanmoins influencer les résultats.

Ainsi, bien qu'elle constitue une amélioration par rapport au design « avant–après », la comparaison non équivalente ne suffit généralement pas à établir une causalité robuste, sauf si elle est complétée par d'autres méthodes (ajustements statistiques, expériences naturelles, etc.).

3. Comment la randomisation crée la norme de référence (“gold standard”) des contrefactuels

La nécessité de meilleurs contrefactuels.

Comme indiqué précédemment, les comparaisons « avant–après » et les groupes de comparaison non équivalents présentent des limites. Elles peuvent permettre d’observer un changement, mais elles ont du mal à déterminer ce qui a réellement causé ce changement — en particulier lorsque des facteurs de confusion, connus ou inconnus, sont en jeu.

Tout comme dans l’exemple des glaces et des attaques de requins, de nombreuses relations observées dans le monde réel sont influencées par des variables cachées. Dans l’évaluation de programmes, ces variables cachées sont souvent nombreuses, complexes et impossibles à mesurer entièrement.

Il est donc nécessaire de disposer d’un moyen permettant de créer des groupes de comparaison équilibrés sur les caractéristiques observables et non observables.

C’est là que l’échantillonnage aléatoire intervient. En commençant par sélectionner de manière aléatoire différents groupes (individus, écoles ou communautés) au sein de la population cible, puis en attribuant aléatoirement lesquels recevront l’intervention, on crée des groupes statistiquement équivalents. En moyenne, ces groupes seront similaires sur l’ensemble des caractéristiques, qu’elles soient observables ou non, puisqu’ils proviennent de la même population de départ et ont été sélectionnés de manière comparable. Cela signifie que des facteurs tels que la motivation, l’état de santé initial, le revenu, les croyances non mesurées ou encore les normes communautaires seront répartis de façon aléatoire entre les groupes.

Lorsqu’elle est correctement mise en œuvre, la randomisation garantit que la seule différence entre les groupes est le fait d’avoir reçu ou non l’intervention. Cette approche rend beaucoup plus probable que les différences observées dans les résultats soient dues au programme lui-même, plutôt qu’à des facteurs externes ou à des différences préexistantes. C’est pour cette raison que la randomisation est qualifiée de « norme de référence » (gold standard) en inférence causale. Lorsqu’elle est bien menée et que la taille de l’échantillon est suffisante, cette méthode offre le niveau de confiance le plus élevé possible pour affirmer que l’intervention est la cause du changement observé.

La randomisation comme méthode de création de contrefactuels

En assignant aléatoirement les unités (c’est-à-dire les individus, ménages, écoles ou communautés) qui recevront une intervention, l’influence des facteurs de confusion est considérablement réduite. Tous les types de facteurs de confusion évoqués précédemment se retrouvent alors répartis de manière équilibrée entre les groupes, par construction. Concrètement, cela signifie que :

- Les variations saisonnières affectent les deux groupes de façon similaire
- Le biais d’auto-sélection est éliminé, car la participation est attribuée et non choisie
- Les tendances préexistantes se déploient de manière comparable dans les deux groupes
- Les programmes simultanés ou les changements de politique publique touchent les deux groupes au même moment
- Les effets liés à la mesure s’appliquent de la même manière dans les deux groupes

Ainsi, la seule différence systématique entre les groupes est le fait d’avoir reçu ou non l’intervention.

Cela permet d'attribuer directement à l'intervention toute différence observée dans les résultats, plutôt qu'à des facteurs externes ou cachés.

Pour illustrer, imaginons la mise en œuvre d'un programme destiné à encourager les parents à envoyer leurs enfants à l'école. Un grand nombre de parents sont éligibles, mais ils présentent une grande diversité de situations :

- Certains sont aisés, d'autres en difficulté financière
- Certains vivent près de l'école, d'autres très loin
- Certains ont fait des études supérieures, d'autres ont peu de scolarité formelle
- Certains ont des emplois flexibles, d'autres des horaires rigides
- Certains accordent une grande importance à l'éducation, d'autres sont plus sceptiques
- Certains ont un vécu scolaire positif, d'autres négatif
- Certains sont très motivés, d'autres moins

En assignant aléatoirement les parents soit au groupe bénéficiant de l'intervention, soit au groupe témoin (qui ne reçoit pas l'intervention), on s'assure que toutes ces caractéristiques sont réparties de façon similaire entre les groupes. Cela inclut les caractéristiques observables, comme le revenu ou la distance à l'école, ainsi que les caractéristiques non observables, comme les croyances ou la motivation.

Grâce à la randomisation, ces divers traits sont distribués de manière équilibrée entre les deux groupes. Dans cet exemple, toute différence constatée dans les taux de fréquentation scolaire après l'intervention peut donc être attribuée à l'intervention elle-même, et non à des différences préexistantes entre les parents qui l'ont reçue ou non.

Le chemin qui mène des bonnes intentions à un impact réel

Une évaluation rigoureuse n'est pas seulement une question de crédibilité scientifique : elle permet de s'assurer que les programmes améliorent réellement la vie des populations. Comme mentionné précédemment, des interventions bien intentionnées peuvent ne pas produire d'impact, gaspiller des ressources précieuses, voire causer des effets négatifs inattendus lorsqu'elles reposent sur des suppositions plutôt que sur des données probantes. La distinction entre la corrélation et la causalité est donc essentielle : elle détermine si l'on met à l'échelle des solutions qui fonctionnent réellement, ou si l'on investit dans des programmes qui ne font que coïncider avec des changements positifs. En construisant des contrefactuels valides — idéalement grâce à la randomisation — l'analyse passe de ce qui semble fonctionner à ce qui fonctionne réellement, pour qui, et pour quelles raisons.

Cette connaissance transforme la manière de concevoir les programmes, d'allouer les ressources et, au final, de servir les communautés. Si l'évaluation rigoureuse peut sembler complexe, poursuivre des interventions sans connaître leur véritable impact est bien plus risqué.

Préoccupations fréquentes concernant la randomisation et pistes de réponse

Bien que la randomisation soit l'une des approches les plus rigoureuses pour répondre aux défis liés à la causalité et aux contrefactuels, elle suscite souvent des préoccupations. Celles-ci peuvent être regroupées en six grandes catégories :

1. Coût et ressources

PRÉOCCUPATION : les évaluations randomisées sont coûteuses et nécessitent beaucoup de ressources.

RÉPONSE :

- Bien que les évaluations rigoureuses demandent un investissement, leur coût doit être comparé à la valeur d'obtenir des données fiables.
- Tous les essais randomisés ne doivent pas être coûteux ni de grande ampleur : des études ciblées peuvent être rentables.
- Le coût de mise en œuvre à grande échelle d'un programme inefficace est largement supérieur aux coûts d'évaluation.
- L'utilisation de sources de données existantes et de conceptions méthodologiques ingénieuses peut parfois réduire considérablement les coûts.

2. Contraintes de temps

PRÉOCCUPATION : Les évaluations randomisées sont trop longues, alors que l'UNICEF doit agir rapidement.

RÉPONSE :

- Les tests rapides itératifs permettent de tester des interventions à petite échelle, de mesurer les résultats rapidement et d'ajuster les solutions.
- Les mises en œuvre par phases permettent une action immédiate tout en garantissant une évaluation rigoureuse.
- Du temps consacré à l'évaluation permet d'éviter des années de déploiement de programmes inefficaces.
- Certains résultats peuvent être mesurés à court terme (ex : observance des ARV), tandis que d'autres nécessitent un suivi plus long (ex : charge virale supprimée).

3. Préoccupations éthiques

PRÉOCCUPATION : Il est contraire à l'éthique de priver des groupes de programmes potentiellement bénéfiques.

RÉPONSE :

- Lorsque les ressources sont limitées, la randomisation peut représenter le mode d'allocation le plus équitable.
- Il est impossible de savoir si un programme est réellement bénéfique sans l'évaluer ; certains peuvent être inutiles voire nocifs.
- Les mises en œuvre par phases garantissent que tous les groupes bénéficient du programme si celui-ci s'avère efficace.
- Il existe une obligation éthique de s'assurer que les programmes améliorent réellement la vie des enfants.

4. Défis politiques et défis liés aux parties prenantes

PRÉOCCUPATION : Les partenaires gouvernementaux ou les communautés peuvent refuser la randomisation.

RÉPONSE :

- L'importance du cadrage : mettre en avant les bénéfices, y compris la possibilité de réaliser des évaluations à faible coût ou limitées dans le temps.
- Impliquer les parties prenantes dès la phase de conception permet de répondre aux préoccupations et de renforcer l'adhésion.
- Définir des procédures claires pour interrompre l'essai si des preuves solides d'efficacité émergent.
- Expliquer que l'évaluation renforce le plaidoyer en faveur des programmes efficaces

5. Pertinence contextuelle et généralisabilité

PRÉOCCUPATION : Les résultats obtenus dans un contexte ne sont pas applicables aux autres contextes où travaille l'UNICEF.

RÉPONSE:

- Une sélection stratégique des sites peut améliorer la généralisabilité.
- Mesurer les facteurs d'implémentation permet de comprendre quels aspects du contexte sont déterminants.
- Des preuves localisées valent mieux que l'absence de preuves.
- Des évaluations menées dans plusieurs contextes contribuent à construire un socle de connaissances plus large.

6. Capacité technique

PRÉOCCUPATION: Le personnel de l'UNICEF pourrait ne pas disposer des compétences techniques nécessaires pour concevoir et analyser des évaluations randomisées.

RÉPONSE:

- Des partenariats avec des institutions académiques peuvent compléter les compétences internes.
- Investir dans la formation du personnel permet de renforcer des capacités durables au sein de l'organisation.
- Des conceptions expérimentales simples peuvent être plus accessibles que certaines méthodes quasi-expérimentales complexes.
- Les spécialistes de l'évaluation au sein de l'UNICEF peuvent fournir un appui technique transversal aux programmes.

Considérations relatives à la conception d'une évaluation d'impact

Après avoir établi pourquoi l'évaluation est essentielle et comment la causalité peut être identifiée de manière crédible, l'attention se porte désormais sur la conception d'évaluations qui formulent les bonnes questions, mesurent les bons résultats et produisent des données probantes permettant d'orienter directement les décisions. La section suivante présente les principaux éléments de conception à prendre en compte pour mener des évaluations d'impact rigoureuses.

1. Définir les questions de recherche

Toute évaluation commence par un défi fondamental : clarifier les questions clés auxquelles il faut répondre. Cela implique de réunir les principales parties prenantes et de se demander :

- Quelles décisions dépendent de cette évaluation ?
- Que ferait-on différemment si l'on savait X plutôt que Y ?

Il convient de se concentrer sur 3 à 5 questions centrales, directement liées à des actions concrètes. Ces questions se répartissent généralement en trois catégories :

- **Efficacité**: L'intervention fonctionne-t-elle ? Au bout de combien de temps ? À quel moment les effets apparaissent-ils ?
- **Mécanisme**: Comment fonctionne l'intervention ?
- **Ciblage**: pour qui fonctionne-t-elle le mieux ?

Pour chaque question, il est essentiel de préciser comment la réponse orientera les décisions. Par exemple, si un programme de formation professionnelle augmente l'emploi de 10 %, faut-il l'étendre ? Et si l'effet n'est que de 5 % ? Et s'il fonctionne pour les hommes mais pas pour les femmes ? Définir ces seuils décisionnels à l'avance permet d'éviter toute rationalisation a posteriori et garantit que l'évaluation produira des éléments concrets d'aide à la décision. Il est également crucial d'anticiper les résultats positifs et les résultats nuls – savoir qu'une intervention ne fonctionne pas est tout aussi précieux pour l'allocation des ressources.

Une fois les questions de recherche clarifiées, l'étape suivante consiste à déterminer comment y répondre. Il s'agit de traduire des objectifs généraux en résultats mesurables et de sélectionner les bons indicateurs pour capter le changement réel.

2. Sélection des indicateurs principaux et de la stratégie de mesure

La réussite de l'évaluation dépend du fait de mesurer les bons éléments, de la bonne manière. Cela commence par la transformation de finalités générales en questions évaluatives formulées avec précision. Une question vague telle que « Ce programme de nutrition fonctionne-t-il ? » conduit à des réponses vagues et à des décisions incertaines. À l'inverse, une question précise telle que : « Le fait de fournir un conseil nutritionnel mensuel augmente-t-il les scores z taille-pour-âge d'au moins 0,2 écart-type chez les enfants de 6 à 24 mois dans les zones rurales, au cours d'une période de 12 mois ? »

En d'autres termes : « L'intervention [X] entraîne-t-elle [un changement mesurable précis, avec seuil si pertinent] chez [une population définie] dans [un délai et un contexte spécifiés] ? » Ce niveau de précision oriente toutes les décisions de mesure ultérieures et garantit que les résultats seront interprétables et exploitables.

Une fois les questions définies, il faut identifier les mesures de résultats qui permettront d'y répondre.

3. Choisir ce qu'il faut mesurer

Choisissez des résultats qui reflètent clairement ce que le programme cherche à accomplir. Il faut privilégier des indicateurs qui montrent directement si l'intervention produit l'effet attendu. Ceux-ci doivent être suffisamment spécifiques pour être mesurés avec rigueur et suffisamment importants pour guider les décisions relatives au programme.

