Revue de psychoéducation

Ce programme d'intervention produit-il vraiment des données probantes?
Does this intervention program really produce valid data?

Frank Vitaro, Mélissa Gauthier-Samuel, Camille Livernoche Leduc, Isabelle Ugnat-Laurin et François Bowen

Volume 48, numéro 2, 2019

URI : https://id.erudit.org/iderudit/1066149ar
DOI : https://doi.org/10.7202/1066149ar

Aller au sommaire du numéro

Éditeur(s)

Revue de Psychoéducation

ISSN

1713-1782 (imprimé)
2371-6053 (numérique)

Découvrir la revue

Citer cet article

Vitaro, F., Gauthier-Samuel, M., Livernoche Leduc, C., Ugnat-Laurin, I. & Bowen, F. (2019). Ce programme d'intervention produit-il vraiment des données probantes? *Revue de psychoéducation*, *48*(2), 397–424. https://doi.org/10.7202/1066149ar

Résumé de l'article

La production et l'utilisation de données probantes sont au coeur des activités scientifiques et cliniques en psychoéducation, comme dans la plupart des disciplines en sciences humaines et de l'éducation. Or, qu'entend-t-on par 'données probantes' et plus encore, quels critères permettent d'avancer qu'un programme d'intervention préventive ou curative produit des données probantes? Cet article propose une grille d'analyse composée de critères pertinents à cette fin. Ces critères sont regroupés en cinq sections : les critères relatifs à la conception du programme ; les critères relatifs à la mise en oeuvre du programme ; les critères relatifs au choix du devis d'évaluation ; les critères relatifs au contrôle des sources d'invalidité interne ; et les critères relatifs à l'évaluation des effets. Ensemble, ces critères sont essentiels pour déterminer jusqu'à quel point les données obtenues à l'issue d'un processus d'évaluation d'un programme sont de nature probante. Une sixième section propose des critères relatifs à la diffusion du programme ; cette section est toutefois optionnelle pour l'obtention de données probantes.

Revue de psychoéducation
Volume 48, numéro 2, 2019,397-424

Ce programme d'intervention produit-il vraiment des données probantes?

Does this intervention program really produce valid data?

F. Vitaro[1]
M. Gauthier-Samuel[1]
C. Livernoche Leduc[1]
I. Ugnat-Laurin[1]
F. Bowen[2]

[1] École de psychoéducation, Université de Montréal
[2] Département de psychopédagogie et d'andragogie, Université de Montréal

Correspondance :

Frank Vitaro
École de psychoéducation, Université de Montréal, C.P. 6128. Succ. Centre-ville, Montréal QC H3C 3J7

frank.vitaro@umontreal.ca

Résumé

La production et l'utilisation de données probantes sont au cœur des activités scientifiques et cliniques en psychoéducation, comme dans la plupart des disciplines en sciences humaines et de l'éducation. Or, qu'entend-t-on par 'données probantes' et plus encore, quels critères permettent d'avancer qu'un programme d'intervention préventive ou curative produit des données probantes? Cet article propose une grille d'analyse composée de critères pertinents à cette fin. Ces critères sont regroupés en cinq sections : les critères relatifs à la conception du programme ; les critères relatifs à la mise en œuvre du programme ; les critères relatifs au choix du devis d'évaluation ; les critères relatifs au contrôle des sources d'invalidité interne ; et les critères relatifs à l'évaluation des effets. Ensemble, ces critères sont essentiels pour déterminer jusqu'à quel point les données obtenues à l'issue d'un processus d'évaluation d'un programme sont de nature probante. Une sixième section propose des critères relatifs à la diffusion du programme ; cette section est toutefois optionnelle pour l'obtention de données probantes.

Mots-clés : Programme, intervention, prévention, évaluation, données probantes

Abstract

Producing and using evidence-based data is an essential part of scientific and clinical work in Psychoeducation as well as in other disciplines of human and educational sciences. However, there is often ambiguity in what constitutes 'evidence-based' data and, importantly, it is not always clear whether such data effectively support a prevention or an intervention program. This paper proposes a set of criteria with the goal of clarifying these issues. These criteria are organized into five sections: those that refer to (1) the program planning phase; (2) the program implementation phase; (3) the evaluation design; (4) the internal validity of the evaluation process; and (5) the program effects. In all, these criteria should be viewed as necessary guideposts in determining whether data can

be designated as evidence-based. A sixth (optional) section presents additional criteria to determine whether the program is ready for large scale dissemination.

Keywords: Program, intervention, prevention, evaluation, evidence-based data

Un objectif central de la recherche en psychoéducation est de développer des interventions préventives ou curatives fondées sur des modèles théoriques et de les soumettre à une démarche d'évaluation rigoureuse visant à produire des données probantes. Une version pragmatique de cet objectif consiste à repérer des pratiques cliniques novatrices et de les soumettre à une démarche d'évaluation rigoureuse afin de démontrer qu'elles génèrent des données probantes et, le cas échéant, en dégager les fondements théoriques. Les chercheurs et les intervenants œuvrant en psychoéducation ont ensuite la responsabilité d'identifier et de diffuser les interventions qui produisent des données probantes et de les placer au cœur de leur pratique. Cette responsabilité partagée constitue un objectif qui reçoit un large appui de la part des chercheurs et des intervenants. L'atteinte de cet objectif serait facilitée si les chercheurs et les intervenants disposaient de critères communs et explicites pour juger du caractère probant des données engendrées par une intervention préventive ou curative. Les auteurs de méta-analyses pourraient également bénéficier de tels critères pour trier et classifier les études empiriques sur la base du caractère probant de leurs résultats. L'objectif de cet article est de clarifier la notion de données probantes et de proposer une liste de critères pour déterminer jusqu'à quel point une intervention produit de telles données.

Au cours des 20 dernières années est apparu un mouvement en médecine, en éducation et en sciences humaines appliquées (c.-à-d. psychologie, criminologie, psychoéducation) visant à repérer et à utiliser de manière consciente, explicite et judicieuse les pratiques qui produisent les 'meilleures données probantes'. L'objectif principal d'un tel exercice était de sélectionner les interventions préventives et curatives qui répondent aux besoins des clientèles propres à chaque discipline et d'écarter les autres (Sackett, Rosemberg, Gray, Haynes et Richarson,1996; Saussez, 2016; Slavin et Chambers, 2016). Un effet corollaire de ce mouvement fut de pousser les chercheurs à vouloir appuyer les programmes dont ils font la promotion sur des données probantes et aux cliniciens d'exiger qu'il en soit ainsi. Il faut dire que la situation le réclamait. En effet, il y a 20 ans, à peine une poignée des nombreux programmes destinés à venir en aide à des enfants et des adolescents en difficulté avaient fait l'objet d'une évaluation et pouvaient donc prétendre à des données probantes (Kazdin, 2000). En outre, plusieurs d'entre eux, même bien évalués, produisaient souvent des résultats modestes, voire néfastes, ce qui a donné lieu à une intensification du mouvement de production de données probantes (Dodge, Dishion et Lansford, 2006; Vitaro et Gagnon, 2003). Des progrès importants ont été accomplis depuis ce temps, mais peut-on affirmer aujourd'hui que les programmes de prévention ou d'intervention en psychoéducation produisent des données probantes (Das et al., 2016; Vitaro et Tremblay, 2017)? Pour répondre à cette question, il faut d'abord clarifier ce que signifie 'données probantes'. De manière générale, il est convenu qu'une chose est dite probante lorsqu'elle permet « de prouver et de conclure » (Larousse, 2018). Dans le contexte d'un programme de prévention ou d'intervention, les données sont jugées probantes lorsqu'elles

témoignent avec certitude de la *capacité* et de la *responsabilité* du programme à atteindre ses objectifs suivant un enchainement d'événements crédibles aux plans théorique et pratique. En somme, on peut 'prouver et conclure' qu'un programme repose sur des données probantes lorsque a) les changements/différences à la suite de l'application du programme sont importants, bénéfiques et valides, b) le programme en est la cause hors de tout doute et c) les mécanismes sous-jacents à ces changements/différences sont compatibles avec un modèle théorique reconnu (Kazdin, 2000). Le caractère probant des données est donc tributaire de l'ampleur et de la pertinence des effets produits par le programme. Il est également tributaire de la qualité du devis et des instruments utilisés pour faire échec aux variables confondantes susceptibles de porter ombrage au lien causal entre le programme et les effets obtenus. Finalement, il est tributaire du modèle logique et du modèle de changement, explicites ou non, sur lesquels le programme repose. L'objectif de cet article est d'avancer une série de critères opérationnels permettant de repérer les pratiques qui produisent les meilleures données probantes (synonyme : pratiques basées sur les preuves; *evidence-based*).

Les critères proposés s'inscrivent dans une perspective positiviste et quantitative, mais n'excluent pas les approches méthodologiques dites mixtes (Chatterji, 2004). Il ne faut donc pas se surprendre si leur application donne lieu à un résultat gradué plutôt qu'à un jugement binaire par rapport au caractère probant des données issues d'un programme d'intervention. Plusieurs des critères proposés dans cet article sont inspirés de textes publiés par Flay et al. (2005), par Charlebois et ses collaborateurs (Charlebois, Vitaro, Normandeau, Lévesque et Rondeau, 2006) ou par la Cochrane Collaborative Foundation (Waters et al., 2006). D'autres font bon usage des connaissances cumulées dans le domaine de la recherche évaluative (voir Shadish, Cook et Campbell, 2002). Un tel exercice n'est pas nouveau. Des critères visant à apprécier le caractère probant des résultats de programmes d'intervention préventive ou curative reliés à divers problèmes ont été proposés dans le passé (par exemple, Farrington, Gottfredson, Sherman et Welsh, 2007; Wong et al., 2014). Quoique pertinents, ces critères nous semblent cependant incomplets. Par exemple, Farrington et al. (2002) proposent une série de critères en lien avec les programmes de prévention de la délinquance alors que Wong et al. (2014) font de même en ce qui a trait aux interventions auprès de personnes autistes. Dans les deux cas, les auteurs s'attardent à la qualité du devis d'évaluation, à la validité des instruments de mesure par rapport aux variables dépendantes, au pouvoir statistique lors des analyses et à l'existence d'études de reproduction. En aucun temps, toutefois, ils ne font mention de l'intégrité de la variable indépendante (c.-à-d. le programme d'intervention), de la durabilité des effets d'une intervention ou de la pertinence théorique des mécanismes sous-jacents à ces effets dans leur appréciation du caractère probant des résultats d'une intervention. Par ailleurs, les critères proposés varient selon les auteurs, même lorsque les programmes concernent le même problème (voir par exemple Pendergast, 2011, pour la délinquance). Il nous est donc apparu utile de proposer une grille qui se veut à la fois inclusive et exhaustive par rapport à toutes les étapes de production de données probantes, de la planification à la reproduction, et applicable à une variété de problèmes; la grille proposée est présentée à l'appendice A.

