Les méthodes d'expérimentation en question

Denis Fougère
CNRS, Centre de recherche en économie et statistique (CREST, Paris),
Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP, Sciences po, Paris),
IZA (Bonn) et CEPR (Londres)

Les expérimentations randomisées, également appelées expériences controlées, sont des méthodes d'évaluation rigoureuses, mais leur mise en œuvre soulève des problèmes qui peuvent biaiser les estimations que l'on en déduit. Ces biais sont le plus souvent inhérents à la démarche expérimentale. Leur nature, et les éventuelles corrections qu'il est possible de leur apporter, sont ici passées en revue. L'accent est également mis sur la question des plans et protocoles d'expérience, trop souvent ignorée par les praticiens de ces méthodes statistiques appliquées aux sciences sociales[1].

Parmi les méthodes utilisées pour évaluer des politiques sociales ou éducatives, les expérimentations randomisées font aujourd'hui l'objet d'un intérêt tout particulier. Dans le champ de l'éducation, elles peuvent être utilisées pour évaluer les effets d'une innovation pédagogique, d'une modification des rythmes scolaires, de l'introduction de cours de soutien, etc., et cela, avant même que de telles interventions soient ou non généralisées. Les plaidoyers en faveur des expérimentations mettent souvent l'accent sur les biais statistiques inhérents aux évaluations conduites à l'aide de données d'observation non expérimentales. L'évaluation par expérimentation randomisée essaie de s'affranchir de ces biais de sélection en proposant d'affecter de manière aléatoire (par tirage au sort) les établissements scolaires, classes ou élèves échantillonnés à un groupe de traitement qui bénéficiera du dispositif ou programme éducatif devant être évalué, ou bien à un groupe de contrôle qui n'en bénéficiera pas.

Cet article ne se propose pas de synthétiser les résultats des expérimentations randomisées conduites en milieu scolaire, à l'étranger ou en France. Son objectif est plus modeste. Il s'agit ici de dresser la liste des principales difficultés auxquelles les praticiens de ces méthodes doivent faire face. Pour la plupart, ces difficultés ne sont pas insurmontables, pour autant qu'elles soient bien identifiées. Certaines sont toutefois très sérieuses et entraînent des biais statistiques significatifs dont l'ampleur ne peut être facilement évaluée.

RÉTICENCES ET OPPOSITIONS

Les réticences et oppositions à la mise en œuvre d'expérimentations randomisées reposent en général sur plusieurs types d'arguments. Le premier type a trait aux coûts élevés des expérimentations. Ces coûts sont de natures diverses. Les expérimentations sont certes coûteuses en termes de budget et de temps, mais l'on doit reconnaître que, dans la collecte des données non expérimentales d'observation qui se présentent sous la forme de fichiers longitudinaux, ces deux types de coûts sont souvent à peine moins importants. Les coûts

NOTE

1. Je remercie Cédric AFSA pour sa lecture attentive et ses remarques constructives.

spécifiques aux expérimentations sont des coûts de nature politique : persuader les chefs d'établissement, les enseignants, les parents d'élèves, des avantages et nécessités d'une expérimentation entraîne certes des dépenses, mais fait également courir un risque de perte de réputation ou de crédibilité aux concepteurs de l'expérience et aux groupes de pression qui les soutiennent. On le sait, les opposants au principe des expérimentations sociales invoquent plus souvent des arguments éthiques que budgétaires pour faire obstacle à leur mise en œuvre. Ils condamnent le déni de traitement imposé aux personnes (ici, aux élèves) du groupe de contrôle et le tort substantiel que ce déni peut leur faire subir. Si le programme s'avère être significativement bénéfique, refuser ce traitement à certains à l'issue d'un tirage aléatoire et de ce fait rationner une ressource publique est contraire à l'éthique. Mais, comme le fait remarquer Burtless **[1]**, l'argument est le même dans les expérimentations cliniques et, bien que les enjeux individuels y soient souvent bien plus importants, le recours à l'expérimentation par randomisation y est plus facilement admis. Devant décider de la justesse du recours à l'expérimentation randomisée, il nous faut donc répondre à une et une seule question : est-il *a priori* plus juste, ou plus bénéfique, de généraliser, éventuellement en le rendant obligatoire, un programme qui risque d'être préjudiciable, de ne pas rendre accessible à tout un chacun un programme qui peut s'avérer profitable, ou bien de mettre en place, préalablement à toute décision de généralisation ou d'abandon définitif du programme, une procédure d'évaluation par expérimentation avec l'aide et le consentement éclairé d'un nombre limité d'individus dont certains acceptent de se voir refuser, à l'issue d'un tirage au sort et à l'exclusion de toute autre considération (de revenu, de sexe, d'origine sociale ou nationale), le droit d'accès au programme durant la période limitée de l'expérience ? Cette dernière solution ne peut être envisagée que sous certaines conditions très précises, admises par les acteurs concernés à l'issue d'un débat préalable. Ces conditions, qui tout à la fois permettent de recourir à l'expérimentation et déterminent son protocole, sont au minimum au nombre de trois :