Tenez compte de la proximité des résultats par rapport à l'intervention. Les résultats proximaux, observables peu de temps après l'intervention, sont plus faciles à

modifier et à mesurer, mais ils ne reflètent pas toujours l'objectif final. Les résultats distaux, situés plus loin dans la chaîne des effets, reflètent l'impact réel, mais nécessitent généralement des échantillons plus importants et davantage de temps pour être détectés.

Par exemple, un programme d'alimentation scolaire peut entraîner rapidement une hausse de l'assiduité (résultat proximal), mais il faudra davantage de données et de temps pour mettre en évidence une amélioration des apprentissages (résultat distal). Les deux types de résultats sont utiles : les effets à court terme permettent de vérifier si le programme progresse comme prévu, tandis que les effets à long terme indiquent s'il produit réellement un changement significatif.

Une fois les résultats clairement définis, il devient essentiel de déterminer comment ils seront mesurés. Les différentes sources de données varient en termes de précision, de coût et de faisabilité. Reconnaître ces arbitrages permet de prendre des décisions de mesure qui renforcent — plutôt que compromettent — la crédibilité de l'évaluation.

Méthodes de collecte de données et compromis

Chaque source de données présente des avantages et des limites spécifiques :

Les enquêtes offrent une grande souplesse pour mesurer exactement ce dont on a besoin, mais elles comportent plusieurs défis. Les comportements autodéclarés sont souvent affectés par le biais de désirabilité sociale. Par exemple, les parents tendent à surestimer la vaccination ou à sous-déclarer les pratiques de discipline sévères. Les périodes de rappel jouent également un rôle majeur : interroger au sujet d'événements survenus la semaine précédente produit des résultats plus fiables qu'en demandant de se souvenir de faits remontant à une année. La fatigue liée aux enquêtes peut également compromettre la qualité des données lorsque les questionnaires sont trop longs. Par ailleurs, les groupes marginalisés peuvent être difficiles à atteindre via des enquêtes téléphoniques, et les personnes issues de communautés à faible niveau d'alphabétisation peuvent avoir du mal à comprendre certaines questions. Il est donc essentiel de réaliser des entretiens cognitifs lors de la phase pilote, afin de s'assurer que les questions sont culturellement appropriées et comprises comme prévu.

Les données administratives, telles que les registres scolaires, les dossiers cliniques ou les bases de données de programmes, constituent une source de mesure objective, souvent moins coûteuse, mais présentent certaines limites. Elles se restreignent aux informations déjà collectées, qui ne correspondent pas toujours aux résultats recherchés. Les données concernant certains groupes ethniques ou populations marginalisées peuvent être incomplètes ou absentes. La qualité de ces données peut également varier considérablement : certains centres de santé tiennent des registres méticuleux, tandis que d'autres ne fonctionnent presque pas. L'utilisation de données administratives impose souvent d'aligner l'unité de randomisation sur les niveaux administratifs (écoles, cliniques), plutôt que sur les individus.

L'observation directe et les mesures comportementales fournissent des évaluations objectives, mais leur mise en œuvre doit être rigoureuse. Les observateurs doivent recevoir une formation approfondie pour garantir la cohérence des mesures. Il est essentiel d'aborder les communautés avec sensibilité, de demander le consentement de manière appropriée, et d'éviter toute intrusion dans des espaces privés ou culturellement sensibles. Les technologies offrent désormais des possibilités de mesures discrètes (géolocalisation par GPS, capteurs), mais ces approches peuvent s'avérer impossibles dans certains contextes en raison de coûts élevés ou de contraintes techniques.

Les biomarqueurs et les mesures anthropométriques fournissent des données objectives dans le cadre d'interventions sanitaires, mais nécessitent des compétences spécialisées et du matériel adéquat. Peut-on garantir la chaîne du froid pour les prélèvements sanguins ? Les participants accepteront-ils des procédures invasives ? Comment les erreurs de mesure provenant de différents enquêteurs seront-elles gérées, tout en veillant à ce que la collecte soit respectueuse et aussi peu contraignante que possible pour les participants ?

L'identification de sources de données appropriées est indispensable, mais insuffisante à elle seule ; le moment et la fréquence de la collecte sont tout aussi déterminants. Une collecte mal synchronisée peut masquer des effets réels ou déformer l'évaluation des performances d'un programme.

4. Calendrier et fréquence des mesures

Les résultats apparaissent selon des échelles de temps différentes. Les connaissances peuvent évoluer en quelques semaines, les comportements sur plusieurs mois, et les impacts sur la santé sur plusieurs années. Mesurer trop tôt expose au risque de conclure à l'absence d'effet pour des interventions qui nécessitent du temps pour produire un changement. Mesurer trop tard peut conduire à ignorer des effets qui s'atténuent ou sont modifiés par d'autres facteurs.

Il est souvent pertinent de prévoir plusieurs points de mesure afin de comprendre la dynamique des effets. L'impact augmente-t-il, se stabilise-t-il ou diminue-t-il avec le temps ? Un effet initial qui s'estompe peut indiquer la nécessité de mesures de renforcement — les phénomènes de « décroissance d'effet » sont fréquents dans les interventions de changement de comportement. Un changement progressif peut signaler que les effets se construisent au fil du temps et éventuellement par différents mécanismes. Idéalement, il convient de prévoir au moins un suivi après la période immédiate post-intervention (par exemple, six mois plus tard) pour évaluer la persistance des effets.

Les variations saisonnières peuvent fausser les résultats si elles ne sont pas prises en compte. Par exemple, les résultats agricoles varient selon les cycles de récolte, les maladies suivent des tendances saisonnières, et les indicateurs scolaires fluctuent selon le calendrier académique. Il faut donc planifier la collecte de données de manière à éviter d'attribuer à l'intervention des changements dus à la saison, ou bien s'assurer que les groupes de traitement et de contrôle sont mesurés au même moment. Dans l'idéal, il faut également intégrer ces variations saisonnières dans l'analyse pour comprendre comment les effets interagissent avec celles-ci.

Le moment de la mesure indique quand le changement se produit, tandis que les mécanismes expliquent pourquoi. En mesurant tout au long de la chaîne causale, il devient possible de comprendre comment l'impact se construit, d'identifier les points de défaillance et d'améliorer la conception des programmes.

5. Mesurer les mécanismes tout au long de la chaîne causale

Comprendre pourquoi les programmes fonctionnent (ou ne fonctionnent pas) est tout aussi important que de savoir s'ils fonctionnent. La mesure des mécanismes remplit plusieurs fonctions :

Validation de la théorie : Les voies d'action supposées opèrent-elles réellement ? Par exemple, un programme de lavage des mains repose sur l'enchaînement suivant : information → connaissances → changement d'attitudes → changement de comportement → amélioration de la santé. Mesurer chacune de ces étapes permet de confirmer ou de remettre en question ces hypothèses.

Diagnostic des échecs : Lorsque les résultats n'évoluent pas, les mécanismes révèlent où la chaîne s'est rompue. Les enseignants ont-ils reçu la formation ? L'ont-ils comprise ? L'ont-ils mise en œuvre ? Les élèves ont-ils été réceptifs ? Chaque point de rupture appelle des solutions différentes.

Amélioration du programme : Plutôt que d'abandonner globalement un programme « inefficace », les données sur les mécanismes permettent d'y apporter des ajustements ciblés. Par exemple, si les parents ont reçu des informations nutritionnelles mais n'ont pas les moyens d'acheter des aliments diversifiés, l'ajout de coupons ou de transferts pourrait déclencher l'impact attendu.

Généralisation : Les programmes qui reposent sur des mécanismes universels, tels que les rappels, sont plus susceptibles d'être transférables dans différents contextes que ceux dépendant de caractéristiques institutionnelles spécifiques.

Ne vous limitez pas à mesurer les résultats finaux : suivez également les étapes intermédiaires. Dans le cadre d'un programme de nutrition visant à réduire la malnutrition infantile, cela pourrait se traduire par : les connaissances

des aidants (immédiat), les pratiques d'alimentation (court terme), la diversité alimentaire de l'enfant (moyen terme) et son statut nutritionnel (long terme). Chacun de ces indicateurs apporte une information précieuse sur la manière dont le programme fonctionne.

Comprendre les mécanismes permet de révéler comment le changement se produit. Toutefois, pour interpréter ces tendances avec exactitude, il est essentiel de veiller à ce que les données reflètent véritablement les populations que l'on souhaite servir. La représentativité et l'inclusion dans la mesure sont indispensables pour produire des preuves qui capturent la diversité des réalités, et non seulement celle des groupes les plus faciles à atteindre.

6. Assurer une mesure représentative et inclusive

Les personnes que l'on mesure comptent autant que ce que l'on mesure. Les enquêtes réalisées en milieu scolaire excluent systématiquement les enfants non scolarisés, souvent parmi les plus vulnérables. Les enquêtes téléphoniques excluent les personnes sans téléphone, et les données cliniques ne tiennent pas compte de celles qui ne sollicitent pas les services de santé ou qui en sont empêchées.

Concevez dès le départ des stratégies de mesure inclusives, en veillant à ce que les communautés participent à l'élaboration des indicateurs et des sources de données. Utilisez plusieurs sources de données pour couvrir différentes populations. Suréchantillonnez les groupes marginalisés afin de garantir leur représentation. Traduisez les outils dans les langues locales et testez-les auprès de répondants diversifiés. Formez des enquêteurs issus des communautés enquêtées afin de renforcer la relation de confiance et la compréhension mutuelle.

Tenez compte du point de vue qui est recueilli. Les enfants, les parents, les enseignants et les agents de santé peuvent faire des déclarations différentes au sujet d'un même phénomène. Même au sein d'un groupe donné (par exemple les mères), les expériences peuvent varier selon l'identité sociale, le parcours ou la position au sein de la communauté. La triangulation entre différents répondants peut révéler des dynamiques importantes, mais elle nécessite des protocoles clairs pour traiter les divergences.

Même les instruments les mieux conçus peuvent s'avérer insuffisants s'ils ne peuvent pas être mis en œuvre efficacement dans des conditions réelles et quotidiennes. Il est donc essentiel de garantir que les approches de mesure soient réalisables, fiables et adaptées aux contraintes du terrain.

7. Considérations pratiques relatives à la mesure

Au cours du processus de mesure, il convient de garder à l'esprit les éléments suivants :

Réaliser des tests pilotes approfondis. Ne jamais présumer que des outils efficaces ailleurs fonctionneront dans un autre contexte. Il faut tester les instruments auprès d'un petit groupe de répondants dans des conditions proches du terrain réel. Vérifiez les logiques de saut, le temps de passage, les traductions et la compréhension. Organisez des débriefings approfondis avec les enquêteurs, car ils repèrent souvent des problèmes que les répondants ne signalent pas.

Trouvez le juste équilibre entre exhaustivité et faisabilité. Des instruments longs produisent des données riches, mais entraînent de la fatigue chez les répondants, des coûts plus élevés et une baisse de la qualité. La plupart des effets peuvent être détectés à l'aide d'outils ciblés. Il est préférable de réserver les mesures longues aux études de mécanismes à petite échelle, plutôt qu'aux évaluations d'impact de grande envergure.

Anticiper l'erreur de mesure. Toute mesure comporte une marge d'erreur : l'évaluation anthropométrique varie d'un enquêteur à l'autre, et les scores de tests dépendent des conditions d'examen. Il faut prévoir des dispositifs de contrôle de qualité, tels que des exercices de standardisation pour les enquêteurs, la répétition de mesures sur des sous-

échantillons et la validation des données à partir de sources externes lorsque cela est possible.

Tout documenter. Élaborer des protocoles détaillant précisément la manière dont chaque résultat est mesuré, codé et construit. Les futurs utilisateurs doivent comprendre et, si nécessaire, reproduire les mesures choisies. Les annexes devraient inclure les questionnaires, les supports de formation, ainsi que le code de construction des variables.

Des pratiques de mesure rigoureuses garantissent la qualité des données ; toutefois, leur valeur dépend également de la capacité de l'étude à détecter de véritables effets. Une puissance statistique adéquate permet d'éviter des conclusions erronées, en s'assurant que les données ne se contentent pas de décrire ce qui a été observé, mais permettent de révéler ce qui a réellement fonctionné.

8. Déterminer la taille de l'échantillon et la puissance statistique

La puissance statistique correspond à la capacité d'une évaluation à détecter un effet réel lorsqu'il existe. On peut la comparer à un radar suffisamment sensible pour repérer un avion qui s'approche. À l'inverse, une étude dont la puissance est insuffisante revient à chercher quelque chose avec une lampe torche trop faible : l'effet peut exister sans jamais être détecté. Cette notion devient particulièrement importante lorsqu'il s'agit de déterminer le nombre de participant·e·s nécessaire. Si l'échantillon est trop petit, on risque de conclure que le programme n'a eu aucun effet alors qu'en réalité il en a eu (ce que l'on appelle un « faux négatif »). À l'inverse, enquêter un nombre excessif de participant·e·s conduit à un gaspillage de ressources. Plusieurs facteurs influencent la puissance, notamment :

- La taille attendue de l'effet du programme : les effets importants sont plus faciles à détecter.
- Le degré de variation naturelle dans la mesure des résultats : plus cette variation est élevée, plus l'échantillon requis sera grand.
- L'unité de randomisation choisie : la randomisation au niveau des communautés exige des échantillons plus importants que la randomisation au niveau individuel.

