Evaluer pour améliorer : les principes à suivre pour mener des évaluations constructives et utiles

Diane Berthelette

1. Définitions :

Reprenons la même définition de l'évaluation que celle utilisée précédemment par François Daniellou : « porter un jugement de valeur sur une intervention ou sur n'importe laquelle de ses composantes dans le but d'aider à la prise de décision » (Contandriopoulos et coll., 2000).

Qu'entend-on par intervention ? Je précise que cette présentation concerne tout type d'intervention et étant donné qu'il y a eu peu d'évaluations d'interventions qui ont fait l'objet de publication, les réflexions viennent du domaine de l'éducation, de la psychologie, de la criminologie et un peu de la santé au travail. Donc, une intervention est « un ensemble de moyens qui sont organisés dans un contexte spécifique à un moment donné pour produire des biens ou des services dans le but de modifier une situation problématique » (Contandriopoulos et coll., 2000). Ce sont, par exemple et par ordre de complexité en termes de service ou d'activité, des outils, des méthodes, des programmes ou des politiques. Le problème à l'origine de l'intervention fait aussi partie de la définition de l'intervention.

Quelles sont les composantes de l'intervention ? C'est la base à partir de laquelle on compose l'évaluation. On compose tout d'abord la description de l'intervention, même si elle n'est pas parfaite, si elle est modélisée. On doit regarder les dimensions suivantes :

- Les objectifs qui sont poursuivis.
- Les sous-objectifs qu'il peut y avoir.
- Le processus, qui correspond à l'ensemble des activités et des services qui sont dispensés.
- La structure, ce sont les ressources humaines, matérielles, financières, symboliques. Pour illustrer les ressources symboliques, il peut y avoir des formations syndicales qui sont toujours dispensées dans le même lieu parce que ce lieu est associé au milieu syndical. On ne verrait pas certains syndicats aller s'installer dans un hôtel 5 étoiles par exemple, il y aurait un problème de dissonance.
- Les effets qui sont attendus ou inattendus, positifs ou négatifs. Lorsque ce sont des effets attendus, on parle de résultats. Lorsqu'on mesure les effets attendus et inattendus, on parle d'impact.
- Le contexte qui est important parce qu'une intervention peut prendre une certaine couleur dans certains milieux et une autre couleur ailleurs. Elle peut être efficace dans un contexte organisationnel et ne pas l'être dans un autre.

– Le problème ou le besoin identifié par les concepteurs de l'intervention. Les types d'évaluation prennent en compte les relations qui existent entre ces composantes quand on fait de la recherche évaluative.

Qu'entend-on par évaluation normative ? Reprenons la définition de Contandriopoulos que j'ai modifiée. C'est « une activité qui consiste à porter un jugement de valeur sur une intervention ». De quelle façon procède-t-on ? On compare, à l'aide de critères, les caractéristiques des ressources mises en œuvre et leur organisation (la structure), des services ou des biens produits (le processus) et des résultats, et on compare ces caractéristiques à des normes.

Qu'est ce qu'on entend par critères ? Les critères peuvent concerner par exemple :

– L'écart par rapport à l'intervention qui était prévue.
– L'étendue de la couverture qui est la proportion d'unités (individus, groupes, etc.) qui ont véritablement reçu les services. On peut alors comparer le nombre de groupes ou d'individus couverts par rapport à ce qui était souhaité à un moment donné de l'intervention.
– La qualité qui est toujours extrêmement difficile à définir parce qu'elle est souvent subjective et multidimensionnelle.
– L'intensité, par exemple le nombre de services ou le nombre d'heures dispensées.

En évaluation, surtout en recherche évaluative, pour expliquer la variation des effets observés, il faut tenter de définir des indicateurs de « dose » d'exposition à l'intervention. Il faut essayer de mettre en relation les caractéristiques de l'intervention ou du service avec les effets. Lorsque l'exposition à l'intervention varie (selon les entreprises, ou dans une même entreprise, selon les ergonomes en charge de l'intervention par exemple) l'intervention doit être conceptualisée de manière à mesurer la variation existante.