À l'instar de certains auteurs qui demeurent critiques vis-à-vis le mouvement des pratiques fondées sur la preuve, nous sommes conscients que la grille proposée est réductionniste à bien des égards (Couturier, Gagnon et Carrier, 2009). Nous croyons toutefois qu'elle est beaucoup moins réductionniste que les grilles existantes parce qu'elle intègre précisément les éléments de flexibilité et de variabilité qui sont absents des autres grilles fondées principalement, rappelons-le, sur le devis d'évaluation. La grille que nous proposons ne s'applique toutefois qu'aux études qui font appel à un devis d'évaluation reposant sur la comparaison de groupes de participants. D'autres critères s'appliquent pour des études qui reposent sur des devis à cas unique (voir Kratochwill et al., 2013). Même si elle se veut la plus complète possible, la grille proposée ici n'est pas définitive. Elle est appelée à être bonifiée à la lumière des avancées méthodologiques et conceptuelles en constante évolution. Sa vertu principale est de proposer des critères explicites dans un format, souhaitons-le, convivial.

Les critères proposés sont regroupés en six sections : les critères relatifs à la conception du programme ; les critères relatifs à la mise en œuvre du programme ; les critères relatifs au choix du devis d'évaluation des effets ; les critères relatifs au contrôle des sources d'invalidité interne ; les critères relatifs à l'évaluation des effets ; les critères relatifs à la généralisation et à la diffusion du programme. Les critères des cinq premières sections correspondent à ce qu'il est convenu d'appeler une épreuve *d'efficacité*, que tout programme devrait subir, qu'il soit issu d'une modélisation théorique ou de la pratique clinique. Les critères de ces cinq premières sections peuvent servir, en principe, à analyser n'importe quel programme qui prétend produire des données probantes. Les critères de la dernière section sont réservés aux programmes qui aspirent à une épreuve *d'effectivité*. Comparée à une épreuve d'efficacité, laquelle vise à déterminer si un programme est en mesure d'atteindre ses objectifs dans des conditions optimales, une épreuve d'effectivité vise à déterminer ce qu'un programme réussit vraiment à produire dans un contexte naturel. Les deux types d'épreuves s'avèrent complémentaires (pour plus de détails, voir Flay et al., 2005; Vitaro, 2003).

Une échelle de type Likert à trois unités est utilisée pour déterminer le degré d'atteinte de chaque critère : complètement, partiellement, pas du tout. Dans certains cas, il est important de considérer l'option 'non applicable' (N/A); nous revenons sur ce point à la fin du texte. Ce système donne lieu à une cote numérique pour chaque section. Les cotes ainsi générées peuvent ensuite être pondérées en fonction du nombre de critères applicables et donner lieu à une cote globale. Cette cote globale numérique peut, à son tour, se prêter à des valeurs catégorielles du type A, B, C… moyennant l'application de seuils de performance préétablis. Compte-tenu du caractère subjectif et arbitraire de cette dernière opération, nous laissons le lecteur décider de son bien-fondé. Afin de vérifier la sensibilité de la grille aux biais perceptuels, nous l'avons soumise à deux reprises à un test de fidélité inter-juges impliquant des étudiants gradués en psychoéducation. Un accord minimal de 80 % a été atteint aux deux occasions (100 % après discussion). Les critères reliés à chaque section sont décrits sommairement dans les prochains paragraphes.

A. Critères relatifs à la phase de conception du programme

Les critères proposés dans cette section se rapportent à la phase de conception du programme. Ils visent spécifiquement à évaluer les assises théoriques et empiriques du programme, ainsi que les stratégies anticipées pour atteindre les objectifs visés. Même si ces critères ne sont pas reliés directement à la démarche d'évaluation et aux résultats qui en découleront, leur importance n'est pas à négliger puisqu'ils touchent aux conditions préalables à la production de données probantes. Malgré leur importance stratégique, ils sont souvent ignorés ou implicites, dans l'appréciation du caractère probant des données issues d'un programme.

Plusieurs des critères proposés dans cette section font écho aux modèles étiologiques proposés dans le champ de la psychopathologie développementale pour décrire et comprendre comment les problèmes d'adaptation se développent (Cicchetti et Cohen, 1995). Ces modèles sont basés sur des études longitudinales de nature prédictive-explicative et comprennent les facteurs de risque/bénéfiques et les facteurs de protection/vulnérabilité dont l'enchainement temporel permet d'expliquer le développement d'un problème sous la forme de chaînes développementales (voir Vitaro et Tremblay, 2017, pour un exemple de modèle développemental relié aux problèmes de comportement). Ils servent d'assise à ce que certains auteurs appellent le Modèle logique et le Modèle de changement d'un programme (Chen, 2005; Clark et Anderson, 2004). Le modèle logique sert à identifier les cibles d'une intervention (c.-à-d. les facteurs de risque/bénéfiques et les facteurs de protection/vulnérabilité). Il sert aussi à anticiper les changements proximaux et intermédiaires requis pour atteindre l'objectif terminal visé par l'intervention. Le modèle de changement, quant à lui, fait référence aux composantes d'un programme et aux stratégies d'intervention utilisées pour leur mise en place optimale. Le modèle de changement touche aussi aux conditions de mise en œuvre des diverses composantes du programme (Tougas et Tourigny, 2013), notamment les étapes d'apprentissages/changements chez les sujets ciblés par l'intervention (Kirkpatrick, 1998).

Le *critère #1* exige que les cibles du programme correspondent à des facteurs de risque/de promotion ou à des facteurs de vulnérabilité/de protection dont l'importance est reconnue au sein des chaînes développementales tentant de rendre compte de l'étiologie du problème d'adaptation à prévenir ou à remédier. Les notions de risque et de promotion renvoient à des facteurs présentant un haut potentiel d'influence causale alors que les facteurs de protection et de vulnérabilité correspondent à des facteurs modérateurs. Les deux catégories de facteurs sont considérées essentielles puisqu'elles correspondent aux deux objectifs d'une intervention préventive ou curative, soit a) d'éliminer les sources du problème en les remplaçant par leur contrepartie bénéfique ou b) de mettre en place/éliminer les facteurs susceptibles, respectivement, d'en atténuer ou d'en exacerber l'impact (voir Vitaro et Tremblay, 2017, pour une description détaillée des notions de risque/promotion ou de vulnérabilité/protection). Évidemment, l'application de ce critère implique que les chaînes développementales soient connues, ce qui est le cas sur un plan nomothétique pour une grande variété de problèmes d'adaptation. Par exemple, Dodge et ses collaborateurs (2008) ont réalisé une étude qui illustre bien la notion de chaîne développementale reliée à la délinquance. Une analyse

fonctionnelle peut cependant s'avérer nécessaire pour confirmer la pertinence de ces facteurs au plan idiographique.

Conformément au premier volet du Modèle logique, les cibles directes d'une intervention correspondent aux objectifs proximaux d'un programme, alors que les cibles à plus long terme auxquelles le programme aspire correspondent aux objectifs intermédiaires ou distaux. Peu importe leur nature, tous les objectifs doivent être formulés clairement, complètement et de manière opérationnelle (*critère #2*). Il faut par ailleurs que les cibles et les objectifs proposés correspondent aux composantes du programme. Vice-versa, les composantes du programme doivent être pertinentes par rapport aux cibles sélectionnées et aux objectifs visés (*critère #3*). Par exemple, il serait illusoire de croire qu'il est possible d'améliorer les pratiques disciplinaires des parents sans qu'une partie du programme leur soit spécifiquement consacrée. En revanche, une composante parentale qui n'a pas de cible ou d'objectif particulier n'a pas lieu d'être.

Une fois les composantes du programme identifiées, encore faut-il que le choix des stratégies pour assurer leur déploiement optimal et leur impact maximal découle d'une démarche réfléchie et rigoureuse. À titre illustratif, quelle est la meilleure stratégie pour aider des parents au chapitre de leurs pratiques disciplinaires et de leurs habiletés communicationnelles, tout en assurant leur collaboration et leur satisfaction? Les réponses à de telles questions sont au coeur du Modèle de changement et doivent prendre appui sur des sources fondées empiriquement, théoriquement ou cliniquement (*critère #4*). Un fondement empirique fait référence aux études pertinentes ayant produit les meilleures tailles d'effet à date et ayant subi le plus grand nombre de reproductions; à cet égard, lorsque disponibles, les méta-analyses constituent un excellent outil (voir Fortin, Lévesque et Vitaro, 2007). À défaut de bases empiriques, les stratégies de changement peuvent prendre appui sur des principes théoriques bien établis ou s'inspirer d'une pratique reconnue. Idéalement, le promoteur du programme proposera une version améliorée des stratégies inspirées des meilleures pratiques disponibles.

Afin d'asseoir davantage la pertinence des cibles, des objectifs et des stratégies retenus, il importe d'anticiper quelle cascade d'événements (c.-à-d. mécanismes) est susceptible d'être mise en branle par chaque composante du programme, conformément aux modèles développementaux connus. Cette opération renvoie au second volet du Modèle logique associé à un programme (*critère #5*). À titre illustratif, un programme peut cibler la capacité de supervision des parents (objectif proximal) dans le but de diminuer la tendance chez leur adolescent à s'affilier à des pairs déviants (objectif intermédiaire) et, ultimement, prévenir l'apparition de comportements délinquants (objectif distal). Dans cet exemple, l'affiliation à des pairs déviants sert à expliquer comment l'amélioration de la capacité de supervision chez les parents permet, à terme, de prévenir l'adoption de comportements délinquants chez leur jeune; on parle alors de variable médiatrice ou intermédiaire en référence aux pairs déviants (voir Hayes, 2018). Par ailleurs, il serait naïf de croire qu'un programme peut avoir un impact équivalent auprès de tous les participants. Il faut donc prévoir dès sa phase de conception quelles variables liées au contexte d'intervention ou aux caractéristiques des participants sont susceptibles d'influencer, à la hausse ou à la baisse, l'efficacité du programme

ainsi que son déploiement (*critère #6*). Ces caractéristiques contextuelles ou personnelles peuvent jouer deux rôles. Premièrement, elles permettent d'identifier des possibles modérateurs par rapport aux effets du programme. Par exemple, une intervention qui serait bénéfique pour les femmes, mais pas pour les hommes pourrait globalement être déclarée inefficace lorsque l'ensemble des participants est considéré. Deuxièmement, les caractéristiques contextuelles ou personnelles peuvent aider à expliquer la variabilité, non souhaitable mais parfois inévitable, au niveau du déploiement du programme (i.e. de sa mise en œuvre).