Il est essentiel d'intégrer des hypothèses réalistes concernant l'attrition (10 à 20 % est courant), la non-observance (le groupe traité ne reçoit pas totalement l'intervention) et la contamination (le groupe témoin accède à l'intervention). Ces éléments réduisent la taille effective de l'échantillon. Il vaut mieux recruter 20 % de participant·e·s supplémentaires que de découvrir, une fois les données collectées, que l'étude est sous-dimensionnée.

9. Architecture de randomisation

Lors de l'évaluation d'une intervention, le choix du niveau de randomisation est essentiel. Il s'agit de déterminer si l'intervention sera attribuée aux individus, aux groupes, aux écoles, aux communautés, ou à toute autre unité pertinente pour le projet. Le niveau choisi doit correspondre à la manière dont l'intervention est réellement mise en œuvre. Par exemple, si un nouveau programme scolaire ne peut être appliqué qu'à une classe entière et non à des élèves individuellement, alors la classe devient l'unité la plus appropriée pour la randomisation.

Le niveau de randomisation dépend également du type d'information pouvant être collectée. Si les données ne peuvent être mesurées qu'à un niveau agrégé — comme les dépenses des ménages ou l'assiduité scolaire — il est alors logique de randomiser au même niveau.

TABLEAU 8. DIFFÉRENTS TYPES D'UNITÉS DE RANDOMISATION

UNITÉ DE RANDOMISATION	AVANTAGES	CONSIDÉRATIONS	QUAND L'UTILISER	EXEMPLE
Randomisation au niveau individuel	Nécessite une taille d'échantillon plus faible que les autres niveaux ; forte efficacité statistique.	Risque d'effets de débordement (« spillovers ») lorsque les individus interagissent ; difficultés logistiques pour fournir des interventions différentes au sein d'un même environnement (ex. communauté, salle de classe).	Adaptée lorsque l'interaction entre individus est limitée et que l'intervention peut être ciblée facilement sur des personnes spécifiques.	Rappels SMS de vaccination envoyés à des aidants choisis aléatoirement dans un grand district urbain.
Randomisation au niveau des ménages	Permet de saisir les décisions prises au niveau du foyer ; correspond à la manière dont de nombreux comportements et résultats sont déterminés.	Les débordements restent possibles lorsque les voisins interagissent ; nécessité de définir clairement ce qu'est un « ménage » ; l'analyse peut nécessiter de tenir compte de la taille du foyer.	Appropriée lorsque les interventions concernent ou impliquent l'ensemble des membres d'un ménage (ex. visites à domicile, transferts monétaires conditionnels).	Transferts monétaires conditionnels accordés à des ménages tirés au sort avec enfants de moins de cinq ans.
Randomisation au niveau de la communauté ou du village (randomisation par grappes)	Réduit le risque de contamination ou de débordement ; souvent plus facile à gérer sur les plans politique et opérationnel.	Nécessite davantage de grappes (communautés) pour obtenir une puissance statistique suffisante ; grande variation entre communautés → augmente la variance ; logistique d'implémentation plus complexe à grande échelle.	Utile lorsque les individus d'une même communauté sont susceptibles de s'influencer mutuellement ou lorsque l'intervention est délivrée de manière publique (ex. mobilisation communautaire).	Campagnes de vaccination menées par des agents de santé communautaires, testées dans des villages sélectionnés aléatoirement.

UNITÉ DE RANDOMISATION	AVANTAGES	CONSIDÉRATIONS	QUAND L'UTILISER	EXEMPLE
Randomisation au niveau des établissements (par exemple, écoles, cliniques)	Pratique dans des dispositifs de prestation institutionnelle ; alignée avec les structures organisationnelles existantes ; adaptée aux interventions ciblant le personnel ou l'ensemble d'une structure.	Les structures peuvent varier en taille, qualité ou taux de rotation du personnel ; chevauchement des zones de recrutement → risques de débordement ; la puissance statistique est limitée par le nombre de structures disponibles.	Appropriée pour évaluer des interventions délivrées via des institutions, en particulier lorsque le ciblage au niveau individuel n'est pas envisageable.	Formation en communication interpersonnelle dispensée au personnel de centres de santé sélectionnés aléatoirement.

10. Lorsque la randomisation n'est pas possible

Il arrive que la randomisation ne soit pas réalisable pour des raisons politiques, éthiques ou pratiques. Les parties prenantes peuvent considérer l'attribution aléatoire comme injuste, les programmes peuvent déjà être déployés, ou les tailles d'échantillon peuvent être trop faibles pour permettre une randomisation pertinente. Dans ces cas, les méthodes quasi-expérimentales visent à créer des comparaisons valides en utilisant des techniques statistiques pour approximer le contrefactuel qu'aurait fourni la randomisation.

Ces approches consistent à identifier et contrôler les facteurs qui influencent à la fois la participation au programme et les résultats, ce que l'on appelle parfois « fermer les chemins rétrospectifs ». Bien que ces méthodes puissent fournir des preuves utiles, elles reposent sur des hypothèses plus fortes concernant les données et le contexte. Elles exigent généralement des échantillons plus importants, une collecte de données plus complète et des analyses plus complexes que les essais randomisés contrôlés (ERC). Surtout, elles demeurent vulnérables aux biais provenant de facteurs non observés — biais que la randomisation aurait éliminés.

Pour davantage de conseils sur les approches quasi-expérimentales, voir l'annexe 1. Il est recommandé de n'utiliser ces méthodes qu'avec l'appui d'experts, car leur validité repose sur des hypothèses spécifiques au contexte, souvent impossibles à vérifier.

Liste de vérification pratique pour la mise en œuvre d'une évaluation d'impact

Protocole d'évaluation

Avant de développer un protocole d'évaluation, il est utile de définir les objectifs d'apprentissage à l'aide d'un outil tel que l'Agenda d'apprentissage. Cet outil permet de formuler et d'organiser les principales questions auxquelles l'évaluation d'impact devra répondre. Son utilisation est illustrée à l'annexe 2 à travers **l'étude de cas sur l'augmentation de la couverture vaccinale des enfants au Liban**.

Une fois l'agenda d'apprentissage établi, il est alors possible d'élaborer le protocole d'évaluation. Le protocole fournit un cadre structuré et un plan détaillé expliquant comment l'intervention sera évaluée. Élaborer un protocole transforme les décisions de conception en un document technique complet qui guide la mise en œuvre et l'analyse. On peut l'envisager comme un contrat avec son "soi futur", empêchant les analyses sélectives et garantissant l'intégrité scientifique. Un protocole solide pré-spécifie toutes les décisions analytiques avant d'observer les résultats, protégeant ainsi contre les biais conscients et inconscients visant à obtenir des effets positifs.

Éléments essentiels d'un protocole:

- Description détaillée de l'intervention : ce qui sera délivré, par qui, et à quelle fréquence
- Théorie du changement avec des chaînes causales explicites
- Questions d'évaluation associées à des hypothèses précises
- Calculs de puissance statistique avec toutes les hypothèses rendues explicites
- Définitions précises des indicateurs, incluant les questions exactes des enquêtes (si les indicateurs proviennent d'enquêtes)
- Procédures de randomisation, y compris les variables de stratification
- Modèles analytiques avec les équations de régression exactes
- Liste des covariables déterminée par la théorie, et non par les données
- Analyses de sous-groupes assorties d'une justification claire
- Procédures pour gérer l'attrition et la non-observance
- Tests de robustesse pour vérifier la sensibilité des résultats

La pré-spécification du plan d'analyse est particulièrement cruciale pour :

- les résultats primaires vs secondaires, afin d'éviter le changement opportuniste de critères d'évaluation
- les analyses de sous-groupes, afin d'éviter la recherche d'effets significatifs a posteriori
- les critères d'inclusion/exclusion de l'échantillon d'analyse
- le traitement des valeurs aberrantes et des données manquantes
- Ajustements pour tests multiples

Toute déviation par rapport au protocole doit être clairement signalée comme exploratoire dans les rapports.

À titre de référence, voir l'étude de cas au Liban, où le protocole d'évaluation complet est disponible.

Envisagez d'enregistrer le protocole dans des registres publics (AEA Social Science Registry, RIDIE, ClinicalTrials.gov) avant le début de la collecte de données. L'enregistrement fournit un horodatage attestant la pré-spécification des analyses et permet à la communauté scientifique de suivre l'ensemble des études, et non uniquement celles ayant abouti à des résultats positifs. Il convient d'inclure suffisamment de détails pour qu'un autre chercheur puisse répliquer l'évaluation, tout en préservant une flexibilité opérationnelle permettant d'apporter les adaptations de terrain nécessaires sans compromettre la conception fondamentale de l'étude.

Plan de mise en œuvre

Le plan de mise en œuvre est la feuille de route opérationnelle détaillée qui traduit la conception de l'évaluation en actions concrètes sur le terrain. Il transforme le protocole technique en instructions quotidiennes indiquant qui fait quoi, quand, où, et avec quelles ressources, sur l'ensemble du cycle de l'évaluation. Ce plan comprend notamment :

- des calendriers détaillés avec dates précises et jalons
- \des attributions de rôles claires, avec une personne responsable identifiée pour chaque tâche
- les besoins en ressources (temps du personnel, matériel, transport, technologie)
- des indicateurs de suivi de l'exécution
- des plans de contingence pour les problèmes fréquents
- des protocoles de communication entre membres de l'équipe et partenaires

L'outil « Plan de mise en œuvre » peut servir de modèle simple et utile pour développer un guide étape par étape. Un exemple d'utilisation de cet outil est présenté à l'annexe 2.

Pourquoi c'est important ?

La différence entre une excellente conception d'évaluation et un échec sur le terrain provient généralement d'une planification opérationnelle insuffisante. Même les conceptions les plus rigoureuses échouent lorsque les tablettes ne sont pas chargées, que les questionnaires ne sont pas correctement traduits ou que les équipes ignorent qui est responsable du recrutement des participants. Le plan de mise en œuvre évite une multitude de petits dysfonctionnements susceptibles d'invalider une évaluation bien conçue. Il garantit la coordination entre de multiples acteurs (équipe de recherche, partenaires de mise en œuvre, autorités publiques, leaders communautaires) aux priorités et modes de travail parfois différents.

Sans planification opérationnelle claire, on risque de découvrir trop tard que la campagne de vaccination coïncide avec la collecte de données, que des membres clés de l'équipe sont indisponibles à un moment crucial, ou que les documents n'ont pas été imprimés à temps. Le plan constitue également un outil de gestion permettant de suivre l'avancement d'opérations complexes et d'identifier rapidement les problèmes avant qu'ils ne compromettent l'évaluation.

Gardez à l'esprit ces considérations clés :

- **Planifiez en tenant compte des contraintes fixes.** Partez des dates non flexibles (saisons agricoles, calendrier scolaire, fêtes religieuses, cycles budgétaires) pour construire un calendrier réaliste, puis ajouter 20–30 % de marge pour les retards inévitables. Si la collecte de données est estimée à trois semaines, en prévoir quatre.
- **Attribuez une responsabilité claire et unique pour chaque activité.** Évitez les responsabilités partagées, qui se traduisent souvent par une absence réelle de responsabilité. Par exemple, remplacer « l'équipe fera la formation » par : qui dirige, qui assiste, et qui est responsable en cas d'échec.
- **Suivez la mise en œuvre, pas seulement les résultats.** Incluez des indicateurs simples, suivis chaque semaine : nombre de participants recrutés vs objectif, enquêtes réalisées par jour, sessions réalisées comme prévu. Ces indicateurs portent uniquement sur l'exécution, pas sur les effets.

- **Prévoyez un budget complet.** Inclure les coûts souvent négligés : services de traduction, transport pour les superviseurs, crédits téléphoniques pour les suivis, rafraîchissements lors de réunions communautaires, remplacement du matériel abîmé, etc.
- **Précisez la chaîne de gestion des données.** Indiquez comment les données seront transférées des formulaires papier à la base numérique, qui les vérifiera, qui y aura accès, à quelle fréquence, et où elles seront stockées.
- **Planifiez des solutions face aux problèmes fréquents rencontrés sur le terrain, à l'aide de mesures de contingence spécifiques.** Que se passera-t-il si de fortes pluies empêchent les déplacements pendant la période d'enquête ? Si des membres clés de l'équipe tombent malades ou démissionnent ? Si les priorités gouvernementales changent soudainement et que l'agence partenaire est réaffectée ? Si les participants sont occupés par les récoltes au moment prévu pour le suivi ? Utilisez l'outil « Risques de mise en œuvre et stratégies d'atténuation » pour documenter de façon systématique les risques anticipés ainsi que les stratégies prévues pour y répondre. Pour un exemple d'utilisation concrète de cet outil, consultez l'annexe 2 pour l'étude de cas au Liban, qui illustre son application pratique.
- **Fournissez des outils pratiques.** Incluez des modèles et des procédures opératoires standard en annexe, afin que les équipes de terrain disposent d'outils concrets et non seulement de plans théoriques — cela signifie, par exemple, des scripts prêts à l'emploi pour le recrutement, des guides détaillés étape par étape pour la saisie des données, ainsi que des listes de contrôle pour la mise en œuvre de l'intervention.

Obtenir l'approbation éthique

L'approbation éthique est le processus formel d'obtention de l'autorisation d'un comité d'éthique ou du comité d'éthique de la recherche (CER), impliquant la navigation des exigences administratives, des délais et des procédures institutionnelles. Ce processus est indispensable pour garantir que l'évaluation puisse être menée légalement et de manière éthique.