1. les participants à l'expérience doivent être pleinement informés des avantages comme des risques que leur participation peut leur procurer ou leur faire courir (principe de consentement éclairé) ;
2. dans les cas où l'on anticipe que le programme peut provoquer des risques ou des torts, il faut prévoir de compenser les élèves qui seront affectés au groupe de traitement; dans le cas où l'on anticipe que le programme peut être bénéfique, il faut prévoir de compenser ceux qui seront affectés au groupe de contrôle (principe de compensation) ;
3. enfin, une expérimentation n'est envisageable que dans les cas où l'on ignore *a priori* si le programme est bénéfique ou non ; s'il s'avère préjudiciable, il est préférable d'observer ses effets sur un nombre limité d'élèves consentants que sur un très grand nombre d'élèves contraints (principe de précaution).

Un des enseignements les plus importants des expériences randomisées qui ont été conduites dans le domaine social, économique ou éducatif, tout d'abord aux Etats-Unis puis plus récemment en Europe, est que leur réussite, c'est-à-dire l'assurance de la fiabilité de leurs conclusions, dépend de la participation étroite des personnels chargés de leur mise en œuvre, à la définition de leurs objectifs et de leurs modalités pratiques. Obtenir cet agrément est, pour l'administration centrale ou locale qui souhaite mettre en place une expérience, chose d'autant plus facile que le plan (*le design*) de l'expérience a été préalablement élaboré avec les acteurs chargés de la réalisation de l'opération et que ce plan est perçu comme respectant les objectifs et contraintes de l'expérience.

Mais, dans ce cas favorable, les résultats d'une expérimentation randomisée sont susceptibles d'être soumis à un certain nombre de biais inhérents, que nous allons maintenant énumérer, et dont il faut avoir conscience.

Le principe des expérimentations randomisées

Traditionnellement, les études statistiques qui tentent d'identifier et d'estimer les effets de politiques éducatives (par exemple, la création des zones d'éducation prioritaire) exploitent des données provenant d'enquêtes non expérimentales recueillies auprès d'établissements comparables. Pour mener l'évaluation, ces études utilisent généralement une comparaison des caractéristiques moyennes (par exemple, le taux de redoublement, les notes aux épreuves de mathématiques ou de français, etc.) de deux types d'établissements, ceux ayant bénéficié de la mesure examinée (en ce cas, la labellisation EP) et ceux n'en ayant pas bénéficié. La différence des moyennes de résul-

tats dans les deux groupes est en fait la combinaison de deux effets, l'effet spécifique du dispositif que l'on cherche à mesurer, mais aussi celui qui provient du fait que les deux groupes d'établissements peuvent se comporter différemment face au dispositif considéré. En effet, ce dispositif peut ne pas intéresser de la même façon tous les établissements. Certains peuvent demander avec force à bénéficier du label et des avantages qu'il procure, d'autres au contraire peuvent être réticents, estimant que la stigmatisation créée par un label de ce type est supérieure aux avantages afférents. En ce cas, la différence de résultats entre les deux groupes d'établissements reflète en partie ces comportements d'autosélection.

Pour qu'une expérimentation randomisée élimine le biais de sélection (ou d'autosélection), deux conditions doivent être vérifiées :

1. la probabilité de candidature au programme considéré ne doit pas être affectée par le fait que l'accès à ce programme est conditionné par le tirage au sort ;

2. les comportements ne doivent pas être modifiés par la participation à l'expérimentation.

Dans une expérimentation randomisée, le plan d'expérience le plus fréquent est celui où les établissements éligibles au dispositif sont d'abord informés des buts et principes de l'expérimentation, ceux acceptant d'y participer étant ensuite affectés aux groupes de traitement et de contrôle par tirage au sort. Le fait que la décision de participation à l'expérience précède le tirage au sort garantit que les groupes de contrôle et de traitement constituent des sous-échantillons aléatoires de la population participante. En ce cas, l'effet moyen du programme pour les établissements qui souhaitent participer au dispositif peut être estimé par la différence entre la moyenne des résultats dans le groupe de traitement et la même moyenne dans le groupe de contrôle.