Finalement, il est peu probable que les résultats d'un programme se maintiennent ou parviennent à activer les médiateurs pertinents sans une provision de stratégies en ce sens au terme du programme. Ceci est particulièrement important lorsqu'il apparaît évident que le programme sera incapable de mettre en branle tous les mécanismes médiateurs requis (*critère #7*). Ces stratégies de maintien ou d'entrainement d'un effet domino peuvent faire appel aux ressources naturelles du milieu ou prendre la forme de sessions de relance (i.e. *boosters*).

En somme, les critères qui composent cette première section ne garantissent pas le succès d'un programme d'intervention ou de prévention. Ils réduisent toutefois les risques d'échec en optimisant la planification de son contenu et de sa mise en œuvre.

B. Critères relatifs à la mise en œuvre d'un programme

Cette section présente les critères relatifs à la mise en œuvre d'un programme afin d'assurer son intégrité. La mise en œuvre d'un programme consiste à favoriser et à documenter la concordance entre, d'une part, les contenus et les stratégies de changement prévus lors de la phase de planification d'un programme et, d'autre part, les composantes et les stratégies de changement effectivement déployées pendant son déroulement. Sans un tel exercice, il n'est pas possible de porter un jugement sur le caractère probant des données issues d'une intervention, ne contrôlant pas ou ne sachant pas ce qui s'est vraiment déroulé durant sa mise en œuvre. De façon plus précise, la mise en œuvre comporte deux volets : un volet qui touche au contenu mis en place par l'intervenant (c.-à-d. le contenu livré, aussi connu sous le terme 'adhésion') et un volet qui touche au contenu auquel les participants auront été exposés ou qu'ils se seront approprié (c.-à-d. le contenu reçu, aussi connu sous le terme 'exposition'), l'un n'étant pas nécessairement l'image-miroir de l'autre. Pour chaque volet, la concordance et les écarts par rapport au contenu prévu sont notés. Cette concordance ou ces écarts peuvent survenir à deux niveaux : au niveau des principes théoriques sur lesquels le programme repose ou au niveau des opérations pratiques déployées sur le terrain (pour plus de détails voir Gearing El-Bassel, Ghesquiere, Baldwin, Gillies et Ngeow, 2011; Dane et Schneider, 1998). Ainsi, la mise en oeuvre d'un programme correspond à un effort délibéré pour assurer et circonscrire la validité de la variable indépendante, c.à-d. le programme. La mise en œuvre inclut aussi la prise en compte des conditions dans lesquelles le programme est déployé (Vitaro, 2003). Ces conditions peuvent être améliorées par la mise en place d'ententes ou de moyens promotionnels. Les critères qui suivent se rapportent aux éléments liés à la mise en œuvre d'un programme et aux conditions susceptibles de l'optimiser.

Selon le premier critère de cette seconde section, le mode de recrutement et le mode de sélection des participants doivent être documentés (*critère #1*). La recherche de participants débute nécessairement par l'établissement d'un échantillon manifestant le problème à remédier (pour les programmes d'intervention) ou exposé aux facteurs de risque à éliminer et à remplacer par leur contrepartie positive (pour les programmes de prévention) (Charlebois et al., 2006). Sans une description de la population dont proviennent ces participants et sans une description de leur mode de recrutement (c.-à-d. probabiliste ou non probabiliste), comment bien apprécier les résultats qui découleront du processus d'évaluation, les reproduire au besoin et en connaître le degré possible de généralisation?

Les individus ou les groupes qui répondent aux critères de sélection sont ensuite invités à participer au programme et ceux manifestant un intérêt sont retenus. Cette opération donne lieu au taux de participation qui, en soi, constitue un premier indicateur de succès ou d'échec. Il constitue aussi notre second critère (*critère #2*), car un programme qui réussit à recruter une proportion élevée des participants ciblés verra sa validité externe et son pouvoir statistique améliorés par rapport à un autre programme peu performant à ce chapitre. Des efforts de recrutement sont parfois requis pour obtenir des taux élevés de participation (et de rétention), surtout s'il s'agit d'un programme de prévention pour lequel il n'y a eu aucune demande d'aide explicite. Évidemment, cela doit se faire dans le respect des règles éthiques, en évitant tout risque de stigmatisation. Diverses stratégies ont été proposées en ce sens (Normand, Charlebois et Vitaro, 2000).

Par ailleurs, il importe de bien décrire les critères de sélection et d'exclusion ainsi que les caractéristiques des participants, notamment le sexe, l'âge, le statut socio-économique et le lieu de résidence (*critère #3*). Ces données sociodémographiques, personnelles et cliniques servent à déterminer l'étendue et le profil de la population à laquelle les résultats sont généralisables (Durlak et DuPre, 2008). Elles permettent également une comparaison avec les études utilisant des programmes similaires. Finalement, elles peuvent, au besoin, servir de facteurs potentiellement modérateurs.

Tel que mentionné précédemment, l'évaluation de la mise en œuvre d'un programme d'intervention implique une adhésion optimale des intervenants au protocole du programme; sinon le contenu livré ne correspondra pas au contenu prévu. Cette recherche de conformité et d'intégrité peut être facilitée par l'accessibilité à une documentation écrite, c'est-à-dire, un manuel qui décrit les diverses composantes d'un programme et les stratégies pour les mettre en place (Gearing et al., 2011) (*critère #4*). Elle peut être facilitée également par une formation préalable et une supervision continue des intervenants (*critère #5*). La mise en place de ces deux éléments permet par ailleurs une standardisation des composantes et des stratégies du programme et une comparaison avec des programmes similaires. Elle facilitera aussi une éventuelle reproduction du programme.

Il serait vain d'optimiser le contenu livré par les intervenants si les participants le boudent. Autrement dit, on doit se préoccuper autant du contenu reçu que du contenu livré. Le contenu reçu renvoie au degré d'exposition et d'implication des participants au programme d'intervention ou de prévention. L'exposition

et l'implication se veulent optimales dans le contexte d'une épreuve d'efficacité. Plusieurs stratégies peuvent aider en ce sens (voir Normand et al., 2003). Il est de la responsabilité du promoteur du programme de les mettre en place afin d'éviter une faible exposition ou une piètre implication des participants, ce qui menacerait la validité interne et la validité externe de l'étude (*critère #6*). Il en va également de la responsabilité du promoteur de minimiser et, le cas échéant, d'évaluer les effets de débordement afin de maintenir une différenciation claire entre la condition d'intervention et la condition de contrôle (ou la condition pré-intervention). La notion de 'débordement' renvoie aux activités complémentaires auxquelles les participants prennent part et qui pourraient affecter les objectifs visés par le programme (Vitaro et Gagnon, 2003) (*critère #7*). Cette notion concerne autant les participants de la condition de contrôle que ceux de la condition d'intervention. En effet, il est peu probable que les participants dans une condition de contrôle s'abstiennent de toute intervention pendant la durée d'une étude. Ceci est d'autant plus vrai si les participants dans la condition de contrôle sont en contact avec ceux de la condition expérimentale ou si le programme a acquis une grande notoriété avant même d'être évalué.

Le promoteur d'un programme ne peut toutefois pas tout contrôler, même dans le contexte d'une étude d'efficacité. Il devient alors utile, voire nécessaire, de connaître les circonstances qui, en cours de déploiement, ont pu moduler à la hausse ou à la baisse la mise en œuvre d'un programme, tant au chapitre du contenu livré qu'à celui du contenu reçu (*critère #8*). Cette information permet de compléter l'examen des caractéristiques contextuelles ou personnelles susceptibles également d'affecter à la hausse ou à la baisse les effets d'un programme tel que noté à la section précédente.

Le contenu livré et le contenu reçu peuvent être évalués de manière qualitative ou quantitative, selon les contextes. Dans tous les cas, une telle évaluation s'accompagne d'instruments de mesure fiables, comme c'est le cas pour l'évaluation des effets (Gearing et al., 2011;Whitley, 2002) (*critère #9*). En termes pratiques, cela signifie que le journal de bord traditionnel rempli par l'intervenant devrait être assorti d'une évaluation externe indépendante, idéalement aveugle par rapport aux objectifs et aux étapes de réalisation du programme afin de minimiser les biais possibles dans l'évaluation de la mise en œuvre (Reid, 1998) (*critère #10*). Les études qui se contentent de faire appel uniquement aux intervenants pour juger de la qualité de la mise en œuvre ne trouvent souvent aucun lien entre les résultats de la mise en œuvre et les effets du programme (Dane et Schneider, 1998). Si aucune source indépendante n'est disponible, un compromis acceptable consiste à croiser les données obtenues auprès des intervenants par différents moyens : journal de bord, entrevues de groupe et individuelles, questionnaires, etc. McGraw et al. (1994) sont d'avis qu'une conformité de l'ordre de 80% est nécessaire et suffisante pour conclure que la mise en œuvre est réussie. Ce critère libéral laisse place à une certaine flexibilité afin d'être mieux adapté à la réalité du milieu et aux besoins des participants (Charlebois et al., 2006) (*critère #11*).

En somme, une mise en œuvre optimale correspond à l'absence d'écart entre le contenu prévu et le contenu livré/reçu, tant sur le plan des concepts théoriques que sur le plan des opérations concrètes. Cette absence d'écart assure

que le programme, une fois livré, est fidèle et intègre par rapport au programme initialement planifié. En pratique toutefois, il y a souvent des écarts d'un participant à l'autre, ouvrant la porte à la possibilité d'analyses visant à expliquer une telle variabilité ou à la relier aux résultats obtenus au terme du programme. Les résultats de ces analyses de type 'dosage-effet' doivent être considérés avec prudence en raison de leur faible validité interne. Si des contenus partiellement variables sont requis entre les participants afin de tenir compte de leurs besoins particuliers, il est préférable qu'ils soient programmés et non laissés à la discrétion des participants ou des intervenants. Le programme *Early Risers* constitue un bon exemple d'un programme qui intègre des composantes particulières ajustées aux besoins individuels des participants, en plus des composantes communes à tous les participants (August, Egan, Realmuto, et Hektner, 2003). Castro Gonzalez, Barrera, & Martinez (2004) offrent d'autres exemples de stratégies visant à harmoniser le besoin d'intégrité dans la mise en œuvre d'un programme et la nécessaire adaptation de son contenu et de son mode de livraison selon le contexte culturel des participants.