Au-delà de la compréhension des principes éthiques, ce processus exige la gestion de tâches administratives concrètes, notamment :

- identifier quel CER ou comité d'éthique a juridiction (une université, un ministère de la santé, l'UNICEF ou plusieurs instances)
- compléter les formations obligatoires en éthique pour tous les membres de l'équipe
- préparer des dossiers documentaires complets dans des formats spécifiques
- répondre aux commentaires des évaluateurs et aux demandes de clarification
- maintenir la conformité tout au long de l'étude, y compris au travers des amendements, des déclarations d'événements indésirables et des renouvellements annuels

Pourquoi est-ce important ?

L'autorisation éthique est juridiquement obligatoire pour toute recherche impliquant des êtres humains : mener une évaluation sans cette approbation peut invalider l'ensemble de l'étude, exposer les institutions à des risques légaux et compromettre durablement la confiance des communautés. De plus, de nombreuses revues scientifiques refusent de publier des résultats sans preuve de validation éthique, et les bailleurs exigent de plus en plus cette autorisation avant de financer les projets. Le défi pratique est que l'examen éthique peut prendre entre 2 et 6 mois, avec plusieurs cycles de révision. Tout retard dans ce processus affecte l'ensemble du calendrier. Ainsi, même la meilleure conception d'évaluation devient inutilisable si la collecte de données ne peut commencer faute d'approbation. Assurer la conformité pendant toute la mise en œuvre impose également de disposer de systèmes permettant de documenter toute déviation du protocole, de signaler les événements indésirables et de garantir que tous les membres de l'équipe suivent les procédures approuvées.

Gardez à l'esprit ces considérations clés :

Commencez tôt. Initiez le processus éthique avant même que tous les détails ne soient finalisés — des amendements pourront être soumis par la suite. Obtenir l'approbation initiale permet de démarrer le calendrier.

Identifiez les autorités compétentes. Déterminez quelles instances doivent approuver le projet. Les IRB universitaires exigent souvent une affiliation ; les comités nationaux (ex. ministère de la santé) sont parfois requis pour les recherches liées à la santé ; plusieurs approbations peuvent être nécessaires dans des projets multi-pays ; certains bailleurs ont également leurs propres exigences éthiques.

Prévoyez des délais réalistes. Comptez : 2 à 3 semaines pour préparer le dossier, 4 à 8 semaines pour l'examen initial (plus long en cas d'examen en comité plénier), 2 à 3 semaines pour répondre aux commentaires, 2 à 4 semaines pour l'approbation finale, auxquels peuvent s'ajouter des délais locaux ou nationaux supplémentaires.

Préparez un dossier complet. Celui-ci doit inclure : un protocole détaillé (contexte, objectifs, méthodes), des formulaires de consentement dans toutes les langues locales avec rétrotraduction, les questionnaires (même encore en cours d'ajustement), les CV et attestations de formation des personnels impliqués, un plan de gestion des données avec mesures de sécurité, les modalités de compensation avec justification, ainsi que les protocoles d'atténuation des risques et d'orientation.

Comprenez le niveau d'examen susceptible d'être requis, car cela influe sur le calendrier. Une revue exemptée (risque minimal, catégories spécifiques) prend 2 à 3 semaines ; une revue accélérée (risque minimal, non exemptée) prend 4 à 6 semaines ; une revue complète en comité plénier (risque plus qu'à minimal, populations vulnérables) peut prendre 2 à 3 mois et ne se tient en général qu'une fois par mois.

Éviter les erreurs courantes telles que : dossiers incomplets, formulaires de consentement trop techniques ou incomplets, évaluation des risques insuffisante, montants de compensation jugés coercitifs, procédures de protection des données peu claires, signatures manquantes, ou absence d'autorisation institutionnelle.

Une fois l'autorisation obtenue, veillez à respecter la conformité. Cela implique de former tout nouveau membre de l'équipe, de documenter et signaler toute déviation au protocole, de soumettre tout amendement avant sa mise en œuvre, de fournir des rapports de suivi annuels, puis de clôturer officiellement l'étude à son terme.

Suivi de la mise en œuvre et évaluation du processus

Il s'agit d'un système complet de suivi de la manière dont l'intervention est réellement mise en œuvre sur le terrain. Il combine des routines de suivi opérationnel régulier avec une documentation systématique des processus. Cela implique des points de contrôle structurés (quotidiens, hebdomadaires ou bihebdomadaires selon l'intensité de l'intervention), à l'aide d'outils standardisés permettant d'évaluer plusieurs dimensions de la mise en œuvre :

- **Fidélité :** l'intervention a-t-elle été délivrée conformément à sa conception initiale ?
- **Couverture :** quelle proportion de la population cible a effectivement reçu l'intervention ?
- **Dose :** quelle a été la fréquence et l'intensité de la mise en œuvre ?
- **Qualité :** dans quelle mesure l'intervention a-t-elle été mise en œuvre correctement ?
- **Engagement des participants :** les participants ont-ils compris, participé et appliqué ce qui leur a été proposé ?
- **Facteurs contextuels :** quels éléments du contexte ont influencé la mise en œuvre ?

Il comprend le suivi en temps réel des taux de recrutement par rapport aux objectifs, le suivi des tendances d'attrition afin de préserver la puissance statistique, la documentation de toutes les adaptations apportées lors de la mise en œuvre, ainsi que la collecte de retours de la part des équipes de terrain et des participant·es sur ce qui fonctionne et ce qui ne fonctionne pas.

Pourquoi est-ce important ?

De nombreuses évaluations n'observent aucun impact — non pas parce que l'intervention est inefficace, mais parce qu'elle n'a jamais été correctement mise en œuvre. L'absence de suivi systématique représente un risque majeur : par exemple, après une coûteuse collecte de données de fin de programme, on peut découvrir que la moitié du groupe traité n'a jamais reçu l'intervention, que des personnes du groupe témoin y ont eu accès, ou encore que les équipes de terrain ont modifié l'intervention au point de la rendre méconnaissable. L'évaluation des

processus permet de distinguer une défaillance de la théorie (l'intervention ne fonctionne pas, même lorsqu'elle est correctement appliquée) d'une défaillance de la mise en œuvre (l'intervention n'a pas été appliquée comme prévu, ce qui empêche d'en évaluer réellement l'efficacité).

Ces informations sont essentielles pour interpréter les résultats. Par exemple, si aucun impact n'est observé, est-ce parce que la théorie était erronée ou parce que seulement 30 % des participants ont réellement reçu l'intervention complète ? À l'inverse, si des effets positifs sont constatés, comprendre ce qui a effectivement été mis en œuvre facilite la reproduction de l'intervention ailleurs. Le suivi en temps réel permet d'apporter des corrections alors qu'il est encore possible d'agir. Par exemple, si le recrutement prend du retard, les efforts peuvent être intensifiés avant que cela ne menace la puissance statistique ; si certains sites n'assurent pas une mise en œuvre adéquate, un soutien supplémentaire peut être fourni ; si des barrières inattendues apparaissent, des solutions peuvent être développées.

Les données issues de l'évaluation des processus fournissent également des informations déterminantes pour les décisions d'extension à plus grande échelle : quels contextes ont facilité une mise en œuvre fluide, quelles adaptations ont été nécessaires, quels défis risquent de persister lors du passage à l'échelle, et quel niveau de qualité est réellement atteignable dans les conditions routinières, par opposition aux conditions contrôlées de recherche.

Gardez à l'esprit ces considérations clés :

- **Mettez en place des routines de suivi adaptées à l'intensité de la mise en œuvre, mais qui ne surchargent pas les équipes de terrain.** Cela peut se traduire par des réunions quotidiennes pour les interventions intensives, des appels hebdomadaires pour les programmes standards ou des bilans mensuels pour les interventions légères.
- **Créez des outils simples.** Des fiches de suivi standardisées peuvent permettre de collecter les informations essentielles sans générer une charge administrative excessive.
- **Suivez les indicateurs clés de mise en œuvre séparément des données de résultats.** Cela peut inclure le pourcentage de sessions prévues réalisées, les taux moyens de fréquentation ou de participation, le pourcentage de personnes ayant reçu la dose complète de l'intervention, le temps entre les différentes composantes de l'intervention ou les notes de qualité issues d'observations standardisées.
- **Mettez en place des systèmes de contrôle de la qualité des données.** Ceux-ci peuvent combiner : des contrôles à haute fréquence (examens automatisés quotidiens ou hebdomadaires pour repérer les valeurs aberrantes, données manquantes ou anomalies), et des contre-vérifications (réinterroger 10 à 20 % des participants pour vérifier l'exactitude des données et détecter d'éventuelles fraudes).
- **Surveillez en permanence les facteurs susceptibles d'affecter la puissance statistique :** examinez les taux de recrutement par rapport aux objectifs (le projet est-il en voie d'atteindre la taille d'échantillon prévue ?), taux d'attrition global (quel pourcentage est perdu de vue ?), attrition différentielle (le taux d'abandon est-il plus élevé dans le groupe traité ou témoin ?), taux de conformité (quel pourcentage du groupe traité reçoit effectivement l'intervention ?), et contamination (le groupe témoin a-t-il accès à l'intervention ?).
- **Documentez chaque adaptation à l'aide de cadres structurés.** Notez précisément ce qui a été modifié par rapport au protocole, la raison de ce changement (barrière rencontrée, demande d'un partenaire, contrainte opérationnelle), le moment de l'adaptation, la personne ayant pris la décision, si le changement était planifié ou réactif, et enfin, si celui-ci remet en cause la théorie d'intervention.
- **Créez des boucles de rétroaction rapides avec les équipes de mise en œuvre.** Par exemple, utilisez des groupes WhatsApp pour résoudre les problèmes en temps réel, organisez de courtes réunions hebdomadaires axées sur les difficultés rencontrées et les solutions possibles, et réalisez des revues mensuelles des données de suivi afin d'identifier les tendances.
- **Distinguez les composantes essentielles des composantes adaptables.** Différenciez les éléments fondamentaux, qui doivent impérativement être maintenus pour garantir la validité de la théorie d'intervention, des éléments périphériques pouvant être ajustés au contexte. Documentez ces deux catégories, mais traitez-les différemment lors de l'analyse.

- **Recueillez le retour des participant·e·s.** Intégrez des mécanismes de rétroaction, tels que de brefs entretiens en fin de session, des groupes de discussion périodiques avec les participant·e·s, ainsi que des dispositifs anonymes comme des boîtes à suggestions ou des lignes téléphoniques dédiées.
- **Suivez les facteurs contextuels susceptibles d'influencer la mise en œuvre ou les résultats.** Notez, par exemple : la présence d'autres programmes ou politiques visant la même population, les facteurs saisonniers (jours fériés, saisons agricoles, conditions climatiques), la situation politique ou sécuritaire, ainsi que les crises sanitaires ou autres perturbations éventuelles.
- **Tenez des registres détaillés.** Ceux-ci sont essentiels pour interpréter les résultats, guider les décisions de mise à l'échelle et contribuer au renforcement de la base de connaissances sur les défis et solutions liés à la mise en œuvre.

Analyse coûts-bénéfices

Il s'agit d'un calcul systématique de l'ensemble des ressources nécessaires pour mettre en œuvre l'intervention et atteindre les impacts mesurés, afin de produire des indicateurs standardisés qui permettent de comparer différentes interventions, modalités de mise en œuvre ou options d'investissement. Cette comptabilité complète dépasse les simples budgets de programme et inclut le coût économique réel de l'obtention des résultats, notamment :

- les coûts directs du programme (personnel, matériel, opérations)
- les coûts indirects souvent répartis dans d'autres budgets (supervision, administration, frais généraux)
- les coûts d'opportunité des ressources mobilisées (temps des volontaires, temps du personnel gouvernemental, temps des participant·e·s)
- les coûts de démarrage versus les coûts de fonctionnement
- les coûts marginaux liés à l'ajout de nouveaux participant·e·s

L'analyse génère des indicateurs tels que : coût par enfant vacciné, coût par point de pourcentage d'augmentation des scores scolaires, coût par année de vie sauvée, ou ratio de retour sur investissement, que les décideurs peuvent comparer à des références ou à d'autres interventions. Consultez l'Outil d'analyse coûts-bénéfices ici pour vous aider à identifier, quantifier et comparer les coûts et les bénéfices des interventions. Pour voir un exemple d'application de cet outil à **l'étude de cas au Liban**, reportez-vous à l'annexe 2.

Pourquoi est-ce important ?

Même des interventions très efficaces peuvent ne pas être adaptées à un passage à l'échelle si elles s'avèrent trop coûteuses par rapport à d'autres options possibles. L'analyse coûts-bénéfices permet alors de transformer l'évaluation, qui ne se limite plus à démontrer qu'une intervention fonctionne, pour devenir une véritable orientation stratégique en répondant à la question essentielle : « Ce programme constitue-t-il un bon usage des ressources ? » Cette analyse est particulièrement déterminante dans un contexte où les bailleurs exigent des preuves combinant impact et efficience, où les gouvernements, contraints par des budgets limités, doivent maximiser les résultats obtenus pour chaque unité dépensée, et où les décisions de mise à l'échelle requièrent une compréhension fine de la manière dont les coûts évoluent lorsque l'intervention change d'ampleur.