Lorsque le tirage au sort précède le choix des établissements de participer ou non au programme proposé, l'effet moyen de ce programme est plus difficilement identifiable. Le fait que certains établissements refusent de participer au programme dans le cadre de l'expérimentation ne permet pas d'estimer la valeur moyenne de l'effet du programme, une fois celui-ci généralisé. Manski **[10]** a néanmoins démontré qu'il est possible de borner la valeur de cet effet moyen, et qu'il est également possible de borner l'effet moyen d'une mesure généralisée mais facultative à partir des résultats d'une expérimentation randomisée.

Le biais de randomisation

Les expérimentations avec assignation aléatoire sont susceptibles d'être soumises à ce que les spécialistes de ces techniques appellent le biais de « randomisation ». Ce biais apparaît lorsque le groupe de traitement, constitué à l'issue de la procédure d'affectation aléatoire des participants, diffère du groupe de personnes auxquelles l'intervention (par exemple, la politique éducative) pourrait être appliquée, une fois l'expérience menée et jugée concluante. Ce biais a été maintes fois observé dans des essais cliniques, mais aussi dans de nombreuses expérimentations sociales. Il est en effet souvent plus difficile de persuader des personnes de participer à une expérience randomisée que de les inciter à participer à une observation non expérimentale. Cette difficulté peut, par exemple, apparaître lorsqu'il s'agit de demander à des chefs d'établissement scolaire, des enseignants, des élèves ou des parents d'élèves de ces établissements, préalablement choisis de manière aléatoire, de participer à une expérimentation randomisée, et qu'un certain nombre d'entre eux refusent, ou bien, pire encore, acceptent, puis cessent de participer à l'expérience avant que celle-ci parvienne à son terme.

Par exemple, il est à craindre que les établissements qui refusent de participer à l'évaluation expérimentale d'une nouvelle politique éducative soient précisément ceux qui pensent en tirer le moins d'avantages. En ce cas, la différence des moyennes de résultat entre les deux groupes de traitement et de contrôle ne mesure pas vraiment ce que l'on souhaite mesurer, à savoir, l'effet moyen de la politique éducative, mais l'effet moyen de cette politique au sein des établissements, des classes, etc., qui ont accepté de participer à l'expérience. L'écart entre ces deux quantités peut être assez substantiel, comme l'ont montré des expériences conduites dans différents domaines.

Les effets de type « Hawthorne »

Les effets de type « Hawthorne » sont une variante du biais de randomisation. Dans le cadre expérimental, ce type d'effet correspond au cas où le comportement des sujets étudiés est modifié par le fait d'avoir été ou non retenu pour participer à l'expérience, et non par le dispositif ou le programme lui-même. En effet, certains peuvent éprouver un sentiment de gratitude

ou au contraire d'injustice selon qu'ils ont été ou non choisis pour participer à une expérience susceptible d'améliorer leur situation matérielle. Le terme provient des expériences menées dans l'établissement d'Hawthorne de la Western Electric Company, à Chicago entre 1924 et 1933. La première de ces expériences montra que des variations de la luminosité sur le lieu de travail avaient pour effet une amélioration de la productivité et de la satisfaction des salariés, indépendamment du fait que la lumière fut plus ou moins forte. En présence d'effets de type « Hawthorne », l'estimation de l'impact du programme risque, là encore, d'être biaisée, le sens du biais (positif ou négatif) étant ici difficile à prévoir. Seule une observation très étroite des élèves, professeurs, chefs d'établissement, etc., permet alors de comprendre comment, et dans quelle mesure, leur comportement peut être éventuellement modifié par leur participation à l'expérimentation.