C. Critères relatifs au devis d'évaluation

Cette section concerne le choix du devis d'évaluation. Un devis d'évaluation correspond aux stratégies méthodologiques mises en place pour contrer ou cerner l'effet de divers facteurs susceptibles d'invalider la relation causale entre le programme d'intervention et les changements/différences au chapitre des variables dépendantes. Ces facteurs sont souvent désignés comme des 'sources d'invalidité interne' ou des 'variables confondantes' (voir la section D du présent texte ainsi que Reid, 1998, ou Shadish et al., 2002). Sans le joug d'un devis d'évaluation performant, leur contrôle n'est pas possible, semant alors le doute par rapport à la responsabilité du programme face aux effets obtenus. Voilà pourquoi le devis d'évaluation constitue l'élément central, parfois unique, auquel les auteurs dans le passé se réfèrent pour juger du caractère probant des données issues d'un programme de prévention et d'intervention. Le choix d'un devis plus ou moins performant dépend de l'accessibilité aux éléments suivants : la possibilité a) de concevoir un groupe de contrôle, b) la possibilité de procéder à l'assignation aléatoire des participants dans les différents groupes et c) la possibilité d'effectuer des mesures répétées (incluant plusieurs prétests).

Ainsi, le premier élément à considérer est la possibilité d'avoir un groupe de contrôle. Le groupe de contrôle est constitué de participants présentant un degré d'équivalence raisonnable avec les participants qui prennent part au programme d'intervention ou de prévention (c.-à-d. les participants du groupe expérimental). Le groupe de contrôle témoigne des effets possibles des sources d'invalidité interne et sert de point de comparaison au groupe expérimental qui, en plus, est exposé au programme à évaluer. Il est donc primordial que le groupe expérimental et le groupe de contrôle soient semblables au prétest avant l'application du programme. Ceci ouvre la porte sur le second élément : l'assignation aléatoire des participants dans le groupe expérimental et le groupe de contrôle. L'assignation aléatoire (ou randomisation) des participants maximise en effet les chances que les groupes soient équivalents au point de départ puisque chaque participant a une probabilité égale d'être affecté au groupe expérimental ou au groupe de contrôle. Toutefois,

pour que les lois du hasard s'exercent correctement, il faut que les effectifs de départ soient suffisamment grands ou relativement homogènes (Mercier et Gagnon, 1998). À défaut d'une assignation aléatoire, il est essentiel de bien repérer les participants du groupe de contrôle afin qu'ils soient le plus semblables possible aux participants du groupe expérimental. Diverses techniques, telles l'appariement ou l'utilisation de scores de propension peuvent servir à cette fin (pour un exemple, voir Petitclerc, Gatti, Vitaro et Tremblay, 2013). Au besoin, un ajustement statistique est possible lors des analyses.

Le troisième élément à considérer est la possibilité de prendre des mesures répétées au prétest comme au post-test. Des mesures fréquentes et à intervalle régulier auprès des participants, en particulier au prétest, permettent d'estimer les effets possibles de diverses sources d'invalidité interne sur les tendances comportementales avant le début de l'intervention et d'en tenir compte dans les analyses. Ensemble, ces trois éléments (c.-à-d. la présence d'un groupe de contrôle, le mode d'assignation des participants aux différents groupes et les mesures répétées des comportements-cibles) déterminent la catégorie à laquelle appartient le devis de recherche : expérimental, quasi-expérimental ou pré-expérimental. Ces trois catégories de devis ne sont pas d'égale valeur pour contrer les sources d'invalidité interne (Reid, 1998).

Les devis expérimentaux

La combinaison du groupe de contrôle et de l'assignation aléatoire donne lieu à un *devis de type expérimental* et permet d'inférer avec confiance que les changements sont attribuables à l'intervention. Il s'agit du type de devis permettant de contrôler le plus grand nombre de facteurs d'invalidité interne avec le plus d'efficacité. L'assignation aléatoire n'est toutefois pas infaillible. Elle peut engendrer un groupe expérimental et un groupe contrôle qui ne sont pas équilibrés au regard des variables d'intérêt, particulièrement si le bassin de participants est petit ou hétérogène. Un prétest est alors souhaitable. Toutefois, même dans des circonstances non optimales, le risque inhérent à la méthode d'affectation aléatoire d'engendrer des groupes biaisés (c.-à-d. non équivalents et tendant vers les hypothèses de recherche) demeure inférieur au risque associé à toute autre méthode d'affectation basée sur des critères fixés par les chercheurs (Campbell et Stanley, 1963).

Les devis quasi-expérimentaux

La présence d'un groupe de contrôle, mais l'impossibilité de répartir aléatoirement les participants dans la condition expérimentale ou contrôle, donne lieu à un *devis quasi-expérimental.* Dans ce cas, un prétest est requis pour s'assurer de l'équivalence initiale des groupes. Tel que proposé précédemment, l'équivalence initiale des groupes est améliorée si a) un effort d'appariement, par pairage ou par scores de propension, a été réalisé au point de départ et si b) un effort est consenti pour recruter les participants du groupe de contrôle et ceux du groupe expérimental au sein de la même population d'origine.

Il existe différents devis de type quasi-expérimental. Tous requièrent un prétest mais certains ne requièrent pas de groupe contrôle (voir Mercier et Gagnon, 1998; Shadish et al., 2002). Par exemple, dans le cas du devis à séries temporelles à groupe unique, la prise de mesures répétées avant et après l'application du programme permet de suivre les effets possibles d'un programme tout en contrôlant/estimant la contribution possible de plusieurs facteurs d'invalidité interne. Le devis à régression discontinue mérite également une attention particulière dans ce contexte puisque l'intervention peut être offerte aux participants qui en ont le plus besoin, sans compromission importante de la validité interne (Shadish et al., 2002).

La combinaison de mesures répétées et d'un groupe de contrôle apparié au groupe expérimental permet même d'atteindre un niveau de protection vis-à-vis les facteurs d'invalidité interne équivalent à celui offert par les protocoles expérimentaux. Pour cette raison, nous ne faisons pas de distinction dans la grille entre un devis expérimental (c.-à-d. avec assignation aléatoire au groupe expérimental et au groupe de contrôle) et un devis quasi-expérimental à séries temporelles multiples (c.-à-d. avec mesures répétées et avec groupe de contrôle apparié). En revanche, nous faisons une distinction tranchée avec les devis de type pré-expérimental.

Les devis pré-expérimentaux

Les devis qui ne comprennent ni groupe de contrôle, ni assignation aléatoire, ni mesures répétées, sont des *devis de type pré-expérimental* (voir Mercier et Gagnon, 1998, pour une description détaillée des devis pré-expérimentaux). Ce type de devis est très vulnérable à l'influence des facteurs d'invalidité interne et n'offre, par conséquent, aucune garantie par rapport au lien de causalité entre l'intervention et les résultats. Il a sa place dans les études pilotes, mais il est difficile de le justifier dans le cadre d'une étude d'efficacité ou d'effectivité visant la production de données probantes. Ceci explique la faible valeur qui est accordée à un devis pré-expérimental dans la grille.

En résumé, la nature du devis constitue un élément central quand vient le temps d'apprécier si un programme produit des données probantes. Par conséquent, nous en avons fait une section à part. Toutefois, nous ne proposons pas d'identifier spécifiquement le type de devis en jeu, même si cela est souhaitable. Nous proposons plutôt d'identifier et de coter les trois éléments essentiels qui définissent tout devis suivant le rationnel décrit ci-haut : la présence d'un groupe de contrôle, l'assignation aléatoire des participants, la présence de mesures répétées, au prétest comme au post-test.

D. Critères relatifs au contrôle des sources d'invalidité interne

Même en présence d'un devis expérimental ou quasi-expérimental, il est impératif de continuer à exercer un jugement critique et à rester vigilant par rapport aux facteurs d'invalidité interne. En effet, même si un devis performant permet, en principe, un contrôle rigoureux des facteurs d'invalidité interne, un tel contrôle en pratique peut se heurter à certaines difficultés (p. ex. une perte différentielle des participants au sein du groupe expérimental et du groupe de contrôle pourrait

remettre en question l'équivalence initiale des groupes même si ceux-ci sont constitués initialement par assignation aléatoire). Il n'est donc pas possible de faire l'économie d'un examen attentif de chaque facteur d'invalidité interne même en présence d'un devis en principe rigoureux, car seule leur élimination effective permettra d'établir un lien de causalité entre le programme et les résultats. Ces facteurs d'invalidité interne peuvent être regroupés en sept catégories (Campbell et Stanley, 1963; Reid, 1998) : les facteurs historiques, la maturation, la sélection des participants, l'instrumentation, la réactivité à la mesure, la régression vers la moyenne et la défection des participants. Le contrôle effectif de chaque facteur d'invalidité interne renvoie à un critère différent (c.-à-d. *critères 1 à 7*)

Facteurs historiques. Les facteurs historiques renvoient à des événements et à des expériences personnelles susceptibles de survenir en même temps que le programme d'intervention ou de prévention et d'affecter les variables dépendantes (Fortin et Gagnon, 2016). Cette catégorie de facteurs n'est pas contrôlable avec un devis de type pré-expérimental. Toutefois, leur survenue différentielle au sein du groupe expérimental et du groupe contrôle pourrait compromettre aussi la validité interne d'un devis expérimental, d'où la nécessité d'une bonne évaluation de la mise en œuvre.

Maturation. La maturation se définit comme un ensemble de processus individuels, à la fois biologiques et psychologiques, perméables au passage du temps. Ces changements internes comprennent, par exemple, le vieillissement, la croissance, la fatigue, la faim et la motivation (Fortin et Gagnon, 2016; Reid, 1998). Un devis de type pré-expérimental ne permet pas de contrôler ou d'apprécier l'effet de la maturation, car cet effet est confondu avec l'effet du programme. Toutefois, même les devis quasi-expérimentaux sont sujets à un effet de maturation différentiel si les participants du groupe expérimental et du groupe contrôle ne sont pas équivalents au prétest ou s'ils ne proviennent pas de la même population de référence.