2. Les buts officiels :

– Aider à la planification et à l'élaboration d'une intervention : Il s'agit d'un but stratégique. Par exemple des projets pilotes mis sur pied avant de généraliser l'implantation de l'intervention dans l'ensemble de la population cible. À noter que lors d'une évaluation, on postule qu'il y a eu une intervention, ce qui n'est pas toujours vrai dans la réalité, d'où l'importance d'aller vérifier si il y a eu effectivement intervention. La démarche d'évaluation peut être entreprise avant l'intervention mais l'évaluation sera complétée une fois l'intervention implantée ou une fois l'ensemble de la collecte de données terminée sur les services qui ont été dispensés.
– Fournir de l'information pour améliorer, modifier ou en général gérer l'intervention : c'est un but formatif.
– Identifier les effets de l'intervention : c'est un but sommatif. Dans la majorité des domaines, on voit des analyses des effets qui se contentent d'aller voir quels sont les effets et si les effets attendus ont été atteints. Ce qui est privilégié dans le domaine de la santé, ce sont des approches expérimentales qui posent des problèmes et qui font le plus l'objet de publications. François Daniellou a parfaitement raison de dire qu'il y a des revues qui refusent d'autres types d'approches bien que ceci commence à changer.
– Contribuer à l'avancement des connaissances : dans un but qui est fondamental.

3. L'historique :

- De 1800 à 1900, l'âge du réformisme : l'évaluation a débuté de façon très réduite dans un contexte de révolution industrielle et de réforme de lois et de politique. Cela a débuté par des commissions royales d'enquête.

  La standardisation a débuté par la mise en place de systèmes d'inspection. Joseph Rice est celui qui est identifié à l'approche expérimentale : des groupes d'enfants, exposés si on peut dire à des méthodes pédagogiques différentes, étaient comparés.

- De 1900 à 1930, l'âge de l'efficience et des tests : par le management scientifique ou la gestion scientifique. Des tests objectifs sont développés pour des comparaisons car en évaluation, à plus forte raison en recherche évaluative, on a une approche comparative pour être en mesure de porter les jugements de valeurs. Des approches de tests et de mesures d'apprentissage ont beaucoup été développés.

- De 1930 à 1945, l'âge des objectifs : Tyler développe l'approche par objectifs et tente de vérifier l'atteinte de ceux-ci.

- De 1945 à 1960, l'âge de l'innocence : on perçoit que les ressources sont abondantes. Par conséquent on n'a pas besoin de justifier l'existence des programmes. Les approches citées précédemment sont raffinées et il n'y a pas vraiment de confrontation entre les différentes approches.

- De 1962 à 1972, l'âge de l'expansion : à cette époque, le sénateur Robert Kennedy avait fait passer une loi selon laquelle l'évaluation des programmes à l'attention des enfants handicapés dans les écoles devait être effectuée. L'hypothèse de Kennedy à l'époque était que la production de résultats d'évaluation informerait les parents de l'utilisation des ressources et donc leur donnerait un pouvoir d'action. Ceci s'est étendu à diverses formes de programmes aux Etats-Unis, mais surtout dans le domaine de la recherche sociale. Les différentes approches coexistent encore.

- Désillusion croissante : lorsque des évaluations étaient effectuées, elles parvenaient rarement à démontrer que les interventions atteignaient leurs objectifs, ce qui amenait à penser qu'elles étaient inefficaces. Les décideurs étaient très souvent insatisfaits des évaluations. Il fallait cesser de se limiter à décrire les interventions et commencer à porter des jugements.

  A propos des méthodes, des buts très différents étaient poursuivis. Un écart entre les programmes prescrits et les programmes réels a été pris en compte, ce que les ergonomes connaissent très bien mais qui était ignoré dans les autres domaines. La multiplicité d'effets (attendus et inattendus) a aussi été prise en compte.

- En 1973, l'âge de la professionnalisation : on pensait que le fait de diffuser de l'information auprès de tous les groupes d'intérêt concernés, par exemple les décideurs politiques, les gestionnaires, les intervenants sur le terrain, les gens qui bénéficiaient des services, ceux qui n'en bénéficiaient pas, donnait un pouvoir équivalent à tout le monde selon la théorie de la démocratie. On s'est rendu compte qu'il y avait des gens qui étaient mieux placés que d'autres pour utiliser cette information. Par conséquent lorsque l'évaluateur doit choisir son mandant, il doit déterminer à qui il souhaite fournir de l'information et qui doit l'aider à clarifier la question d'évaluation. C'est aussi le début de la prise en compte des méthodes qualitatives et de l'existence d'intérêts divergents entre les groupes.