La faiblesse des effectifs

Le biais de randomisation peut également résulter de la faiblesse des effectifs participant à l'expérience, les résultats étant en ce cas très imprécis. Par exemple, de trop petits échantillons conduisent trop souvent à rejeter l'hypothèse d'un effet positif (ou négatif) de l'intervention ou du programme expérimental, non parce que la différence des moyennes de résultat observés dans les deux groupes est effectivement positive (ou négative) mais parce que son écart-type est trop large pour aboutir à une conclusion suffisamment fiable. Étayant leurs analyses par des exemples concrets, Heckman **[4]**, Heckman et Smith **[6]**, et Manski **[9]** ont ainsi pu montrer que des expériences conduites à l'aide d'échantillons trop faibles peuvent conduire à des résultats sensiblement différents de celles réalisées à l'aide de groupes de beaucoup plus grande taille. Les deux difficultés peuvent se cumuler dès lors que le consentement des participants pressentis peut dépendre de la taille des échantillons constitués ; pour eux, il peut être plus facile de participer à une expérience qui impliquera un grand nombre d'élèves ou d'établissements, les écarts individuels à la moyenne passant alors plus vraisemblablement inaperçus et les éventuels effets de stigmatisation étant de ce fait atténués. Mais cet avantage a une contrepartie : une expérimentation de grande dimension est tout à la fois plus coûteuse et plus difficile à mettre en œuvre.

Le biais d'attrition

Ce biais résulte du fait que certains participants, membres du groupe de traitement ou du groupe de contrôle, décident de quitter l'expérience avant la fin de celle-ci. Ce phénomène se produit plus fréquemment dans les expériences qui se déroulent sur une échelle de temps assez longue, par exemple sur une ou plusieurs années. L'importance de ce biais, inhérent aux expérimentations qui portent sur des sujets humains, a été reconnue très tôt. Hausman et Wise **[3]** faisaient remarquer que le problème de l'attrition provient *de facto* de l'introduction du facteur temporel dans les expérimentations sociales ou cliniques et que ce problème n'apparaît pas dans les expérimentations généralement conduites en physique, en chimie ou en biologie. On a pu, par exemple, observer un tel phénomène dans le cadre du programme STAR, qui a permis de tester, aux États-Unis, les effets d'une réduction de la taille des classes dans le cadre d'un programme expérimental conduit sur une très grande échelle **[2]**. Le biais d'attrition est difficilement réductible. Heckman, Smith et Taber **[7]** ont précisé la condition sous laquelle il est possible d'identifier l'effet moyen du programme sur le résultat des membres du groupe de traitement qui participent pleinement à l'expérimentation, c'est-à-dire jusqu'à son terme (cette quantité n'est pas exactement l'effet moyen du programme sur ceux qui en ont bénéficié, cette dernière quantité ne pouvant être identifiée en présence d'attrition). Cette condition stipule que le résultat moyen des membres du groupe de traitement qui abandonnent l'expérimentation avant qu'elle ne prenne fin doit être égal au résultat moyen des membres du groupe de contrôle qui auraient abandonné de la même façon l'expérimentation s'ils avaient été placés dans le groupe de traitement. Heckman, Smith et Taber **[7]** proposent des tests statistiques permettant de vérifier la validité de cette condition.

Le biais de substitution

Ce biais peut apparaître dès lors que certains membres du groupe de contrôle, n'ayant pas accès au programme éducatif durant le temps de l'expérience, cherchent à se procurer des substituts à ce programme avant la fin de l'expérience. Pensons à un programme éducatif qui accroît le nombre d'heures de mathématiques. Il est possible que les parents des élèves des classes faisant partie du

groupe de contrôle recourent alors à des cours privés, ou bien que les professeurs en charge de ces classes, hostiles au principe de l'expérimentation, décident unilatéralement d'intensifier le nombre et le contenu de l'enseignement de mathématiques. Il est clair qu'il est difficile de leur interdire d'agir ainsi, leur argument étant de ne pas défavoriser les élèves qui n'ont pas accès à ce programme qu'ils supposent bénéfique. Là encore, réduire ce biais, qui tend à sous-estimer les effets du programme ou de la politique éducative que l'on souhaite évaluer, est chose difficile. Heckman **[5]** a toutefois précisé les conditions sous lesquelles il reste possible d'identifier l'effet moyen du dispositif pour ceux qui en bénéficient dans le cas où le groupe de contrôle a accès à un substitut. La première condition stipule que, pour chaque participant, membre du groupe de traitement ou du groupe de contrôle, l'effet du dispositif devant être évalué est le même que celui de son substitut. La seconde condition est que l'effet du dispositif évalué doit être le même pour chaque participant. Enfin, il faut que l'effet moyen du dispositif pour les membres du groupe de traitement qui participent à l'expérimentation jusqu'à son terme soit égal à l'effet moyen du substitut au dispositif pour les membres du groupe de contrôle qui ont accès à ce substitut. On le voit, ces conditions sont extrêmement strictes et n'ont que peu de chances d'être vérifiées en pratique. En conséquence, la seule solution pragmatique, quoique difficile à mettre en œuvre, est de limiter, autant que faire se peut, les possibilités d'accès à des substituts au sein du groupe de contrôle.