Sélection des participants. La sélection des participants fait référence à l'équivalence des groupes expérimental et contrôle en début d'intervention (Reid, 1998). Lorsque les groupes ne sont pas formés selon un processus garantissant leur équivalence, il y a présence possible d'un biais de sélection, lequel risque d'entraîner des déformations systématiques dans les résultats au post-test (Fortin et Gagnon, 2016). La méthode la plus robuste pour prévenir ce biais de sélection est l'assignation aléatoire, tel que discuté précédemment (Reid, 1998). Toutefois, même le hasard pourrait ne pas constituer des groupes équivalents lorsque les effectifs sont restreints, d'où la nécessité d'un prétest pour s'en assurer. Un prétest est d'autant plus nécessaire, voire incontournable, si les groupes ne sont pas répartis au hasard comme dans le cas des devis de type quasi-expérimental. Il est important toutefois de réaliser que même la présence d'un prétest ne garantit pas l'équivalence des groupes par rapport à des variables non mesurées. En ce sens, une auto-désignation des participants dans le groupe expérimental ou le groupe de contrôle est à proscrire, car elle alimente le biais de sélection des participants.

Fluctuation de l'instrument de mesure. La fluctuation de la mesure renvoie aux changements dans la calibration de l'instrument de mesure ou dans la manière dont l'évaluateur recueille ses données (Fortin et Gagnon, 2016; Reid, 1998). Que la mesure soit de nature qualitative ou quantitative, elle comporte toujours un certain degré d'erreur. Cette erreur peut être de nature systématique ou de nature aléatoire. Même dans le contexte d'une étude avec un devis expérimental ou quasi-expérimental, les résultats peuvent être invalidés si les instruments de mesure comportent un risque d'erreur systématique qui les fait fluctuer de manière différentielle entre le prétest et le post-test ou entre le groupe expérimental et le groupe de contrôle. Ce risque est élevé si, par exemple, les participants ou les évaluateurs connaissent leur groupe d'appartenance et se comportent de manière à confirmer ou à justifier leurs actions ou leurs attentes; on parle alors d'un effet différentiel de désirabilité sociale ou de biais d'attribution.

Réactivité à la mesure. Le seul fait de répéter une mesure peut influer sur le résultat (Reid, 1998). Ce phénomène peut s'expliquer par la familiarisation, par l'apprentissage ou par l'augmentation de l'intérêt ou de la motivation du participant à l'égard de l'instrument de mesure. Un effet inverse d'habituation, d'ennui ou de fatigue est également possible. L'opération même de mesure peut donc porter à conséquence. Par exemple, le simple fait de demander à un participant de noter le nombre de verres d'alcool consommés peut modifier sa consommation. La seule façon d'éviter un tel effet serait de se priver de mesures, ce qui n'est pas réaliste. Plusieurs techniques de contrôle sont cependant disponibles afin de contrer ou de neutraliser ces exemples de réactivité à la mesure (par exemple, en prévoyant une période d'habituation ou en contre-balançant l'ordre de passation des instruments de mesure).

Régression statistique. La régression statistique se manifeste principalement lorsque les participants sélectionnés pour une étude se situent aux extrémités d'une échelle de mesure. Il s'agit d'un phénomène de nature statistique (Fortin et Gagnon, 2016), où tout résultat extrême d'une distribution au prétest tend à régresser vers un résultat moyen dans les tests subséquents, indépendamment de l'intervention (Reid, 1998). Évidemment, un devis de type pré-expérimental n'est pas en mesure de contrer ou d'apprécier un tel effet de régression à la moyenne. Toutefois, certains devis de type quasi-expérimental et même expérimental ne sont pas à l'abri de cette source d'invalidité interne si les groupes ne sont pas parfaitement équivalents au prétest.

Défection des participants. La défection de participants entraîne une diminution de la validité externe d'une étude. Elle peut aussi remettre en question la validité interne d'une étude même dans le contexte d'un devis de type expérimental ou quasi-expérimental où on aurait pris bien soin de s'assurer de l'équivalence initiale des groupes de participants (Fortin et Gagnon, 2016). En effet, la défection peut être de nature différentielle, ce qui renvoie à deux scénarios possibles. Dans le premier scénario, il peut arriver que le nombre d'abandons diffère selon les conditions expérimentales. Dans le second scénario, il pourrait y avoir une différence marquée dans les caractéristiques des participants qui quittent le groupe expérimental et le groupe de contrôle, même si le nombre d'abandons dans les groupes est similaire (Reid, 1998). La meilleure façon de contrôler cette source

d'invalidité interne, comme toutes les autres sources d'invalidité interne d'ailleurs, consiste à empêcher sa survenue. Ceci n'est évidemment pas toujours possible. Dans le cas spécifique de la défection des participants, des stratégies de contrôle sont disponibles (voir Sedgwick, 2015; Ten Have et al., 2008). Par exemple, il est possible d'adopter une stratégie d'intention de traitement (*intent-to-treat*) à l'occasion des analyses. Dans ce cas, les participants qui ont quitté le programme sont quand même inclus dans les analyses statistiques, selon leur assignation initiale (Gupta, 2011; Ten Have et al., 2008). L'*intent-to-treat* permet ainsi de conserver les bénéfices associés à l'assignation aléatoire et de rendre un estimé juste des effets potentiels de l'intervention (Shadish et al., 2002). À l'*intent-to-treat* peut s'ajouter l'analyse par l'utilisation d'une variable instrumentale ou l'analyse *per protocol* (voir Ten Have et al., 2008). L'analyse *per protocol* permet de comparer les participants ayant reçu l'intervention avec ceux ne l'ayant pas reçu. Comme les participants peuvent se retirer d'une intervention à la suite de l'assignation, l'analyse *per protocol* peut toutefois être biaisée si l'attrition est différentielle au sein du groupe expérimental et du groupe de contrôle (Sedgwick, 2015). Les opérations relatives à la détection et à la remédiation d'un possible problème de défection sont représentées par les *critères 8, 9 et 10*.

En somme, après avoir assuré la validité d'un programme par une évaluation rigoureuse de sa mise en œuvre, il est nécessaire pour la production de données probantes de s'assurer de la validité du lien qui unit l'application du programme aux effets obtenus. Cet objectif est atteint lorsque chaque facteur d'invalidité interne est mis en échec ou contrôlé (méthodologiquement ou statistiquement). Autrement, leur influence risque d'être confondue avec les effets du programme. Il reste à s'assurer que ces effets soient eux-mêmes valides.

E. Critères relatifs aux effets du programme

Rappelons qu'un programme d'intervention préventive ou curative vise à prévenir ou mettre fin à un état ou une situation problématique et le-la remplacer par un état de bien-être ou une situation normalisée. Conformément au modèle logique d'un programme et aux modèles théoriques en psychopathologie développementale, l'atteinte de tels objectifs distaux implique souvent (mais pas toujours) l'atteinte d'objectifs proximaux et d'objectifs intermédiaires. Les objectifs proximaux découlent directement des effets présumés des composantes d'un programme, alors que les objectifs intermédiaires correspondent aux variables susceptibles de jouer un rôle médiateur (voir Rose, Holmbeck, Coakley et Franks, 2004). Par ailleurs, il faut garder à l'esprit qu'un programme n'aura pas les mêmes effets ou ne mobilisera les mêmes mécanismes médiateurs chez tous les participants ou dans tous les contextes sociaux, d'où la nécessité de prévoir aussi des variables susceptibles de modérer à la hausse ou à la baisse les effets d'un programme. Ces éléments recouvrent ceux présentés à la première section de ce texte; ici, il s'agit de prévoir des mesures pertinentes par rapport à ces divers éléments (*critère #1*). Par ailleurs, il ne sert à rien de prévoir des mesures pertinentes par rapport à tous ces éléments si celles-ci ne reposent pas sur des instruments possédant de bonnes propriétés psychométriques en termes de validité et de fidélité (Whitley, 2002) (*critère #2*). Simultanément, il est important de ne pas multiplier à outrance le nombre d'instruments de mesure afin de maintenir un équilibre entre une erreur

de type 1 (c.-à-d. rejeter l'hypothèse nulle de façon erronée) et une erreur de type 2 (c.-à-d. conserver l'hypothèse nulle de manière erronée). D'une part, il peut s'avérer nécessaire d'ajuster le seuil de rejet de l'hypothèse nulle (*critère #3*); d'autre part, il faut s'assurer d'une puissance statistique suffisante (*critère #4*). Cependant, il n'est pas permis d'ignorer une mesure une fois administrée, mais il est possible, voire souhaitable, de regrouper différentes mesures afin de créer des cotes composites ou des variables latentes (*critère #5*).

Il est par ailleurs important de planifier un suivi par rapport au maintien des résultats. Un intervalle de six mois apparaît comme une durée minimale permettant d'affirmer que les effets perdurent dans le temps, mais des intervalles variant entre un et trois mois sont également possibles. Pour les programmes de prévention qui visent des objectifs à long terme, il est souhaitable, voire nécessaire, de prévoir une mesure au moment où il est prévu que la variable distale s'actualise (Flay et al., 2005) (*critère #6*). Sans une mesure de la diplomation réelle, comment savoir, par exemple, si un programme de prévention du décrochage scolaire en première secondaire atteint son objectif distal.

Plusieurs problèmes d'adaptation partagent des facteurs de risque communs. Ainsi, il est possible qu'une intervention produise des résultats allant au-delà des objectifs proximaux, intermédiaires ou distaux visés. Ces effets, assimilables à des effets collatéraux, renvoient souvent à des effets bénéfiques dans d'autres domaines que ceux d'intérêt ou chez d'autres individus que ceux ciblés (*critère #7*). Par exemple, des programmes de prévention du décrochage scolaire pourraient aussi avoir un impact à la baisse sur la criminalité, produisant ainsi un rapport coût-bénéfice doublement avantageux. Malheureusement, les effets collatéraux sont rarement évalués. En outre, ils ne sont pas toujours bénéfiques. À titre d'exemple, une intervention visant à augmenter la supervision parentale pourrait entraîner une augmentation des conflits familiaux. Il est également possible qu'une prise en charge préventive d'enfants à risque entraîne un effet de marquage négatif ou un faux sentiment de sécurité chez les parents qui ne chercheront pas à utiliser les ressources d'aide disponibles dans leur milieu (voir McCord, 1992, pour un exemple percutant à cet effet).

Tel que proposé précédemment, il est nécessaire que les résultats relatifs à l'ensemble des mesures réalisées soient soumis à l'analyse. Ce critère de transparence est doublé d'un critère de performance, car il est attendu que la majorité de ces résultats se trouvent statistiquement significatifs (*critère #8*). La tradition et la convention imposent une valeur frontière de signification à 5%. Cette valeur correspond au seuil de tolérance du chercheur quant au risque d'obtenir des résultats dus au hasard. Lorsque le résultat du test statistique est inférieur à 5% ($p < 0,05$), le résultat est significatif, c'est-à-dire qu'il y a moins de 5% de chance que le lien entre le programme et les résultats soit attribuable au hasard. Les tests statistiques visent donc à écarter le hasard comme source d'explication des changements/différences au chapitre des variables dépendantes. Ils ne permettent pas de déterminer si le programme est responsable des résultats. C'est uniquement la rigueur avec laquelle le devis de recherche (voir section précédente) permet de contrer les facteurs d'invalidité interne qui permet de tirer cette seconde conclusion. En somme, le *critère #8* se limite à vérifier si les effets significatifs attendus sont

au rendez-vous, moyennant l'application des tests statistiques appropriés (selon la nature des variables en jeu et le modèle d'analyse projeté).