Un programme qui génère une amélioration de 10 % peut sembler performant, jusqu'au moment où l'on découvre qu'il coûte cinq fois plus cher qu'une alternative permettant une amélioration de 8 %. Comprendre la structure des coûts permet également de repérer des marges d'efficience : il est parfois possible d'atteindre 80 % de l'impact pour 50 % du coût en simplifiant certaines composantes, ou de constater que des coûts fixes élevés lors d'un projet pilote deviennent négligeables à grande échelle. Sans une analyse rigoureuse des coûts, des programmes risquent d'être abandonnés à tort parce qu'ils paraissent « trop chers » sur la base d'informations incomplètes, ou au contraire d'être étendus avec enthousiasme sans que leur structure de coûts ne soit viable à long terme.

Gardez à l'esprit ces considérations clés :

- **Commencez tôt.** Il est essentiel de collecter les données de coûts dès le premier jour de mise en œuvre : reconstruire les coûts a posteriori est souvent peu fiable et parfois impossible lorsque les reçus sont égarés, que le personnel oublie comment son temps a été alloué ou que les contributions en nature n'ont pas été documentées.
- **Saisissez tous les coûts, pas seulement ceux du budget.** Il peut s'agir, par exemple, du temps du personnel, y compris pour la préparation et la formation (même si celles-ci sont financées par des partenaires), du temps des bénévoles valorisé au taux salarial local pour un travail équivalent, du temps du personnel gouvernemental même s'il n'est pas rémunéré par le projet, ainsi que des coûts supportés par les participants (transport, perte de revenus, garde d'enfants). Les matériaux ou locaux mis à disposition doivent être valorisés au prix du marché, et les frais généraux imputables au projet doivent également être inclus.
- **Distinguez les différentes catégories de coûts, car elles évoluent différemment à l'échelle.** Cela inclut les coûts fixes (développement de la formation, mise en place initiale) qui ne varient pas en fonction du nombre de participants, les coûts variables (matériels par participant, incitations) qui augmentent de manière linéaire, et les coûts « en paliers » (supervision, ouverture de nouveaux sites) qui augmentent à certains seuils.
- **Suivez les coûts selon plusieurs perspectives, car les parties prenantes ne se préoccupent pas des mêmes indicateurs.** Il est utile de considérer la perspective de l'exécutant (quel est le coût pour l'équipe qui met en œuvre ?), celle du gouvernement (combien coûterait l'intégration dans les systèmes existants ?), celle de la société (en incluant tous les coûts, quel que soit celui qui les finance), ainsi que celle des participants (quel est le coût de leur participation ?).
- **Calculez plusieurs mesures de coût-efficacité afin de permettre différentes comparaisons.** Cela peut inclure le coût par participant atteint/inscrit/ayant complété l'intervention ; le coût par unité de changement sur l'indicateur principal ; le coût par taille d'effet standardisée pour des comparaisons académiques ; ou encore les ratios de coût-efficacité incrémentale lorsque plusieurs variantes sont comparées.
- **Comparez les résultats à des références pertinentes.** Tenez compte d'interventions similaires dans le même contexte, de la disposition du gouvernement à payer pour des résultats comparables, de normes internationales (comme les seuils de l'OMS pour les interventions de santé) et d'alternatives pouvant atteindre les mêmes objectifs.
- **Intégrez des analyses de sensibilité montrant comment le rapport coût-efficacité varie selon différents scénarios.** Par exemple, les effets persistent-ils un an ou deux ? L'intervention est-elle mise en œuvre avec des salaires gouvernementaux ou des ONG ? Que se passe-t-il à différentes échelles (100, 1 000 ou 10 000 participants) ? Comment les résultats changent-ils avec des niveaux différents de conformité ou d'attrition ?
- **Documentez de façon transparente les facteurs qui influencent les coûts.** Qu'est-ce qui rend cette intervention coûteuse ou abordable ? Certains éléments peuvent-ils être modifiés pour réduire les coûts sans compromettre l'efficacité ? Quelles économies ou inefficiences apparaîtront à grande échelle ? Quels coûts cachés pourraient survenir dans une mise en œuvre en routine plutôt qu'en contexte de recherche ?
- **Présentez les résultats de manière accessible pour les décideurs.** Cela peut inclure des indicateurs simples tels que le coût par résultat obtenu, plutôt que des modèles économiques complexes, ainsi que des comparaisons visuelles avec des interventions connues, des indications claires sur les intervalles de confiance, et des implications concrètes pour la planification budgétaire.
- **Interprétez avec discernement les résultats en matière de coût-efficacité.** Le moins coûteux n'est pas toujours le meilleur : certaines interventions plus coûteuses peuvent être justifiées si leurs effets sont proportionnellement plus importants, ou si elles permettent d'atteindre des populations que des alternatives moins onéreuses ne parviennent pas à toucher.

Des tests rigoureux renforcent le lien entre la conception des programmes et la prise de décision. En fondant les

conclusions sur des preuves plutôt que sur des suppositions, les évaluations d'impact permettent à l'UNICEF et à ses partenaires d'identifier les interventions efficaces, celles qui nécessitent des adaptations et celles qui devraient être abandonnées. Grâce à des mesures systématiques et à une analyse transparente, les connaissances comportementales et les hypothèses de conception sont transformées en données probantes crédibles et exploitables, qui éclairent les décisions relatives à la mise à l'échelle, aux politiques publiques et à l'allocation des ressources.

À mesure que les programmes passent à la phase *« Mettre à l'échelle »* du processus DEPTHS, les preuves générées par *les hypothèses de test* sont appliquées. Les résultats issus d'évaluations rigoureuses orientent la manière dont les interventions sont ajustées, intégrées aux systèmes existants et développées de manière responsable. Cela garantit que les décisions d'élargissement ou de reproduction des interventions reposent sur des impacts démontrés, et non sur des suppositions.

Annexe

ANNEXE 1 :

Alternatives quasi expérimentales lorsque la randomisation n'est pas possible

Cette annexe fournit des orientations techniques supplémentaires pour les praticiens qui conçoivent des évaluations d'impact dans des conditions réelles et routinières. Bien que le chapitre principal mette en avant les dispositifs randomisés comme la méthode la plus fiable pour établir une relation de causalité, certaines contraintes pratiques peuvent parfois en limiter l'usage. Dans ces cas, les évaluateurs peuvent être amenés à envisager des approches alternatives à la randomisation, tout en cherchant à produire des résultats crédibles et fondés sur des données probantes.

La priorité dans la conception d'une évaluation doit être de recourir à la randomisation, car celle-ci demeure la méthode la plus fiable pour établir un lien de causalité. La randomisation élimine les différences systématiques entre les groupes, offrant ainsi une plus grande confiance dans le fait que les effets observés sont attribuables à l'intervention elle-même. Toutefois, il existe des situations où la randomisation n'est pas possible en raison de contraintes politiques, éthiques, logistiques ou pratiques. Dans ces cas, des approches alternatives peuvent être envisagées.

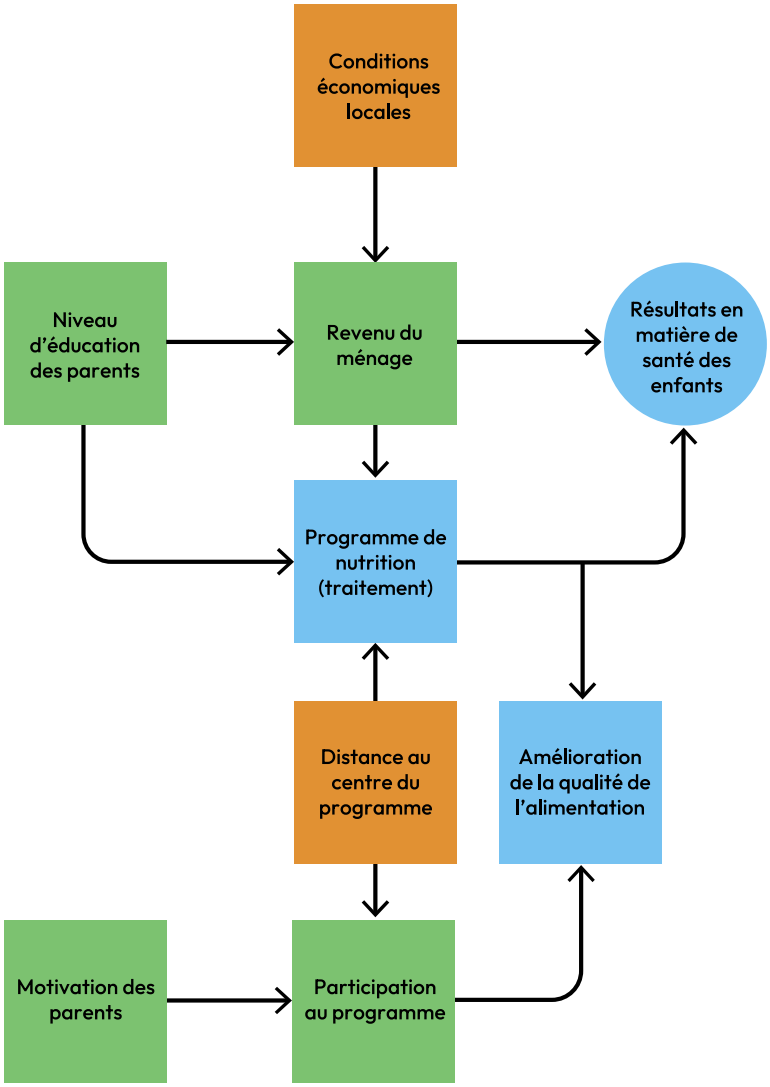
Bien que ces approches puissent fournir des enseignements utiles, elles impliquent une complexité opérationnelle accrue et des exigences statistiques plus importantes. Surtout, elles présentent un risque plus élevé de biais. Il est donc fortement recommandé de mobiliser des experts en inférence causale et en conception d'évaluation lorsque des dispositifs non randomisés sont utilisés.

1. **Fermer les chemins rétrospectifs : une perspective basée sur les DAG**. Lorsque la randomisation n'est pas possible, il devient essentiel d'identifier et de contrôler les facteurs de confusion — une approche connue, en inférence causale, sous le nom de « fermeture des chemins rétrospectifs ». Les graphiques acycliques dirigés (DAGs) offrent un cadre utile pour comprendre ce processus.

   Considérons un DAG illustrant l'évaluation d'un programme communautaire de nutrition ciblant les enfants :

   Dans un tel diagramme :

   - Les flèches représentent des effets causaux directs
   - Les variables (nœuds) désignent les facteurs qui influencent les résultats ou la participation au programme
   - Les chemins entre les variables représentent des associations potentielles, qu'elles soient causales ou non causales



2. **Identification des chemins rétrospectifs.** Un « chemin rétrospectif » est tout chemin reliant le traitement (participation au programme nutritionnel) et le résultat (état de santé de l'enfant) qui ne suit pas la direction causale directe. Ces chemins introduisent des associations susceptibles de biaiser l'estimation de l'impact.

   Dans l'exemple de DAG, plusieurs chemins de biais peuvent exister:

   - **Programme → Revenu du ménage → Santé infantile** : les familles ayant un revenu plus élevé peuvent être à la fois plus susceptibles de participer au programme et d'avoir des enfants en meilleure santé, indépendamment du programme.
   - **Programme ← Niveau d'éducation des parents → Santé de l'enfant** : des parents plus éduqués peuvent s'inscrire plus fréquemment au programme tout en fournissant de meilleurs soins à leur enfant.
   - **Programme ← Niveau d'éducation des parents → Revenu du ménage → Santé de l'enfant** : le niveau d'éducation des parents influence le revenu, qui affecte à son tour la participation et les résultats sanitaires.