Les gouvernements sont les principaux clients et en général ils le sont toujours. Le marché est entre les mains des acheteurs donc ce sont eux qui décident des questions. Un besoin de bases scientifiques apparaît et l'American Evaluation Association est fondée. Elle est composée à 40% d'universitaires et à 21% d'organismes gouvernementaux. Les bureaucraties ont besoin de légitimer leurs activités et l'évaluation est valorisée en raison de sa valeur scientifique. Les méthodes non scientifiques vont devenir marginalisées. L'accent est mis sur des devis expérimentaux et plus l'évaluation est objective, plus elle est utile et plus elle menace l'autorité.

- En 1980, les années Reagan : il y a une diminution des budgets consacrés aux programmes sociaux et moins d'opportunités pour l'évaluation. Les grandes démocraties ont développé leurs propres bureaux d'évaluation en interne. Des questions d'éthique qui apparaissent : De qui les évaluateurs relèvent-ils ? A quel point leur travail doit-il être public ? Les évaluations négatives doivent-elles être cachées ou discutées sur la place publique ? Quels sont les contrôles scientifiques qui doivent être mis en avant ?

Selon Shadish, il existe trois grands courants en ce moment, sans beaucoup de tiraillements entre eux :

- Le groupe de Scriven et Campbell met l'accent sur les méthodes essentiellement utiles. Il est important d'avoir des connaissances valides.
- L'accent est mis beaucoup plus sur le pragmatisme avec Weiss, Wholey et Stake. Il faut générer des alternatives utiles.
- Il faut intégrer les éléments du passé avec Cronbrack et Patton, qui adoptent plutôt une approche qualitative et historique.

4. L'évaluation normative :

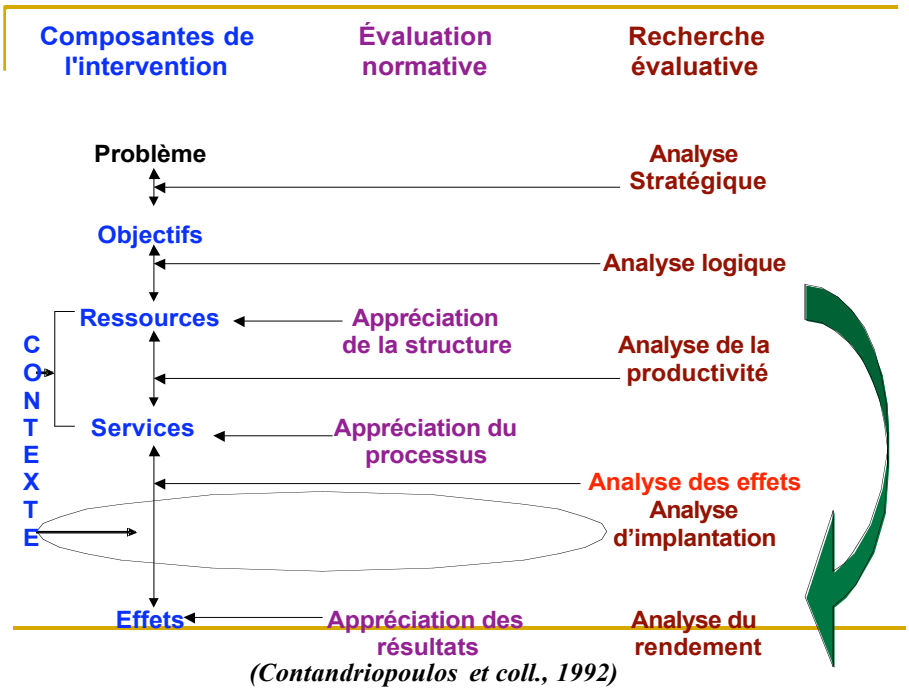
L'appréciation normative pose les principales questions suivantes :

- Est-ce que les ressources sont adéquates pour atteindre les résultats escomptés ? Adéquat étant toujours un terme à définir, dépendamment de la perspective.
- Est-ce que les services sont adéquats pour atteindre les résultats escomptés ?
- Est-ce que les résultats correspondent aux résultats attendus ?