L'efficacité d'une intervention préventive ou curative ne devrait toutefois pas être jugée en fonction de ses seuls effets significatifs (ou non significatifs) au plan statistique, d'autant que le seuil traditionnel de .05 est remis régulièrement en question (voir Lakens et al., 2018). Elle devrait être également jugée en fonction de son impact clinique ou social, moyennant des indicateurs tels le fonctionnement social des participants ou la taille des effets de l'intervention (Cohen, 1977). La taille des effets d'une intervention permet de standardiser les effets à travers différentes études utilisant le même programme ou des programmes différents et donc de les comparer entre elles. Elle sert aussi d'ingrédient principal pour réaliser des méta-analyses (voir Fortin et al., 2007). Finalement, elle permet de juger de l'ampleur du changement ou de la différence qu'un programme est en mesure d'engendrer. Les balises numériques proposées par Cohen pour juger si les tailles d'effet sont modestes, moyennes ou élevées s'avèrent utiles en ce sens (Cohen, 1990; Sullivan et Feinn, 2012). Un tel jugement demeure néanmoins tributaire de certains autres paramètres, notamment l'importance des objectifs en jeu (Rosnow et Rosenthal, 2003). Il est donc de la responsabilité du chercheur de fournir les tailles d'effet de son programme par rapport aux différentes variables dépendantes (proximales, intermédiaires ou distales). Alternativement ou additivement, il est possible de fournir des indicateurs cliniques de nature diagnostique ou des indicateurs pratiques liés au fonctionnement social des participants (*critère #9*). Ces indicateurs constituent souvent la cible ultime d'un programme et, à terme, représentent les données possiblement les plus probantes par rapport auxquelles un programme peut être jugé.

Les prochaines lignes concernent l'application effective des critères qui précèdent et l'obtention de résultats favorables. Par exemple, le *critère #10* propose que le résultat de la démarche de suivi proposée au critère #6 ne se solde pas par un constat d'effritement des résultats. Dans le même ordre d'idées, l'examen souhaité des effets collatéraux au point #7 devrait se traduire par une absence d'effets iatrogènes ou pervers. Un effet iatrogène renvoie aux variables ciblées par l'intervention alors qu'un effet pervers renvoie aux effets collatéraux. Dépendamment de leur nombre et de leur importance, des effets iatrogènes ou pervers pourraient disqualifier un programme par ailleurs performant au chapitre des autres objectifs ciblés (*critère #11*). Par ailleurs, à quoi sert-il de planifier une série de variables potentiellement modératrices et médiatrices dans le Modèle logique et de prévoir des instruments de mesure appropriés, si les analyses qui s'y rapportent ne sont pas réalisées de manière formelle (*critères #12 et #13*) (voir Hayes, 2013, ainsi que Fairchild et McKinnon, 2009)? Une analyse des facteurs modérateurs permet de bien circonscrire la portée du caractère probant des résultats obtenus à l'issu d'une intervention. Sans une telle analyse, un programme pourrait être déclaré inefficace de façon erronée. Une analyse des mécanismes de médiation, quant à elle, augmente la crédibilité du processus de changement sous-jacent à une intervention et par conséquent la valeur probante des résultats qui en découlent. Elle permet également d'appuyer (ou de remettre en question) les éléments du modèle logique et des modèles théoriques qui leur sont sous-jacents (Vitaro, 2003). Sans une telle

analyse, il n'est pas possible d'expliquer les effets obtenus ou l'absence d'effets au terme d'une démarche d'évaluation, créant une situation de 'boîte noire' indésirable.

L'application unique d'un programme peut être vulnérable aux idiosyncrasies des participants ou aux conditions particulières de sa mise en œuvre. Par conséquent, la reproduction d'un programme constitue un atout nécessaire afin d'asseoir son caractère probant, en plus de sa fiabilité (*critère #14*). Si cette reproduction est menée par un groupe de chercheurs différents de ceux ayant mené l'étude originale, la valeur probante du programme s'en trouve rehaussée (Koole et Lakens, 2012) (*critère #15*). Finalement, la qualité de la tribune de diffusion des résultats d'un programme constitue un autre gage de la qualité de la démarche d'évaluation sur laquelle les résultats reposent et donc de leur valeur probante, en raison du processus d'évaluation par les pairs qui y est associé (*critère #16*). Les programmes mal fondés ou mal évalués sont rarement publiés.

F. Critères relatifs à la généralisation et à la diffusion du programme

Les critères précédents s'appliquent surtout dans le contexte d'une épreuve d'efficacité. Tel que décrit précédemment, cette épreuve d'efficacité consiste à évaluer les effets d'une intervention dans des conditions contrôlées et idéales qui assurent une mise en œuvre optimale et une validité interne maximale (Flay et al., 2005). Par opposition, une étude d'effectivité vise à tester les effets d'une intervention dans des conditions naturelles. Autrement, on ne connaîtra pas son efficacité réelle, seulement son efficacité potentielle. Par conséquent, après avoir subi une épreuve d'efficacité, un programme doit subir une épreuve d'effectivité avant de pouvoir conclure au caractère probant de ses effets en milieu naturel. Cette épreuve d'effectivité comporte des critères d'efficience, de généralisation et de diffusion, en plus des critères d'efficacité déjà discutés (voir Flay et al., 2005). La section suivante y est consacrée.

Contrairement à une épreuve d'efficacité qui est menée par des promoteurs motivés et des collaborateurs bien formés, une épreuve d'effectivité implique que le programme est implanté dans des conditions naturalistes par des intervenants au profil souvent diversifié. On ne devrait donc pas s'étonner que sa mise en œuvre soit variable. Toutefois, même une épreuve d'effectivité doit s'attendre à une mise en œuvre suffisante, c'est-à-dire qu'au moins deux tiers (cote 2) ou la moitié (cote 1) des indicateurs de sa mise en œuvre originale obtenus dans le cadre d'une étude d'efficacité soient maintenus. Un programme a beau être efficace dans des conditions idéales, il ne saurait produire de données probantes si les intervenants et les participants n'y adhèrent pas (*critère #1*). Il est donc de la responsabilité du promoteur de l'épreuve d'effectivité de faciliter minimalement la mise en œuvre d'un programme qui a démontré son efficacité en milieu contrôlé en mettant à la disposition des intervenants des moyens simples et peu coûteux comme un manuel ou une formation à distance (*critère #2*). Un tel exercice permet aussi de vérifier, et au besoin d'ajuster, l'adéquation entre la philosophie du programme et l'idéologie des intervenants. Il permet aussi de vérifier si les participants sont rejoints et s'ils adhèrent au programme sans un effort exceptionnel en ce sens. Malgré tout, l'implantation d'un programme a de bonnes chances de ne pas être homogène dans le contexte d'une étude d'effectivité. Par conséquent, le promoteur devrait en

tirer profit en examinant les caractéristiques des intervenants et des participants qui adhèrent le mieux au programme et qui sont associées à de meilleurs résultats (*critères #3 et #4*) (Charlebois, Vitaro, Normandeau, Brendgen, et Rondeau, 2004). Finalement, pour que les données d'un programme soient jugées probantes dans le cadre d'une épreuve d'effectivité, il importe de se demander si les effets demeurent suffisants compte tenu des possibles variations dans sa mise en œuvre. Par suffisant, nous entendons qu'au moins deux tiers (cote 2) ou la moitié (cote 1) des résultats significatifs ou des tailles d'effet obtenus dans le cadre d'études d'efficacité soient maintenus (*critère #5*).

Rappelons qu'une épreuve d'efficacité suffit pour qu'un programme puisse être jugé capable de produire des données probantes. Toutefois, une épreuve d'efficacité et une épreuve d'effectivité sont toutes deux nécessaires pour qu'un tel programme puisse aspirer à une diffusion à grande échelle. Pour cela, des critères additionnels sont requis, dont le critère d'efficience. En effet, un programme qui n'est pas efficient ne recevra probablement pas l'aval des décideurs (Mirelman et al., 2012). L'efficience d'un programme repose sur son rapport coûts/bénéfices : à efficacité égale, le programme le moins coûteux devrait être privilégié. Par ailleurs, il importe de démontrer qu'un programme rapporte des bénéfices monétaires ou humains supérieurs aux investissements qu'il commande s'il veut aspirer à une certaine pérennité et à un déploiement à grande échelle (*critère #6*). Pour éviter que les effets d'un programme dans un contexte d'épreuve d'effectivité ne soient trop dépendants des acteurs impliqués, il s'avère utile de procéder à une reproduction indépendante du programme en contexte naturel (*critère #7*). Les tailles d'effet ou les résultats significatifs dans ces études de reproduction devraient également être au rendez-vous, en plus d'une absence d'effets nuisibles (c.-à-d. iatrogènes) (*critère #8*). La capacité de généralisation du programme sera par ailleurs accrue si les études de reproduction impliquent des participants avec des caractéristiques variées ou représentatifs de leur population d'origine (*critère #9*). Par ailleurs, il importe de prévoir un monitorage du déploiement du programme à grande échelle ainsi que de ses effets afin de s'assurer, comme dans le cas d'une épreuve d'efficacité, que les résultats à ce double chapitre continuent à être suffisants (critère *#10*). Une mesure du degré d'intégration d'un programme dans les pratiques quotidiennes constitue, par ailleurs, un bon indicateur de sa pérennité (*critère #11*). Des stratégies susceptibles de favoriser le transfert et l'appropriation du programme et de son rationnel constituent un atout en ce sens (*critère #12*). Pour terminer, un programme doit être équitable s'il veut prétendre à une véritable diffusion à grande échelle. L'équité fait référence à sa capacité à rejoindre et à bénéficier à tous les participants potentiels (*critère #13*; Reinke, 1999).

Tel qu'indiqué précédemment, les critères qui composent cette section ne sont pas requis pour décréter qu'un programme a produit des données probantes. Ils sont toutefois nécessaires pour qu'un programme fondé sur des données probantes puisse prétendre à un déploiement à grande échelle (pour une illustration de déploiement à grande échelle, voir Christensen et Griffiths, 2007).