3. **L'objectif : fermer tous les chemins rétrospectifs.** Pour isoler l'effet causal du programme nutritionnel, tous les chemins de biais doivent être fermés. Un chemin est considéré comme fermé lorsqu'une des conditions suivantes est remplie :

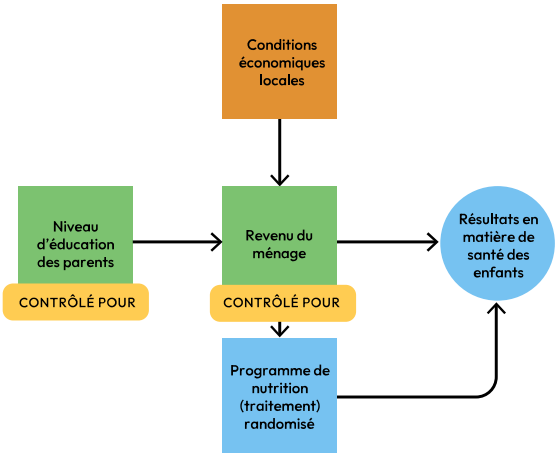
    - **Une variable située sur ce chemin est contrôlée (conditionnée).** Par exemple : contrôler le revenu du ménage ferme le chemin Programme → Revenu du ménage → Santé de l'enfant.
    - **Le chemin contient un collisionneur.** Un collisionneur est une variable influencée par deux variables ou plus. Par exemple, « Participation au programme » peut être influencée à la fois par la « Distance au centre » et la « Motivation des parents ». Ce chemin est fermé naturellement, sauf si l'on conditionne à tort sur le collisionneur, ce qui aurait pour effet de le rouvrir.
    - **Le chemin comprend un médiateur qui est intentionnellement laissé sans contrôle.** Par exemple, « Amélioration de la qualité de l'alimentation » se trouve sur la chaîne causale entre la participation au programme et l'état de santé de l'enfant. Si l'objectif est d'estimer l'effet total du programme, cette variable ne doit pas être contrôlée

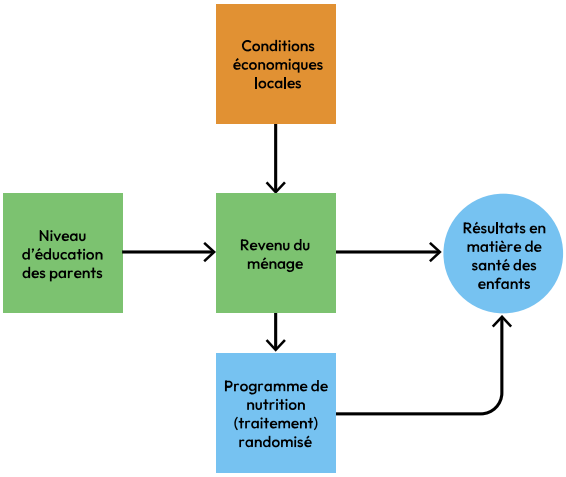
    Lorsque la randomisation est appliquée, toutes les flèches pointant vers le nœud « Participation au programme » sont, en pratique, coupées. La participation devient alors indépendante de tous les facteurs de confusion, ce qui ferme simultanément l'ensemble des chemins rétrospectifs. C'est l'avantage central de la randomisation : elle supprime la nécessité d'identifier et de mesurer chaque source potentielle de biais.

4. **Fermer les chemins rétrospectifs sans randomisation.** En l'absence de randomisation, des méthodes statistiques sont utilisées pour fermer les « backdoor paths ». Deux stratégies courantes incluent :

   a. **Contrôle des facteurs de confusion observés.** Cela consiste à mesurer et à ajuster les variables telles que le niveau d'éducation des parents ou le revenu du ménage à l'aide de méthodes comme les ajustements par régression ou l'appariement. Cette approche exige :

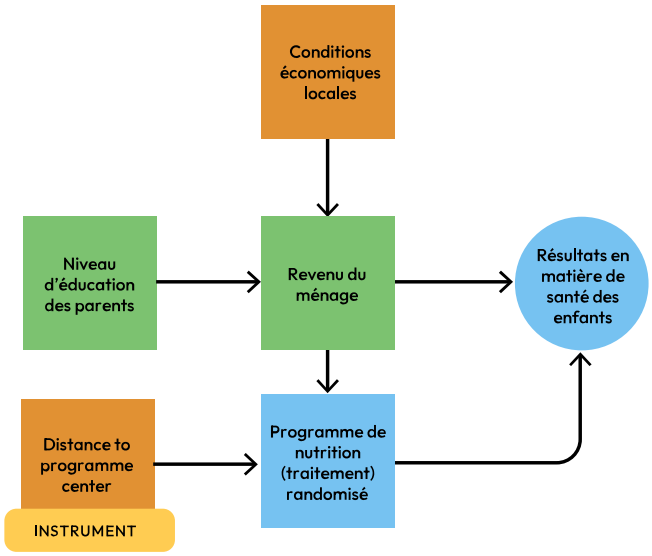
      - l'identification de tous les facteurs de confusion pertinents
      - la mesure précise de ces facteurs
      - la modélisation correcte de leurs relations avec le traitement et les résultats.





b. **Utilisation des variables instrumentales.** Cette méthode repose sur une variable qui influence la participation au programme sans être directement liée au résultat. Par exemple, la distance au centre du programme peut déterminer la participation, sans affecter directement la santé de l'enfant. Cette variation peut être utilisée pour estimer des effets causaux même en présence de facteurs de confusion non observés.

Les opportunités et les limites de chacune de ces méthodes dépendent du contexte et de la disponibilité des données. Différentes situations offrent différents leviers pour fermer les chemins de biais, et le choix de l'approche appropriée nécessite une analyse minutieuse des contraintes liées à la conception de l'étude et aux données disponibles.



Le tableau 3 de l'étape 1 de ce chapitre comprend une description des différentes approches à utiliser lorsque la randomisation n'est pas possible.

ANNEXE 2 :

Le cas de l'augmentation de la couverture vaccinale des enfants au

Cette phase s'appuie également sur l'étude de cas présentée dans les phases précédentes, portant sur l'augmentation de la vaccination des enfants dans des contextes à faibles ressources au Liban. Elle illustre l'utilisation d'outils au sein du processus DEPTHS.

Contexte de l'étude de cas :

Au cours des phases *Définir, Explorer et Prototyper*, l'équipe au Liban a identifié le comportement cible à modifier — le retour des aidants avec leurs enfants pour l'administration de la prochaine dose de vaccin — ainsi que les principaux barrières comportementales associées : surcharge cognitive liée aux multiples tâches quotidiennes, stress émotionnel lors des visites en clinique, et influence des perceptions sociales façonnées par la communauté. Ils ont également testé un prototype de solution potentielle : une carte de rendez-vous papier, conçue pour rendre la prochaine visite vaccinale plus saillante et plus facile à mémoriser.

L'intervention a été évaluée au moyen d'un essai randomisé contrôlé incluant 12 332 enfants non ou sous-vaccinés, répartis dans 6 160 ménages. Les ménages ont été assignés aléatoirement soit à un groupe de traitement recevant la carte de rendez-vous lors des visites de sensibilisation, soit à un groupe de contrôle n'en recevant pas. Le résultat principal a été défini au niveau du ménage : au moins un enfant éligible du ménage recevait-il le vaccin requis dans un délai de 21 jours après son échéance ?

Cette mesure binaire, au niveau du ménage, a été retenue car elle reflète mieux la prise de décision des aidants, sachant que le coût marginal de la vaccination supplémentaire d'un second enfant dans un même foyer est généralement faible. L'analyse a utilisé une régression logistique pour estimer l'effet en intention de traiter, avec erreurs standards robustes regroupées au niveau des agents de sensibilisation. Les variables de base contrôlées incluaient : la taille du ménage, la nationalité et l'historique vaccinal antérieur.

L'essai a montré que les ménages recevant la carte de rendez-vous étaient significativement plus susceptibles de revenir pour la vaccination, avec une augmentation de 6,7 points de pourcentage par rapport au groupe de contrôle. Ces résultats indiquent qu'une intervention comportementale simple et peu coûteuse peut améliorer de manière mesurable la couverture vaccinale infantile lorsqu'elle vise des moments clés de décision. Les conclusions ont alimenté des discussions politiques en cours et démontrent l'utilité d'un processus structuré de test d'hypothèses dans les programmes de santé publique comportementale.

Conformément au guide, cette annexe présente un exemple complet afin d'illustrer l'application pratique des outils et d'aider les équipes à les utiliser facilement et de manière cohérente. Ces outils incluent :

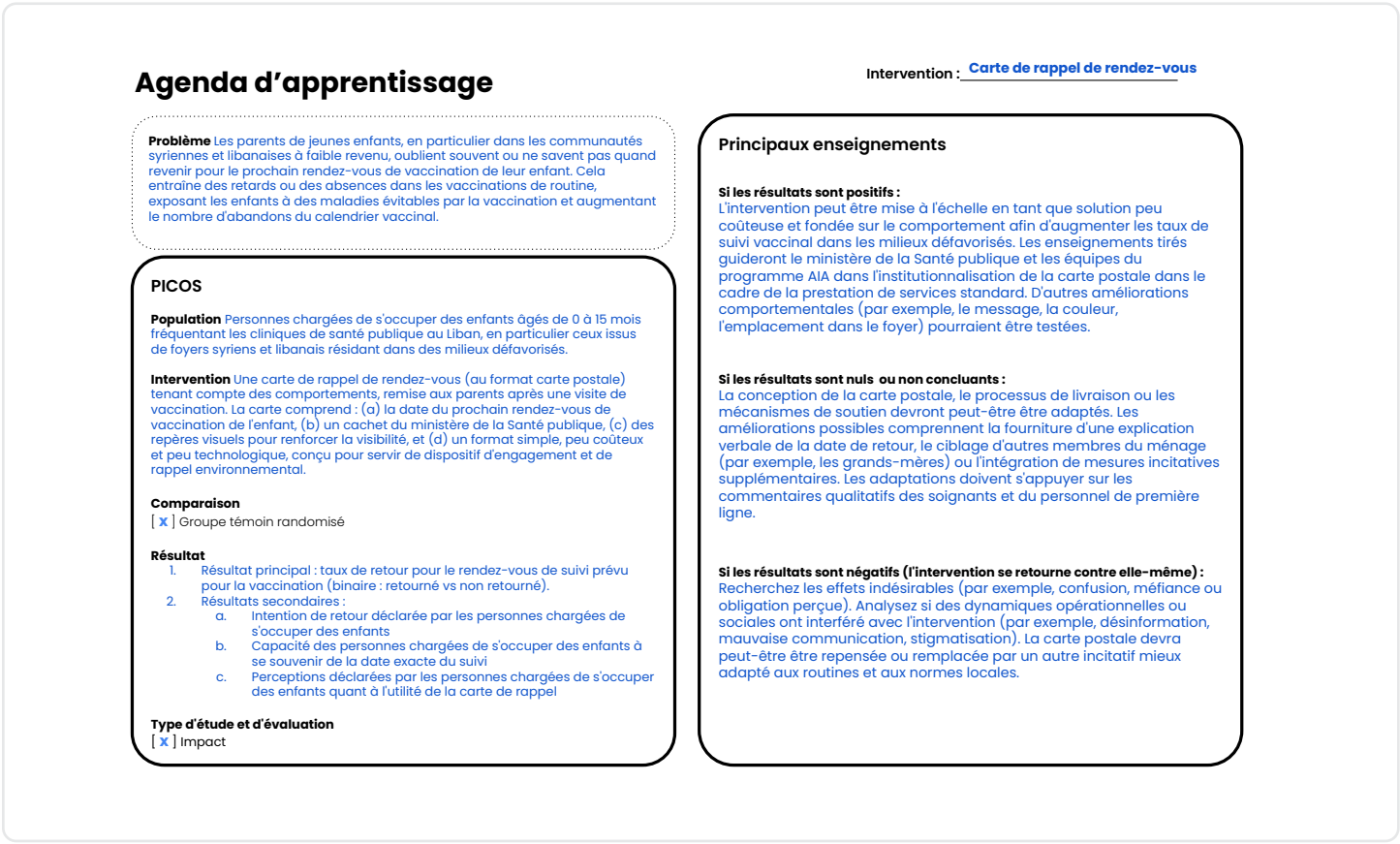
- Agenda d'apprentissage
- Plan de mise en œuvre
- Analyse des risques liés à la mise en œuvre
- Analyse coûts-bénéfices

Application de l'Agenda d'apprentissage

Cet Agenda d'apprentissage n'a pas été élaboré par l'équipe de projet initiale. Il s'agit d'un exemple reconstruit à partir de données et du contexte réel du projet.

La section suivante illustre comment appliquer **l'Agenda d'apprentissage** présenté dans le guide, en utilisant l'étude de cas menée au Liban. Cet outil montre comment formuler et hiérarchiser les principales questions d'apprentissage qui orientent à la fois la conception de l'évaluation et l'interprétation des résultats. L'exemple met en évidence qu'un Agenda d'apprentissage clair — centré sur ce que l'équipe doit apprendre, pourquoi cela est important et comment les résultats seront utilisés — garantit que l'évaluation soit intentionnelle, exploitable et adaptable.

Agenda d'apprentissage



Agenda d'apprentissage

Intervention : **Carte de rappel de rendez-vous**

Problème Les parents de jeunes enfants, en particulier dans les communautés syriennes et libanaises à faible revenu, oublient souvent ou ne savent pas quand revenir pour le prochain rendez-vous de vaccination de leur enfant. Cela entraîne des retards ou des absences dans les vaccinations de routine, exposant les enfants à des maladies évitables par la vaccination et augmentant le nombre d'abandons du calendrier vaccinal.

PICOS

Population Personnes chargées de s'occuper des enfants âgés de 0 à 15 mois fréquentant les cliniques de santé publique au Liban, en particulier ceux issus de foyers syriens et libanais résidant dans des milieux défavorisés.

Intervention Une carte de rappel de rendez-vous (au format carte postale) tenant compte des comportements, remise aux parents après une visite de vaccination. La carte comprend : (a) la date du prochain rendez-vous de vaccination de l'enfant, (b) un cachet du ministère de la Santé publique, (c) des repères visuels pour renforcer la visibilité, et (d) un format simple, peu coûteux et peu technologique, conçu pour servir de dispositif d'engagement et de rappel environnemental.