5. La recherche évaluative :

Dans ce cas, ce n'est pas un jugement de valeur en appliquant des critères et des normes qui est apporté. D'ailleurs, une question que doit toujours se poser l'évaluateur concerne l'origine des critères et des normes. Souvent, il s'agit de critères et de normes qui ne sont pas produits par des recherches évaluatives, alors que pour être utiles ils doivent provenir de recherches qui ont permis d'établir une relation entre les composantes des interventions.

Il s'agit de « l'utilisation de méthodes scientifiques pour investiguer le lien entre les composantes de l'intervention » (Contandriopoulos et coll., 1992).



(Contandriopoulos et coll., 1992)

A gauche, il y a les composantes de l'intervention citées précédemment. A droite, il y a l'évaluation normative. Donc, le processus et les résultats peuvent être évalués. A l'extrême droite, il y a la recherche évaluative qui porte sur les liens.

- L'analyse stratégique : l'intervention permet-elle de répondre au besoin ?
- L'analyse logique : est-ce que pour atteindre les objectifs, les meilleurs moyens ont été mis en œuvre ?
- L'analyse de la productivité : c'est la relation entre les ressources et les services produits.
- L'analyse des effets : elle porte sur la relation entre les services et les effets.
- L'analyse d'implantation : il s'agit d'expliquer la variation des effets d'une intervention. Celle-ci peut être liée à la variation des caractéristiques des services dispensés ou encore du contexte dans lequel ils ont été offerts. Imaginons plusieurs ergonomes intervenant autour du même genre de problématique. Les interventions des ergonomes peuvent être comparées par rapport aux effets qui doivent être produits. Ceci demande une modélisation, une réflexion et une description en profondeur des façons de faire, des interventions, tout en tenant compte du contexte.
- L'analyse de rendement : elle porte sur la relation entre les ressources impliquées dans l'intervention et ses effets. Lors d'une analyse économique, il y a des effets tangibles et intangibles. Ces derniers sont plus difficiles à mesurer et il est plus ardu de leur attribuer une valeur monétaire.
- Il y a aussi plusieurs groupes : si on voulait évaluer des politiques de compensation de lésions professionnelles, il faudrait au moins se dire au départ qu'il y a différents groupes d'intérêt. Il y a l'Etat, les gestionnaires, les travailleurs, il peut y avoir leur famille et tous les groupes qui sont concernés par les effets directs et indirects de l'intervention que l'on veut évaluer. Il est important de recueillir des données auprès de ces groupes. Un choix peut être effectué, il n'est pas obligatoire de les faire tous. Il faut néanmoins se poser cette question pour effectuer consciemment ces choix.

Pourquoi mener des évaluations normatives ? Plusieurs étapes sont nécessaires pour opérationnaliser une intervention. En fait, il y a eu des demandes croissantes d'imputabilité au niveau du public et de l'Etat, et il était très important de recueillir des données et de faire des rapports continus sur la performance des fonctions critiques. L'accent était mis surtout sur les entrants et les effets qui étaient produits.

Il y a 3 perspectives :

- Celle de l'évaluateur qui veut souvent évaluer, interpréter les résultats d'une évaluation d'impact.
- Il y a l'imputabilité : la perspective selon laquelle on doit rendre des comptes parce que les ressources sont rares.
- Celle du gestionnaire qui veut corriger des situations qui sont problématiques.

L'évaluation du processus :

C'est une appréciation ponctuelle qui porte sur les services qui sont dispensés.

La surveillance du processus :

C'est peut-être ce qui est préférable pour des interventions ergonomiques étant donnée la variation puis les modifications qui peuvent arriver au cours de l'intervention. Par des journaux de bord par exemple, il est possible de recueillir des données de façon continue sur l'intervention au fur et à mesure qu'elle se développe. Des questions comme « Atteint-elle ses groupes cibles? » ou « Est-ce que le processus est cohérent avec les spécifications de ses concepteurs ? » peuvent être posées.