Conclusion

La psychoéducation se veut une discipline en constante évolution qui se situe à la croisée des sciences du développement et des sciences de l'intervention. Depuis toujours, les chercheurs et intervenants qui s'en réclament développent, évaluent et font la promotion de programmes de prévention et d'intervention pour des jeunes en difficulté. Au-delà de leurs fondements théoriques souvent implicites et de leurs promesses parfois hyperboliques, ces programmes doivent ultimement être jugés en fonction de la nature probante des résultats qu'ils génèrent. Un tel jugement doit reposer sur des critères explicites et partagés afin d'éviter la partialité et l'arbitraire. La crédibilité des divers programmes de prévention et d'intervention dont les chercheurs et intervenants font la promotion en dépend. Toutefois, l'application de ces critères doit demeurer flexible et s'ajuster en fonction du stade de développement d'un programme. Par exemple, on ne saurait tenir rigueur à un jeune programme prometteur de ne pas avoir encore été soumis à une reproduction indépendante; dans ce cas, le critère de reproduction serait soustrait de l'analyse de son caractère probant par l'utilisation de l'option 'N/A; ne s'applique pas'. De même, il ne faudrait pas juger trop sévèrement un programme qui permet de réduire un problème important, même de façon modeste, lorsqu'aucune alternative n'est disponible.

Ce texte avait pour objectif de proposer un certain nombre de critères opérationnels dans le but de clarifier et de quantifier le niveau de preuve que les données issues de programmes de prévention ou d'intervention offrent en regard de leurs promesses et ultimement de leurs fondements théoriques. Comme la psychoéducation, ces critères et la grille qui les accompagne sont appelés à évoluer.

Références

August, G. J., Egan, E. A., Realmuto, G. M. et Hektner, J. M. (2003). Four years of the early risers early-age-targeted preventive intervention: Effects on aggressive children's peer relations. *Behavior Therapy, 34* (4), 453-470.

Campbell, D. T. et Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand-McNally.

Castro, F. G., Barrera, M. Jr. et Martinez, C. R. Jr. (2004). The cultural adaptation of prevention interventions: Resolving tensions between fidelity and fit. *Prevention Science, 5*(1), 41 45.

Charlebois, P., Vitaro, F., Normandeau, S., Brendgen, M. et Rondeau, N. (2004). Trainers' behavior and participants' persistence in a longitudinal preventive intervention for disruptive boys. *Journal of Primary Prevention, 25*(3), 375-388.

Charlebois, P., Vitaro, F., Normandeau, S., Lévesque, J. et Rondeau, N. (2006). Illustration des principes et des procédures d'implantation et d'évaluation d'un programme de prévention de type ciblé. *Revue de psychoéducation, 35*(1), 65-94.

Chatterji, M. (2004). Evidence on "What Works": An argument for extended-term mixed-method (ETMM) evaluation designs. *Educational Researcher, 33*(9), 3-13.

Chen, H. T. (2005). *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*. Thousand Oaks, CA: Sage.

Christensen, H. et Griffiths, K. (2007). Reaching standards for dissemination: A case study. Dans K. A. Kuhn, J. R. Warren et T.-Y. Leong (dir.), *Medinfo*

2007 - Proceedings of the 12th World Congress on Health (Medical) Informatics (pp. 459-463). Amsterdam, Pays-Bas: IOS Press.

Cicchetti, D. et Cohen, D. J. (1995). Perspectives on developmental psychopathology. Dans D. Cicchetti et D. J. Cohen (dir.), *Developmental psychopathology: Volume 1. Theory and method*s (pp. 3-20). New York, NY: Wiley.

Clark, H. et Anderson, A. A. (2004). *Theories of change and logic models: Telling them apar*t. Communication à l'American Evaluation Association Conference.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York, NY: Academic Press.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312.

Couturier, Y., Gagnon, D. et Carrier, S. (2009). Management des conduites professionnelles par les résultats probants de la recherche. Une analyse critique. *Criminologie, 42*(1), 185-199.

Dane, A. V. et Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*(1), 23-45.

Das, J. K., Salam, R. A., Lassi, Z. S., Khan, M. N., Mahmood, W., Patel, V. et Bhutta, Z. A. (2016). Interventions for adolescent mental health: an overview of systematic reviews. *Journal of Adolescent Health, 5*9(4), S49-S60.

Dodge, K. A., Dishion, T. J. et Lansford, J. E. (2006). *Deviant peer influences in programs for youth: Problems and solutions.* New York, NY: Guilford Press.

Dodge, K. A., Greenberg, M. T., Malone, P. S. et CPPRG. (2008). Frequency distribution, 10,000 replications, no mediated effect, all continuous variables on-line. C*hild Development, 79*(6), 1907-1927.

Durlak, J. A. et DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3-4), 327-350.

Fairchild, A. J. et MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science, 10*(2), 87-99.

Farrington, D. P., Gottfredson, D. C., Sherman, L. W. et Welsh, B. C. (2002). The Maryland Scientific Methods Scale. Dans L. W. Sherman, D. P. Farrington, B. C. Welsh et D. L. MacKenzie (dir.), *Evidence-based crime prevention* (pp. 13-21). New York, NY: Routledge.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., . . . Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science, 6*(3), 151-175.

Fortin, F., Lévesque, J. et Vitaro, F. (2007). La méta-analyse au service de l'intervention préventive et curative. *Revue de psychoéducation, 36*(1), 167-194.

Fortin, M. F. et Gagnon, J. (2016). *Fondements et étapes du processus de recherche : méthodes quantitatives et qualitatives* (3e éd.). Montréal, QC: Chenelière Éducation.

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J. et Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review, 31*(1), 79-88.

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research, 2*(3), 109-112.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis.* New-York, NY: Guilford Press.

Hayes, J. (2018). *The theory and practice of change management*. London, United Kingdom: Palgrave.

Kazdin, A. E. (2000). *Psychotherapy for children and adolescents: Directions for research and practice*. New-York, NY: Oxford University Press.

Kirkpatrick, D. L. (1998). *Evaluating training programs: The four levels*. San Francisco, CA: Berrett-Koehler.

Koole, S. L. et Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science, 7*(6), 608-614.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. et Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*(1), 26-38.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., . . . Bradford, D. E. (2018). Justify your alpha. *Nature Human Behaviour, 2*(3), 168-173.

Le Petit Larousse illustré 2018. (2017). Paris, France: Larousse.

McCord, J. (1992). The Cambride-Somerville Study: A pioneering longitudinal-experimental study of delinquency prevention. Dans J. McCord et R. E. Tremblay (dir.), *Preventing antisocial behavior: Interventions from birth through adolescence* (pp. 196-206). New York, NY: Guilford Press.

McGraw, S. A., Stone, E. J., Osganian, S. K., Elder, J. P., Perry, C. L., Johnson, C. C., . . . Luepker, R. V. (1994). Design of process evaluation within the Child and Adolescent Trial for Cardiovascular Health (CATCH). *Health Education Quaterly, Suppl 2*, S5-26.

Mercier, P. et Gagnon, M. (1998). Les protocoles de recherches pré, quasi et expérimentaux. Dans S. Bouchard et C. Cyr (dir.), *Recherche psychosociale : pour harmoniser recherche et pratiques* (2e ed., pp. 77-135). Sainte-Foy, QC: Presses de l'Université du Québec.

Mirelman, A., Mentzakis, E., Kinter, E., Paolucci, F., Fordham, R., Ozawa, S., . . . Niessen, L. W. (2012). Decision-making criteria among national policymakers in five countries: A discrete choice experiment eliciting relative preferences for equity and efficiency. *Value in Health, 15*(3), 534-539.

Normand, C., Vitaro, F. et Charlebois, P. (2000). Comment améliorer la participation et réduire l'attrition des participants aux programmes de prévention. Dans F. Vitaro et C. Gagnon (dir.), *Prévention des problèmes d'adaptation chez les enfants et les adolescents. Tome 1 : Les problèmes internalisés* (Vol. 1, pp. 101-140). Sainte-Foy, QC: Presses de l'Université du Québec.

Petitclerc, A., Gatti, U., Vitaro, F. et Tremblay, R. E. (2013). Effects of juvenile court exposure on crime in young adulthood. *Journal of Child Psychology and Psychiatry, 54*(3), 291-297.

Prendergast, M. L. (2011). Issues in defining and applying evidence-based practices criteria for treatment of criminal-justice involved clients. *Journal of Psychoactive Drugs, 43*(Sept, suppl. 7), 10-18.

Reid, L. (1998). Les sources d'invalidité et de biais. Dans S. Bouchard et C. Cyr (dir.), *Recherche psychosociale. Pour harmoniser recherche et pratique* (2e éd., pp. 19-75). Sainte-Foy, QC: Presses de l'Université du Québec.

Reinke, W. A. (1999). A multi-dimensional program evaluation model: considerations of cost-effectiveness, equity, quality and sustainability. *The Canadian Journal of Program Evaluation, 14*(2), 145-160.

Rose, B. M., Holmbeck, G. N., Coakley, R. M. et Franks, E. A. (2004). Mediator and moderator effects in developmental and behavioral pediatric research. *Journal of Developmental and Behavioral Pediatrics, 25*(1), 58-67.

Rosnow, R. L. et Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Revue canadienne de psychologie expérimentale, 57*(3), 221-237.

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B. et Richardson, W. S. (1996). Evidence based medicine: what

it is and what it isn't. *British Medical Journal, 312*(7023), 71-72.

Saussez, F. (2016). *Les données probantes en éducation et l'arrimage recherche pratique : un regard critique*. Document web : Université de Sherbrooke.

Sedgwick, P. (2015). Intention-to-treat analysis versus per-protocol analysis of trial data. *British Medical Journal, 350*, h681.

Shadish, W. R., Cook, T. D. et Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.

Slavin, R. E. et Chambers, B. (2017). Evidence-based reform: Enhancing language and literacy in early childhood education. *Early Child Development and Care, 187*(3-4), 778-784.

Sullivan, G. M. et Feinn, R. (2012). Using effect size—or why the P value is not enough. J*ournal of graduate medical education, 4*(3), 279-282.

Ten Have, T. R., Normand, S. L. T., Marcus, S. M., Brown, C. H., Lavori, P. et Duan, N. H. (2008). Intent-to-treat vs. non-intent-to-treat analyses under treatment non-adherence in mental health randomized trials. *Psychiatric Annals, 38*(12), 772-783.

Tougas, A.-M. et Tourigny, M. (2012). L'étude des mécanismes de changement, une avenue de recherche prometteuse pour optimiser les programmes de traitement destinés aux jeunes en difficulté : enjeux conceptuels et méthodologiques. *The Canadian Journal of Program Evaluation, 27*(2), 61-86.