Comparaison
[x] Groupe témoin randomisé

Résultat
1. Résultat principal : taux de retour pour le rendez-vous de suivi prévu pour la vaccination (binaire : retourné vs non retourné).
2. Résultats secondaires :
   a. Intention de retour déclarée par les personnes chargées de s'occuper des enfants
   b. Capacité des personnes chargées de s'occuper des enfants à se souvenir de la date exacte du suivi
   c. Perceptions déclarées par les personnes chargées de s'occuper des enfants quant à l'utilité de la carte de rappel

Type d'étude et d'évaluation
[x] Impact

Principaux enseignements

Si les résultats sont positifs :
L'intervention peut être mise à l'échelle en tant que solution peu coûteuse et fondée sur le comportement afin d'augmenter les taux de suivi vaccinal dans les milieux défavorisés. Les enseignements tirés guideront le ministère de la Santé publique et les équipes du programme AIA dans l'institutionnalisation de la carte postale dans le cadre de la prestation de services standard. D'autres améliorations comportementales (par exemple, le message, la couleur, l'emplacement dans le foyer) pourraient être testées.

Si les résultats sont nuls ou non concluants :
La conception de la carte postale, le processus de livraison ou les mécanismes de soutien devront peut-être être adaptés. Les améliorations possibles comprennent la fourniture d'une explication verbale de la date de retour, le ciblage d'autres membres du ménage (par exemple, les grands-mères) ou l'intégration de mesures incitatives supplémentaires. Les adaptations doivent s'appuyer sur les commentaires qualitatifs des soignants et du personnel de première ligne.

Si les résultats sont négatifs (l'intervention se retourne contre elle-même) :
Recherchez les effets indésirables (par exemple, confusion, méfiance ou obligation perçue). Analysez si des dynamiques opérationnelles ou sociales ont interféré avec l'intervention (par exemple, désinformation, mauvaise communication, stigmatisation). La carte postale devra peut-être être repensée ou remplacée par un autre incitatif mieux adapté aux routines et aux normes locales.

L'équipe au Liban a d'abord élaboré un Agenda d'apprentissage simple visant à clarifier ce qu'elle souhaitait apprendre grâce à l'évaluation et pour quelles raisons. Elle a identifié un problème spécifique : les aidants libanais et syriens ne respectaient pas les calendriers de vaccination de leurs enfants, souvent parce qu'ils oubliaient ou comprenaient mal la date du prochain rendez-vous. L'intervention — une carte de rendez-vous conçue selon des principes comportementaux — visait à répondre à ce problème en servant de rappel physique.

La principale question d'apprentissage formulée était : « Le fait de fournir une carte personnalisée de rappel de rendez-vous augmente-t-il la réalisation en temps opportun des vaccinations infantiles au sein des communautés libanaises et syriennes à faible revenu ? »

Les questions secondaires portaient sur la manière dont l'intervention influençait l'intention de revenir à la clinique et le rappel de la date du suivi. Des interprétations anticipées des résultats ont également été cartographiées.

Par exemple, si la couverture vaccinale augmentait, l'intervention pourrait être mise à l'échelle. Si les résultats étaient mitigés ou nuls, des ajustements de conception et des tests supplémentaires pourraient être nécessaires. En cas d'effet inverse, l'équipe étudierait les effets indésirables tels que la méfiance ou les interprétations erronées.

Application du plan de mise en œuvre

Ce Plan de mise en œuvre n'a pas été élaboré par l'équipe de projet initiale. Il s'agit d'un exemple reconstruit à partir de données et du contexte réel du projet.

La section suivante illustre comment appliquer le Plan de mise en œuvre présenté dans le guide, en utilisant l'étude de cas menée au Liban. Cet outil montre comment traduire la conception de l'évaluation en une planification opérationnelle structurée — en définissant les activités clés, les rôles, les calendriers et les indicateurs de suivi. L'exemple fournit une référence pratique pour élaborer des plans de mise en œuvre favorisant la coordination, la transparence et la responsabilité tout au long du processus d'évaluation.

Élaboration d'un Plan de mise en œuvre

Pour opérationnaliser l'évaluation, l'équipe au Liban a également élaboré un Plan de mise en œuvre clair, décrivant les étapes pratiques, les responsabilités et le calendrier du déploiement. Ce plan détaille chaque phase, depuis la finalisation de la conception de la carte postale et la réalisation d'un pré-test restreint, jusqu'à la collecte des données de référence et à l'assignation des participants aux groupes de traitement ou de contrôle. Chaque étape était associée à des responsables désignés (par exemple : l'équipe de projet de l'UNICEF, J-PAL MENA, le Ministère de la Santé Publique), à un calendrier défini, ainsi qu'à des indicateurs spécifiques pour suivre les progrès, tels que les taux d'inscription, la couverture de distribution ou les contrôles de qualité des données. Ce plan a servi non seulement d'outil de coordination, mais également de base pour assurer transparence et responsabilisation tout au long de la mise en œuvre.

Plan de mise en œuvre

Intervention : Carte de rappel de rendez-vous

Utilisez cette fiche de travail pour décomposer votre intervention en étapes réalisables. Pour chaque domaine prioritaire, définissez les activités clés, attribuez les responsabilités, fixez un calendrier et identifiez les ressources et les indicateurs de suivi qui seront nécessaires pour suivre les progrès.

Domaine prioritaire *[Insérer la priorité]*	**Activités clés** *[Énumérer 1 à 2 actions]*	**Responsable** *[Responsable et soutiens]*	**Calendrier** *[MM/AA]*	**Ressources et budget** *[Contributions nécessaires]*	**Indicateurs de suivi** *[par exemple, % des activités réalisées]*
Essai pilote réalisé	Finalisation de la conception de la carte postale et réalisation d'un petit test préalable dans 1 à 2 cliniques	Équipe du programme AIA, unité de conception du ministère de la Santé publique, équipe du projet UNICEF	23/03	Consultant en conception, transport	Finalisation de la conception de la carte postale et intégration des commentaires
Collecte des données de référence	Collecte des données de référence sur les taux de retour des soignants et les performances des cliniques	Équipe d'évaluation (J-PAL MENA), ministère de la Santé publique, équipe de projet de l'UNICEF	03-04/23	Temps consacré par les enquêteurs, déplacements, tablettes	Pourcentage d'enquêtes de référence réalisées, contrôles de la qualité des données
Recrutement effectué	Identifier et inscrire les personnes chargées de s'occuper des enfants éligibles lors des visites régulières à la clinique	Agents de santé, personnel administratif de la clinique, équipe du projet UNICEF	05-07/23	Orientation du personnel, listes des cliniques	Nombre de soignants inscrits ; taux de consentement
Mission effectuée	Répartir de manière aléatoire les personnes chargées de s'occuper des enfants entre le groupe « carte postale » (traitement) et le groupe « soins standard » (contrôle)	Équipe d'évaluation, soutien à la mise en œuvre de l'AIA, équipe de projet de l'UNICEF	05-07/23	Protocole de Randomisation, formulaires de données	Randomisation terminée, vérification des contrôles d'équilibre
Début de l'intervention	Début de la distribution des cartes postales aux soignants du groupe de traitement après la vaccination	Personnel clinique, supervisé par l'équipe de terrain de l'AIA et l'équipe du projet de l'UNICEF	23/05	Cartes postales imprimées, formulaires de suivi	Pourcentage de soignants éligibles ayant reçu une carte postale
Vérification de la mise en œuvre	Effectuer des appels de supervision à mi-parcours et des visites sur le terrain pour évaluer la fidélité et la portée	Équipe de terrain de l'AIA, points focaux de district du MoPH, équipe de projet de l'UNICEF	23/06	Coûts liés aux communications téléphoniques/données et aux visites sur site	Pourcentage de cliniques rendant compte de la routine, écarts constatés
Fin de l'intervention	Conclure la phase de distribution et arrêter l'inscription de nouveaux participants	Personnel de la clinique, équipe d'évaluation informée, équipe du projet UNICEF	23/07	N/A	Date limite appliquée sur tous les sites pilotes
Données collectées	Réalisation d'enquêtes de suivi auprès des soignants ; extraction des dossiers de retour des cliniques	Enquêteurs J-PAL, personnel M&E des cliniques, équipe de projet UNICEF	08-09/23	Outils d'enquête, transport, incitations	Pourcentage d'enquêtes de suivi réalisées, dossiers extraits
Nettoyage et analyse des données	Nettoyage des ensembles de données, réalisation d'analyses statistiques sur les résultats primaires et secondaires	Analystes d'évaluation (J-PAL MENA), équipe de projet de l'UNICEF	10-11/23	Temps consacré par les analystes, logiciels	Plan d'analyse final achevé ; résultats validés

Munie du Plan d'évaluation et du Plan de mise en œuvre, l'équipe était prête à soumettre le projet à un comité d'éthique afin d'obtenir l'examen éthique et l'autorisation nécessaires avant de mettre en œuvre et d'évaluer l'intervention.

Application de l'outil d'atténuation des risques

Ce qui suit illustre comment utiliser **l'outil Risques de mise en œuvre et stratégies d'atténuation** présenté dans le guide principal, en s'appuyant sur l'étude de cas menée au Liban. Cet outil permet d'identifier et de gérer de manière systématique les risques opérationnels, contextuels et comportementaux susceptibles d'affecter l'évaluation. L'exemple ci-dessous constitue une référence pratique pour les équipes souhaitant anticiper et surmonter les défis de mise en œuvre, afin de préserver l'intégrité de l'évaluation et la réussite du programme.

Gestion des risques pendant la mise en œuvre

L'équipe a utilisé l'outil Risques de mise en œuvre et stratégies d'atténuation de façon dynamique, non seulement lors de la phase de planification, mais aussi tout au long de la mise en œuvre active. Trois principaux risques ont émergé :

RISQUES	PROBABILITÉ	IMPACT	MESURES D'ATTÉNUATION
Les cartes n'arrivent pas à temps	Moyenne	Élevé	l'équipe de l'UNICEF a travaillé avec un service de messagerie pour échelonner les calendriers de livraison et fournir un stock tampon.
Les infirmières ne remettent pas les cartes les explications appropriées	Élevée	Moyen	Un module de formation audio (note vocale) a été diffusé et des fiches explicatives plastifiées ont été distribuées.
Perte ou égarement de la carte par les aidants	Moyenne	Moyen	Utilisation de pochettes en plastique transparent et rappels de coller la carte postale près du calendrier à la maison.
Mise à jour irrégulière des registres de suivi	Moyenne	Élevé	Un rappel rapide a été intégré en fin de service via un message automatisé sur WhatsApp.
Perturbations contextuelles (grève des transports, conditions météorologiques, etc)	Faible à moyenne	Élevé	Des points focaux locaux ont été désignés afin d'adapter rapidement les itinéraires de distribution en fonction des contraintes.

Chaque risque a été attribué soit à l'équipe de l'UNICEF, soit à un superviseur du ministère de la Santé publique pour le suivi, avec des échéances alignées sur le calendrier de mise en œuvre.

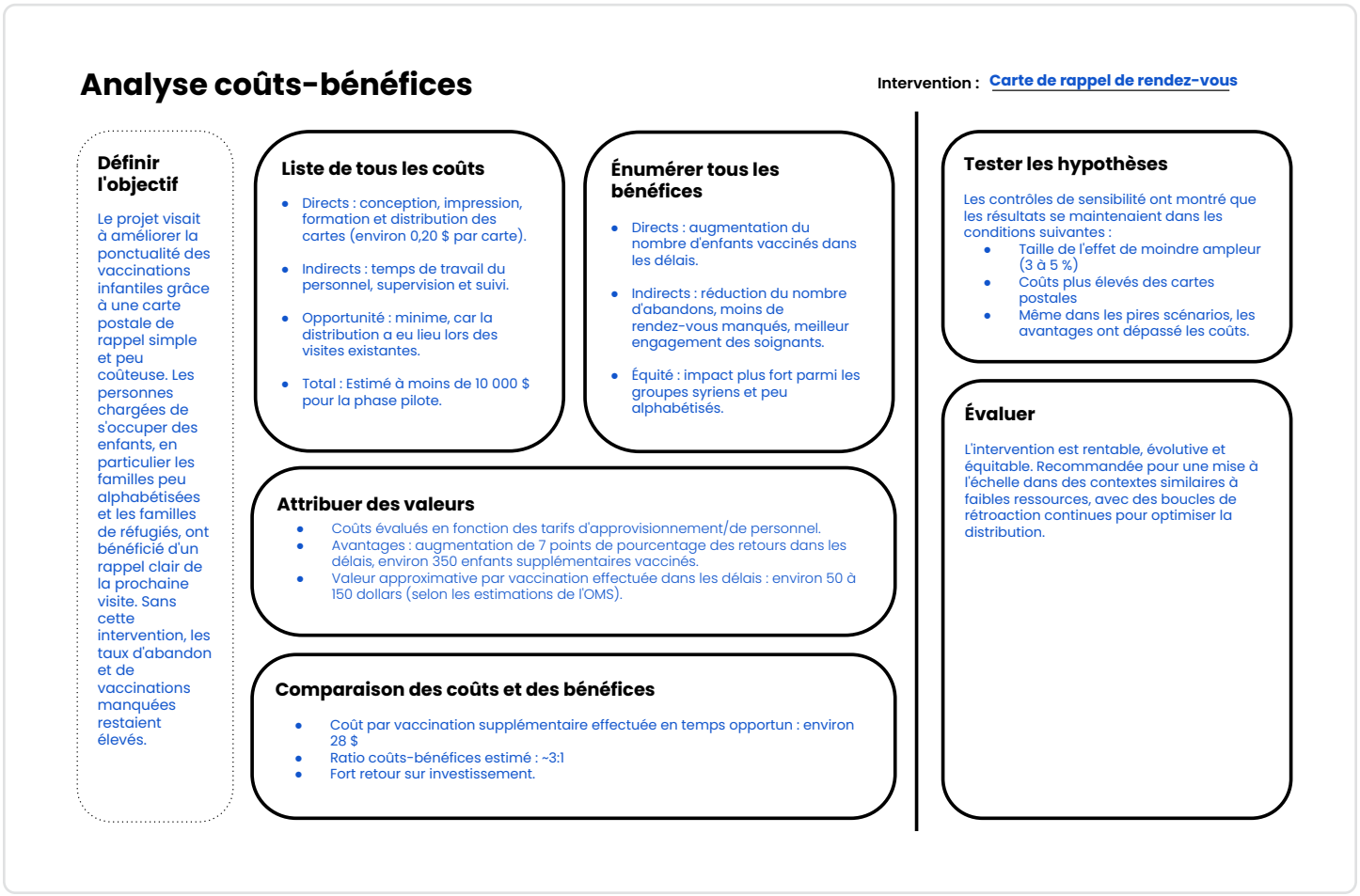
Application de l'analyse Coûts-Bénéfices

Ce qui suit illustre comment appliquer l'outil **d'Analyse Coûts-Bénéfices** présenté dans ce guide, en utilisant l'étude de cas menée au Liban. Cet outil montre comment identifier, quantifier et comparer de manière systématique les coûts et les bénéfices d'un programme afin d'évaluer sa rentabilité. L'exemple fournit une référence pratique pour les équipes cherchant à estimer l'efficacité économique d'interventions comportementales et à orienter les décisions de mise à l'échelle, d'adaptation ou d'allocation des ressources.