Les ergonomes sont les concepteurs et ils mettent en œuvre les interventions donc ils peuvent se poser des questions sur leur cohérence interne, « est-ce que le contexte leur a fait dévier leur route » par exemple. Ceci implique un jugement de valeur donc des critères et des normes sont nécessaires. Il faut aussi établir des critères opérationnels, ce qui est difficile.

La définition de normes :

Elles peuvent être établies selon différentes perspectives, soit grâce à l'expérience antérieure, des interventions comparables, des jugements de gestionnaires ou d'experts, des normes d'ordres professionnels, des législations ou encore des résultats de recherche évaluative sur des interventions similaires, d'où l'importance, lors de la publication ou de la diffusion d'une information, de bien décrire ce qui a été fait pour que les autres puissent se comparer et puissent apprendre de l'évaluation dont les résultats sont présentés. Malheureusement lors d'une intervention dans les organisations, les gestionnaires ou les groupes ont souvent peur de l'évaluation et retardent l'identification des critères et les précisent une fois la collecte de données terminée. Ils veulent obtenir les premiers résultats des observations avant de fixer les normes. Celles-ci ne sont pas très sévères de manière à ce qu'un jugement de valeur soit positif.

Un autre exemple est l'utilisation de service : à quel point les groupes reçoivent les services ? C'est critique et particulièrement lorsque la participation est volontaire et/ou lorsque les participants doivent apprendre de nouvelles procédures, changer leurs attitudes, surtout lorsqu'il y a un écart entre ce qu'ils faisaient et ce qu'on leur demande de faire.

Un autre exemple est l'accessibilité :

Une foule de questions peuvent se poser. Est-ce que les aménagements structurels et organisationnels facilitent la participation au programme ou à l'intervention ? Est-ce que le

lieu dans lequel j'interviens est adéquat pour que les participants utilisent les services ? Par exemple, est-ce que les formations sont dispensées au moment où les gens sont disponibles et à un endroit auxquels les gens peuvent avoir accès facilement ? Il y a aussi l'aspect culturel : est-ce qu'il y a des groupes qui utilisent des vocabulaires différents ? Est-ce que le langage utilisé est accessible aux différents groupes.

L'analyse des effets :

On ne peut pas dire qu'un programme est efficace s'il ne produit pas de changement positif et ce, même s'il répond aux questions suivantes : Répond-il aux besoins ? Possède-t-il un bon plan d'intervention ? Atteint-il sa population cible ? Dispense-t-on les services adéquats ?

Est-ce que le programme est évaluable ?

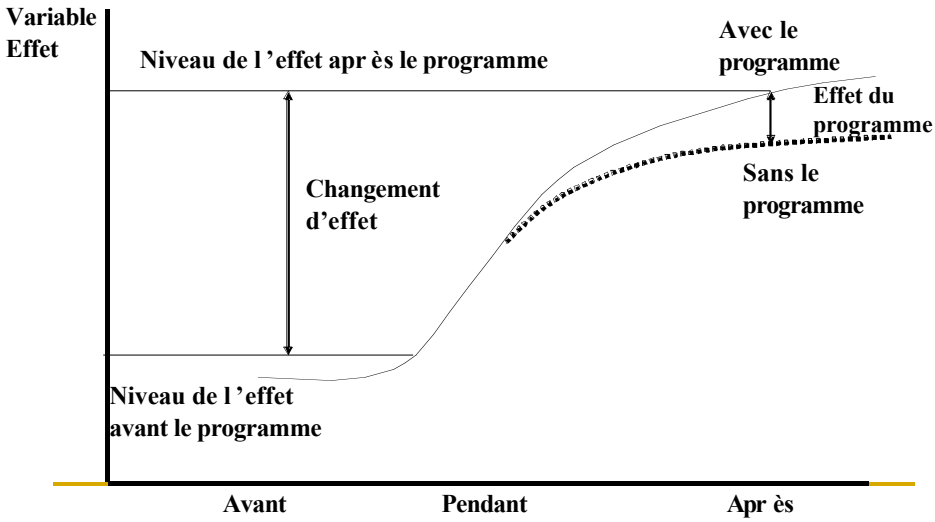
- Est-ce que les buts, les objectifs, les effets secondaires importants sont bien définis ?
- Est-ce que les besoins prioritaires d'information sont clairs ? Dans la perspective d'une évaluation de l'intervention d'une autre personne, il est nécessaire de cheminer avec les gens qui ont implanté l'intervention et qui veulent connaître ses effets ou son processus.
- Est-ce que les buts, les objectifs de l'intervention sont plausibles ? Il est inutile de faire une évaluation si dans la littérature une telle intervention est reliée à répétition à l'absence d'effets, à moins que les évaluations aient été mal faites, c'est-à-dire sans avoir véritablement modélisé l'intervention, décrit son implantation, ou sans avoir d'instruments valides et fiables.
- Possède-t-on des données pertinentes sur la performance de l'intervention ? Sur les services dispensés ? Sur l'étendue de la couverture ?
- Est-ce que les utilisateurs de l'évaluation visée s'entendent sur la façon dont les résultats de l'évaluation seront utilisés ? Pour ne pas être manipulés ou réduire au minimum le risque de la manipulation.