Vitaro, F. (2000). Évaluation des programmes de prévention : Principes et procédures. Dans F. Vitaro et C. Gagnon (dir.), *Prévention des problèmes d'adaptation chez les enfants et les adolescents. Tome I* (pp. 67-99). Sainte-Foy, QC : Presses de l'Université du Québec.

Vitaro, F. et Gagnon, C. (Eds.). (2003). *Prévention des problèmes d'adaptation chez les enfants et les adolescents. Tome II.* Sainte-Foy, QC : Presses de l'Université du Québec.

Vitaro, F. et Tremblay, R. E. (2017). Developmental targeted prevention of conduct disorder and their related consequences [Online]. Dans H. Pontell (dir.), *Oxford Research Encyclopedia of Criminology.* Oxford, United Kingdom: Oxford Press.

Waters, E., Doyle, J., Jackson, N., Howes, F., Brunton, G. et Oakley, A. (2006). Evaluating the effectiveness of public health interventions: The role and activities of the Cochrane Collaboration. *Journal of Epidemiology and Community Health, 60*(4), 285-289.

Whitley, B. (2002). *Principles of research in behavioral science* (2e éd.). Boston, MA: McGraw-Hill.

Wong, C., Odom, S. L., Hume, K. A., Cox, A. W., Fettig, A., Kucharczyk, S., . . . Schultz, T. R. (2015). Evidence-based practices for children, youth, and young adults with Autism Spectrum Disorder: A comprehensive review. *Journal of Autism and Developmental Disorders, 45*(7), 1951-1966.

Appendice A – Grille pour déterminer la valeur probante des données issues d'un programme d'intervention ou de prévention

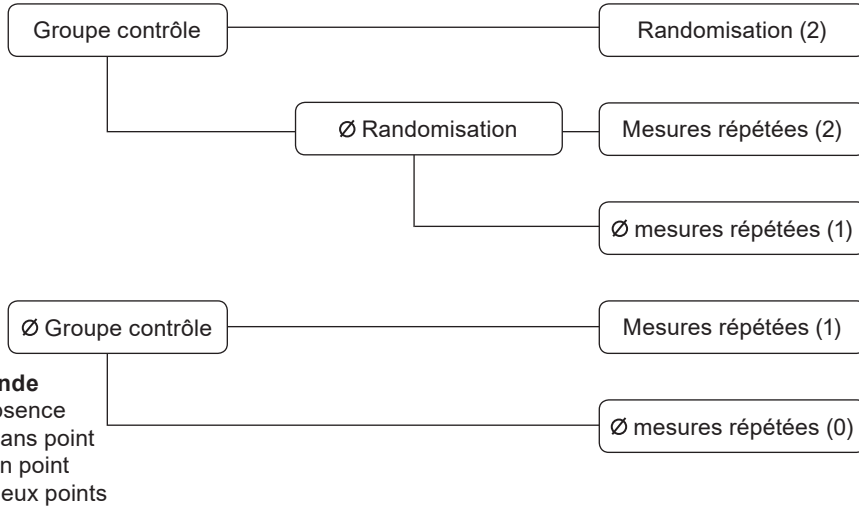
Grille pour l'analyse de programmes de prévention ou d'intervention		
Définition	Pointage	Nom du programme et auteurs :
Critère rempli complètement	2	Année et lieu de publication :
Critère rempli partiellement	1	
Critère non rempli	0	Total de l'évaluation (% moyen) :
Critère non applicable à l'étude	N/A	

A. Conception du programme					
Critères	2	1	0	N/A	Notes
1. Les cibles du programme correspondent à des facteurs de risque/promotion ou à des facteurs de vulnérabilité/protection dont la présomption de contribution causale a été établie.					
2. Les objectifs proximaux et distaux du programme sont formulés clairement.					
3. Il y a adéquation entre les objectifs et les composantes du programme.					
4. Les stratégies de changement visant la mise en œuvre de chaque composante sont appuyées au plan empirique, théorique ou clinique.					
5. Les auteurs décrivent les mécanismes médiateurs par l'intermédiaire desquels le programme devrait atteindre ses objectifs distaux.					
6. Les auteurs précisent et justifient les variables susceptibles d'exercer un effet modérateur sur les effets du programme.					
7. Les auteurs ont prévu des stratégies pour favoriser le maintien et la généralisation des effets du programme.					
Nombre de critères retenus :	*Total de points obtenus sur max possible de:*			*Pourcentage (%) :*	

B. Mise en œuvre du programme					
Critères	2	1	0	N/A	Notes
1. Le mode de recrutement et les critères de sélection des participants sont spécifiés et documentés.					
2. Les auteurs rapportent des stratégies pour favoriser le recrutement des participants et le taux de participation est précisé.					
3. Les caractéristiques des participants sont décrites clairement					

4. Les composantes du programme sont décrites avec suffisamment de détails et un document de référence est disponible.					
5. Les auteurs ont déployé les moyens appropriés pour assurer la formation des intervenants et leur adhésion au programme.					
6. Les auteurs ont déployé les moyens requis pour favoriser l'exposition et l'engagement au programme de la part des participants.					
7. Les auteurs ont fait un examen des effets de débordement au sein du groupe expérimental et du groupe de contrôle.					
8. Les auteurs évaluent les éléments contextuels et personnels qui ont pu faciliter ou hypothéquer le déploiement du programme.					
9. Les auteurs utilisent des instruments adéquats pour évaluer les divers éléments de la mise en œuvre (c.-à-d. adhésion des intervenants et exposition des participants).					
10. Les auteurs ont prévu au moins deux sources pour l'évaluation de la mise en œuvre, dont une source indépendante non biaisée.					
11. Les auteurs ont évalué la mise en œuvre ainsi que l'adhésion et l'exposition; ces éléments atteignent un niveau jugé suffisant (80% ou plus).					
Nombre de critères retenus :	*Total de points obtenus sur max possible de:*			*Pourcentage (%) :*	

C. Choix du devis de recherche



Légende
Ø : absence
(0) : sans point
(1) : un point
(2) : deux points

	Total de points obtenus sur max possible de:	*Pourcentage (%) :*

D. Contrôle des sources d'invalidité interne					
Critères	2	1	0	N/A	Notes
1. L'influence possible des facteurs historiques est contrôlée.					
2. L'influence possible de la maturation est contrôlée.					
3. L'influence possible de la sélection des participants est contrôlée.					
4. L'influence possible de l'instrumentation est contrôlée.					
5. L'influence possible de la réactivité à la mesure est contrôlée.					
6. L'influence possible de la régression vers la moyenne est contrôlée.					
7. L'influence possible de la défection des participants est contrôlée.					
8. Les auteurs ont conservé tous les participants dans l'analyse.					
9. Les auteurs ont vérifié l'équivalence des groupes au prétest.					
10. Les auteurs ont fait l'analyse d'attrition différentielle (qualitative et quantitative) dans le cas de perte de participants.					
Nombre de critères retenus :	*Total de points obtenus sur max possible de:*			*Pourcentage (%) :*	

E. Évaluation des effets					
Critères	2	1	0	N/A	Notes
1. Les auteurs ont prévu des mesures en rapport avec chaque objectif proximal, intermédiaire et distal.					
2. Les auteurs ont utilisé des instruments valides et fidèles par rapport au point précédent.					
3. Les auteurs ont tenu compte du risque d'erreur de type 1 lorsque plusieurs mesures sont soumises aux analyses.					
4. Les auteurs ont procédé à un calcul de puissance statistique afin de s'assurer que le nombre de participants est suffisant en regard d'un risque de type 2.					
5. Les auteurs ont rapporté des résultats pour tous les instruments de mesure qu'ils ont administrés.					
6. Les auteurs ont fait un suivi d'au moins six mois des effets du programme et, dans le cas d'un programme de prévention, inclus une mesure de la variable distale.					
7. Les auteurs ont prévu une mesure des effets collatéraux, et possiblement iatrogènes, du programme.					

8. Les résultats sont statistiquement significatifs pour la majorité des mesures, à la suite de l'application des tests statistiques appropriées.					
9. La taille d'effet ou l'impact clinique de la majorité des résultats est de niveau élevé.					
10. Les résultats se maintiennent sur une période d'au moins six mois, ou jusqu'à l'atteinte de la variable distale.					
11. Les auteurs rapportent peu ou pas d'effets iatrogènes ou pervers, en particulier au chapitre des objectifs distaux.					
12. Les auteurs ont examiné formellement les variables potentiellement modératrices liées aux caractéristiques des participants ou aux conditions de mise en œuvre du programme.					
13. Les auteurs ont analysé formellement les mécanismes potentiellement médiateurs et ceux-ci sont compatibles avec le Modèle logique					
14. Il existe au moins une reproduction du programme par les mêmes auteurs.					
15. Il existe au moins une reproduction indépendante du même programme par une autre équipe de chercheurs.					
16. Le programme a fait l'objet d'une publication dans une revue avec comité de lecture.					
Nombre de critères retenus :	*Total de points obtenus sur max possible de:*			*Pourcentage (%) :*	

F. Généralisation et diffusion (optionnel)					
Critères	2	1	0	N/A	Notes
1. La participation au programme et sa mise en œuvre dans un contexte d'épreuve d'effectivité sont satisfaisants.					
2. Il existe un manuel ou une formation pour aider les intervenants à appliquer le programme de manière la plus fidèle possible.					
3. Les auteurs ont vérifié si les résultats diffèrent selon les catégories d'intervenants.					
4. Les auteurs ont évalué les caractéristiques des intervenants pour déterminer si elles sont associées au degré d'adhésion au programme.					
5. Les auteurs ont procédé à une épreuve d'effectivité en milieu réel et ont obtenu des tailles d'effet suffisantes.					
6. Les auteurs ont analysé les coûts/bénéfices (ou les coûts/efficacité) et les résultats de cette analyse sont avantageux.					
7. Il existe au moins une reproduction indépendante du programme dans le cadre d'une épreuve d'effectivité.					

8. Chaque reproduction du programme montre des effets de taille minimalement modérée ou significatifs au plan statistique et aucun effet iatrogène.					
9. Le recrutement des participants a fait appel à une technique d'échantillonnage représentative de la population de référence.					
10. Il existe des instruments pour monitorer le suivi de la mise en œuvre du programme et de ses effets.					
11. Le programme est intégré dans les pratiques quotidiennes des intervenants.					
12. Il existe une stratégie de transfert et d'appropriation des éléments efficaces du programme et de leur rationnel théorique.					
13. L'offre de programme est équitable.					

Nombre de critères retenus :	*Total de points obtenus sur max possible de:*	*Pourcentage (%) :*