Analyse Coûts-Bénéfices

L'objectif principal de l'intervention était d'accroître la couverture vaccinale en temps opportun dans des cliniques urbaines à faibles ressources au Liban. En l'absence d'intervention, les données indiquent que de nombreux aidants retardaient leur retour à la clinique, augmentant ainsi le risque de vaccinations manquées ou incomplètes. La carte-rappel, conçue comme un signal comportemental simple, visait à modifier cette trajectoire.

Les coûts de l'intervention étaient relativement modestes. Bien que l'étude n'ait pas rapporté directement les données financières, une estimation raisonnable situe le coût par carte imprimée à moins de 0,20 USD, y compris la conception



et la distribution. Pour un pilote de petite envergure (par exemple, moins de 10 000 aidants), le coût total serait probablement inférieur à 10 000 USD, en incluant les matériaux, la supervision et le temps de travail du personnel.

Du côté des bénéfices, l'étude a montré que la carte augmentait les retours en clinique de sept points de pourcentage. Selon la littérature en santé publique, chaque vaccination réalisée en temps opportun contribue, à long terme, à la prévention des maladies, à la réduction de la mortalité infantile et à la diminution des dépenses de santé. Selon des estimations de l'OMS, la valeur sociétale d'une vaccination réalisée à temps peut être évaluée entre 50 et 150 USD. Appliquée à l'ensemble du groupe exposé, cette estimation se traduit par des gains importants, avec un ratio bénéfice-coût vraisemblablement compris entre 3:1 et 5:1.

Afin d'évaluer la robustesse de ces résultats, l'équipe a testé plusieurs hypothèses. Même dans les scénarios les plus pessimistes (par exemple, coûts plus élevés ou effets plus faibles), l'intervention restait rentable, en grande partie grâce à son faible coût unitaire et à son potentiel de mise à l'échelle.

L'équipe a également pris en compte les enjeux d'équité et d'inclusion. L'intervention a été particulièrement bénéfique pour les aidants marginalisés, notamment les familles syriennes, ce qui souligne son potentiel comme stratégie à faible coût réduisant les inégalités d'accès à la vaccination. La carte-rappel n'était donc pas seulement rentable, mais aussi porteuse d'équité — un critère clé pour les décisions futures d'expansion.

Pour en savoir plus

Ce guide fournit des outils pratiques, des cadres et des supports pour aider les équipes à appliquer les sciences comportementales à des défis réels de développement. Cependant, aucun guide ne peut tout couvrir. Les sciences comportementales se situent à l'intersection de plusieurs disciplines, notamment le design centré sur l'humain, la science de l'implémentation, l'éthique, la mesure et l'évaluation. Cette section vise donc à offrir des ressources complémentaires à celles et ceux qui souhaitent approfondir leurs connaissances, intégrer davantage de rigueur éthique, améliorer la conception de l'implémentation ou affiner la sélection des mesures de résultat. Les sources ci-dessous constituent des points de départ pour un apprentissage autonome.

« Je souhaite obtenir un guide plus détaillé, étape par étape, sur la manière de conduire des expérimentations. »

Il existe de nombreux manuels, ressources et cours dédiés aux expérimentations appliquées aux programmes sociaux. Parmi les ressources gratuites utiles, citons le site web du J-PAL sur l'introduction aux évaluations randomisées et le guide pratique de l'Évaluation d'impact de la Banque mondiale.

« Je souhaite améliorer mon approche de l'éthique dans les sciences comportementales appliquées. »

L'éthique est essentielle pour tout projet impliquant des personnes, qu'il s'agisse de rédiger des formulaires de consentement, d'évaluer des risques ou de gérer des asymétries de pouvoir. Les ressources suivantes offrent des orientations pratiques :

- La boîte à outils éthique de l'UNICEF pour les projets de sciences comportementales appliquées aide les équipes à réfléchir dès le début aux risques éthiques et à intégrer des mesures de protection tout au long de la mise en œuvre.
- La procédure de l'UNICEF sur les normes éthiques en matière de recherche, d'évaluation, de collecte et d'analyse de données qui présente les protocoles et exigences institutionnels.
- La liste de contrôle pour le consentement éclairé (J-PAL) est un modèle annoté expliquant les éléments à intégrer dans les formulaires de consentement.
- Les modèles de consentement de l'UNICEF (voir page 41) comprenant des formulaires éditables pour les participants, les aidants et les responsables légaux.

« Je dois suivre un cours d'éthique pour un comité d'éthique. »

Plusieurs organismes proposent des formations donnant droit à un certificat reconnu par divers comités éthiques. Pour des formations externes :

- Formation sur la protection des personnes participant à la recherche du HHS (basée aux États-Unis, certification gratuite, environ 5 à 6 heures)
- Cours sur la politique des trois conseils en Éthique de la recherche (basé au Canada, certification gratuite, environ 4 heures)

« Je souhaite améliorer la manière dont je conçois et mesure l'implémentation. »

Comprendre ce qui a été fait et comment cela a été fait est essentiel pour savoir si une intervention comportementale a fonctionné. L'Implémentation Outcome Repository propose des conseils et des exemples pour mesurer des concepts tels que la faisabilité, la fidélité et l'acceptabilité..

« Je souhaite améliorer ma manière de sélectionner ou d'adapter les mesures des résultats. »

Une bonne mesure ne se limite pas à tester l'efficacité ; elle doit capturer le comportement cible de manière appropriée. Pour les mesures portant sur le changement comportemental ou sur des indicateurs indirects, la

ressource Psychometric Properties of Implementation Measures analyse la validité et la fiabilité d'outils fréquemment utilisés en science de l'implémentation.

« Je souhaite évaluer la qualité et la rigueur des rapports et des études d'évaluation. »

Lorsqu'on examine, commande ou interprète des études, il est essentiel de savoir si les résultats sont fiables. Les outils suivants aident à évaluer la rigueur des études, qu'elles soient quantitatives, qualitatives ou mixtes.

Évaluer la conception globale et la rigueur des rapports

- L'outil d'évaluation DAC (DAT) permet d'évaluer si une étude a été correctement conçue, analysée et rapportée.
- L'article « Publier des articles quantitatifs avec rigueur et transparence » propose des recommandations accessibles pour produire des résultats transparents et robustes.

Évaluer la rigueur des revues systématiques

L'outil Evidence Project Risk of Bias Tool aide à évaluer la rigueur des études randomisées et non randomisées dans les revues systématiques et les synthèses mixtes.

Évaluer la rigueur de la recherche qualitative

- L'article Indicators of Rigor in Qualitative Research décrit comment évaluer la crédibilité, la transférabilité et la fiabilité des recherches qualitatives.
- La puissance informationnelle dans l'échantillonnage qualitatif propose une alternative au concept de « saturation » pour justifier les tailles d'échantillons en entretiens.

Ressources:

1. Akbari, M., Nikijoo, I., Khodapanah, B., Foroudi, P., & Padash, H. (2025). Forty Years of Microfinance Research and Its Impact on Consumers: A Review and Research Agenda Using the ADO-TCM Framework. International Journal of Consumer Studies, 49(4), e70101.
2. Behavioural Insights Team. How to Run Simple Behavioural Insight Projects. 2022. https://www.bi.team/wp-content/uploads/2022/11/BIT-Handbook-How-to-run-simple-BI-projects.pdf.
3. Blanc, J. (2014). Microfinance, Debt and Over-Indebtedness: Juggling with Money, Isabelle Guérin, Solène Morvant-Roux et Magdalena Villarreal (dir.). Editions Routledge, Londres, Royaume-Uni, 2014, 316 pages. Revue internationale de l'économie sociale: recma, (334), 122-124.
4. Bloomberg. "Big Money Backs Tiny Loans That Lead to Debt, Despair and Even Suicide." Bloomberg.com, May 3, 2022. https://www.bloomberg.com/graphics/2022-microfinance-banks-profit-off-developing-world/.
5. Clemens, Michael A., and Gabriel Demombynes. "When Does Rigorous Impact Evaluation Make a Difference? The Case of the Millennium Villages." Journal of Development Effectiveness 3, no. 3 (2011): 305–339. https://doi.org/10.1080/19439342.2011.587017.
6. Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín. "Technology and Child Development: Evidence from the One Laptop per Child Program." American Economic Journal: Applied Economics 9, no. 3 (2017): 295–320. https://doi.org/10.1257/app.20150385.
7. Evaluation Hub. "Run Evaluations." https://www.bitevaluationhub.com/run-evaluations.
8. John, B. (2024, November 14). Challenges and limitations of microfinance in achieving large-scale poverty reduction and job creation [Working paper].
9. J-PAL. "Design and Iterate the Implementation Strategy." https://www.povertyactionlab.org/resource/design-and-iterate-implementation-strategy.
10. J-PAL. "Ethical Conduct of Randomized Evaluations." https://www.povertyactionlab.org/resource/ethical-conduct-randomized-evaluations.
11. J-PAL. "Impact Evaluation Methods Table." https://www.povertyactionlab.org/sites/default/files/research-resources/impact-evaluation-methods-table.pdf.
12. J-PAL. "Power Calculations Exercise." https://www.povertyactionlab.org/sites/default/files/Exercise-PowerCalcs_0.pdf.
13. J-PAL. "Questionnaire Piloting." https://www.povertyactionlab.org/resource/questionnaire-piloting.
14. J-PAL. "Data Security Procedures for Researchers." https://www.povertyactionlab.org/resource/data-security-procedures-researchers.
15. J-PAL and IPA. Implementing Impact Evaluations: Case Study. 2023. https://poverty-action.org/sites/default/files/2023-03/Case-Study-Implementing-Impact-Evaluations.pdf.
16. Karlan, Dean, and Jacob Appel. More Than Good Intentions: Improving the Ways the World's Poor Borrow, Save, Farm, Learn, and Stay Healthy. Penguin, 2011.
17. NYU Office of Research. "IRB Decision Tree." https://www.nyu.edu/content/dam/nyu/research/documents/IRB/IRBDecisionTree.pdf.

18. Shelly, Sarah, et al. “Improving Communication with Participants in Behavioural Trials.” 2023.

19. UNICEF. An Evaluation of the PlayPump® Water System as an Appropriate Technology for Water, Sanitation and Hygiene Programmes. 2007. http://www-tc.pbs.org/frontlineworld/stories/southernafrica904/flash/pdf/unicef_pp_report.pdf.

20. UNICEF. Ethical Considerations When Applying Behavioural Science to Programmes with Children. Innocenti, 2021. https://www.unicef.org/innocenti/media/5186/file/UNICEF-Ethical-Considerations-Behavioural-Science-Children-2021.pdf.

21. UNICEF. “UNICEF Procedure for Ethical Standards in Research, Evaluation, Data Collection and Analysis.” https://www.unicef.org/evaluation/documents/unicef-procedure-ethical-standards-research-evaluation-data-collection-and-analysis.

22. UK What Works Evaluation Hub. “Pilot Impact Studies.” https://evaluationhub.eif.org.uk/pilot-impact-studies/.

23. University of California Santa Barbara Library. “Data Evaluation Checklist.” https://www.library.ucsb.edu/sites/default/files/attachments/data-curation/resources/DataEvaluationChecklist.pdf.

24. World Health Organization. Monitoring the Building Blocks of Health Systems: A Handbook of Indicators and Their Measurement Strategies. 2010. https://iris.who.int/bitstream/handle/10665/44708/9789241502320_eng.pdf.

25. Australian Institute of Family Studies. Process Evaluation. 2025. https://aifs.gov.au/sites/default/files/2025-03/2502%20EES%20process%20evaluation_1.pdf.

26. Tableau. “What Is Data Cleaning?” https://www.tableau.com/learn/articles/what-is-data-cleaning.

27. BetterEvaluation. “Data Cleaning.” https://www.betterevaluation.org/methods-approaches/methods/data-cleaning.