Lors de l'analyse des effets d'une intervention, les défis pour l'évaluateur sont les suivants :

- Mesurer les effets produits ;
- Déterminer à quel point les changements observés au niveau des effets sont attribuables au programme.

La définition des effets peut varier et nécessite de s'entendre sur le sujet. Imaginons que la question porte sur l'incidence des troubles musculo-squelettiques. Est-ce que l'évaluateur devra s'intéresser au niveau atteint après l'intervention ? Aux changements produits depuis le début de l'intervention ? Au gain net ? Généralement, l'accent est mis sur le gain net et tout sera mis en œuvre au niveau méthodologique pour tenter de le mesurer. Un graphique permet de conceptualiser la question :

Selon Rossi et al., 2004



La ligne en bas est continue « avant » « pendant » « après », il y n'a pas de temps zéro absolu. Au niveau conceptuel puisque l'on doit réduire, il y a un taux d'incidence de lésions musculo-squelettiques avant le programme. Si le programme fonctionne, la courbe atteint un certain niveau avec le programme, mais il faut au plan méthodologique essayer de voir ce que ça aurait été sans le programme. Bien sûr les gens qui travaillent avec des méthodes expérimentales diront que la seule façon de s'en assurer est d'assigner aléatoirement les participants à l'étude à un groupe exposé à l'intervention et à un groupe non exposé. C'est généralement impossible dans notre milieu.

De manière générale, les évaluateurs qui ont recours à des stratégies de recherche expérimentale omettent de décrire la théorie sous-jacente au programme c'est-à-dire les relations qui sont censées exister d'une part, entre les ressources investies dans le programme (structure) et les services produits (processus) et d'autre part, entre les services dispensés dans le cadre du programme et les effets qu'ils visent à produire (Bickman, 1987). Selon Bickman de telles omissions risquent d'entraver sérieusement l'interprétation des résultats des évaluations lorsque celles-ci semblent indiquer qu'un programme ne produit pas les effets attendus. L'absence d'effet peut être due à une théorie sous-jacente inadéquate et par conséquent à une intervention inefficace, à la présence d'écarts entre les caractéristiques du programme prescrit et de celui qui est implanté par les intervenants ou encore aux limites méthodologiques de l'évaluation. L'absence de données sur la théorie sous-jacente au programme et sur son implantation ne permet donc pas d'interpréter les résultats d'une étude qui indiquent qu'un programme ne produit pas d'effet. Ces problèmes réduisent considérablement la portée et l'utilité des résultats de telles recherches évaluatives.

Il faut mettre en avant d'autres méthodes permettant de documenter ce qui s'est passé en dehors de l'intervention. Le travail ne consiste pas uniquement à regarder en détail l'intervention mais à rechercher d'autres éléments contextuels qui pourraient expliquer la différence au niveau des effets.

Il peut être difficile d'identifier des effets pertinents, tout comme des mesures valides et fiables. On doit également tenter d'identifier les effets inattendus, positifs ou négatifs, qui peuvent être quand même assez importants.

Comment faire pour identifier les effets pertinents ?

En utilisant la théorie du programme : ceci consiste à préciser chacune des composantes de l'intervention et à décrire les mécanismes selon lesquels elle est censée produire ses effets intermédiaires, puis ultimes.

Ceci nécessite aussi de consulter les groupes d'intérêt, parce qu'ils ont des opinions à prendre en compte pour définir les effets intermédiaires. Ils peuvent expliquer le processus par lequel ils sont passés depuis le début de l'intervention jusqu'à la fin, ce qui enrichit notre théorie.

Les résultats d'études antérieures nous donnent aussi des pistes à ce sujet.

Un exemple de théorie du programme

Lors de l'évaluation d'un programme de formation syndicale, des observations ont été enregistrées et le verbatim des sessions de formation entièrement retapé pour comprendre la logique d'action des formateurs syndicaux. Pour résumer leurs théories, c'est un programme qui visait à faire en sorte que des délégués syndicaux mènent des actions syndicales pour améliorer la santé et la sécurité des salariés de leur entreprise. Leur théorie sous-jacente était qu'il fallait utiliser l'expérience et les connaissances des travailleurs sur les facteurs de risque d'accident de travail car les travailleurs sont inconscients des connaissances qu'ils possèdent. L'approche pédagogique devait être basée sur un échange, une confrontation des connaissances et des perceptions, inclure peu d'exposés, et privilégier les discussions et les travaux d'équipe avec des gens du milieu sous la forme d'une communication simple et directe. Ceci était censé permettre aux délégués d'attribuer entièrement les accidents aux employeurs uniquement, aucunement aux travailleurs. Selon cette théorie, cette dernière condition était essentielle pour que les délégués soient motivés à s'investir davantage dans l'action syndicale en matière de santé et de sécurité du travail.

Des mesures spécifiques :

Si les instruments de mesure des composantes de l'intervention à l'étude ne sont pas disponibles, on doit les développer, puis les valider. Il faut que le temps et les ressources soient suffisants pour ce faire.

La gravité du problème avant exposition peut également être mesurée, de même que l'exposition à d'autres phénomènes susceptibles de produire le même effet afin d'éviter les biais et donc de documenter ce qui entoure l'intervention, et de distinguer le groupe qui a vraiment bénéficié de l'intervention du groupe ciblé qui n'en a peut-être pas bénéficié.

Il ne faut pas oublier qu'il y a des programmes complexes donc il peut y avoir différentes théories pour le même phénomène. Il est important de mettre à jour les théories, les expliciter, les confronter. Il faut prendre pour acquis et non pas pour cas rare une variation de l'implantation. Une même personne peut varier ses approches parce qu'elle apprend de ce qui s'est passé précédemment. Elle peut aussi la varier en fonction du contexte, et entre des groupes d'individus, il peut y avoir également une variation. Il est très fréquent lors de l'évaluation d'un programme que l'intervention évolue donc il faut prendre cette évolution en compte.

Par exemple, un programme concernait le déplacement sécuritaire des bénéficiaires pour la prévention des maux de dos. Il y a quatre types d'activité : la formation du personnel

soignant, la prévention faite par les formateurs, la formation d'agents de suivi, puis les activités de suivi des agents. Tout ceci est censé contribuer à la réduction de l'incidence des maux de dos d'origine professionnelle. Donc chaque type d'intervention doit être documenté.

Conclusion

Pour optimiser l'utilisation des résultats, autrement dit faire en sorte que les résultats soient utiles aux groupes ciblés, il faut impliquer les groupes d'intérêt dans le choix des questions d'évaluation. Dans la mesure du possible, on doit leur présenter la liste des questions d'évaluation, et expliquer la portée et les limites des évaluations.

Il faut informer périodiquement les groupes d'intérêt de l'avancement des travaux. Une approche est de développer les méthodes avec les gens du terrain et non pas de les leur imposer.

Il faut enfin optimiser la rigueur de l'évaluation, c'est-à-dire indépendamment du contexte faire en sorte que les méthodes soient les plus valides possibles pour assurer des résultats qui permettent de voir si les relations entre les composantes de l'intervention existent vraiment mais aussi permettent une généralisation, en notant qu'une certaine généralisation peut être possible dans certains contextes différents.