Mieux définir les protocoles et plans d'expérience

Dans de trop nombreux cas, les chercheurs conduisant des expérimentations randomisées en sciences sociales négligent de définir le protocole statistique qui assurera l'efficacité maximale des résultats de l'expérimentation, et ce, avant même que celle-ci ne débute sur le terrain. Ce souci est depuis longtemps au cœur de la démarche des statisticiens conduisant des expérimentations dans le cadre clinique ou dans d'autres disciplines (agronomie, physique, chimie, etc.), mais il semble avoir été jusqu'à aujourd'hui largement absent des préoccupations des chercheurs en sciences sociales ou en sciences de l'éducation. De ce fait, on a souvent pu constater que les estimations déduites des expérimentations sociales étaient statistiquement peu significatives et/ou peu robustes.

Le protocole expérimental doit être avant tout calibré par un calcul de puissance[2]. Préalablement au lancement de l'expérimentation, les chercheurs doivent s'efforcer d'énoncer clairement la question à laquelle ils souhaitent répondre et spécifier l'hypothèse nulle correspondante (par exemple, qu'il n'existe pas de différence de résultats entre les deux groupes à l'issue de la mise en œuvre du programme expérimental dans le groupe de traitement). Ils doivent également faire des hypothèses concernant l'espérance mathématique de la variable d'intérêt dans le groupe de traitement (par exemple, le taux de redoublement est en moyenne plus faible dans ce groupe). Ces hypothèses leur permettent ensuite de calculer, compte-tenu du plan d'expérience, la taille minimale des échantillons permettant de détecter avec une certaine puissance (souvent fixée à 80 %) un écart d'une certaine valeur entre les deux groupes. Plus l'écart postulé est élevé, plus l'hypothèse nulle est facile à rejeter, et plus la taille de l'échantillon est faible. Ce genre d'exercice permet de savoir *a priori* si le budget de l'expérience permet ou non de détecter un écart d'une certaine ampleur. Lorsque les chercheurs rédigent leur rapport, joindre le calcul de puissance permet d'informer le lecteur sur les hypothèses faites *ex ante* par les chercheurs.

Dans une étude récente, List, Sadoff et Wagner **[8]** énoncent quatre règles permettant d'améliorer l'efficacité statistique des protocoles expérimentaux :

1. lorsque la variable de résultat est continue (par exemple, une note à un examen ou à une évaluation) et lorsque la variance de la moyenne de cette variable est supposée être la même dans les deux groupes (de traitement et de contrôle), il faudrait affecter les sujets (ici les élèves) de

NOTE

2. La puissance statistique est la probabilité que l'hypothèse nulle (ici, l'absence de différence de résultats entre les deux groupes) soit rejetée, et que l'expérimentation puisse donc permettre de repérer l'association réellement existante entre le programme éducatif évalué et la variable de résultat considérée (par exemple, le taux de redoublement ou la moyenne des notes en mathématiques). La puissance est déterminée par différents facteurs, parmi lesquels la fréquence de la variable de résultat considérée, le protocole de l'expérience et la taille de l'échantillon. Lors de la mise en place de l'expérience, les chercheurs doivent opter pour une certaine puissance en fonction de laquelle la taille de l'échantillon est ensuite déterminée. Une puissance statistique de 80 % est généralement considérée comme le minimum exigible. Ce qui signifie qu'il y a 80 % de chance que l'expérimentation puisse mettre en évidence l'effet recherché.

manière égale dans les deux groupes ;

2. lorsque les variances ne sont pas identiques, le rapport des effectifs des deux groupes devrait être égal au rapport des écarts-types de la moyenne des résultats dans les deux groupes ;

3. lorsque le coût d'échantillonnage varie d'une cellule expérimentale à l'autre (une cellule pouvant être ici un établissement, une classe, un groupe d'élèves homogène), le rapport des tailles d'échantillon devrait être inversement proportionnel à la racine carrée des coûts relatifs ;

4. lorsque l'unité que l'on tire aléatoirement (par exemple, un établissement) est différente de l'unité d'analyse (par exemple, une classe), il faudrait tenir compte du coefficient de corrélation entre classes d'un même établissement.

La qualité des évaluateurs

Cet aspect, souvent ignoré, n'est pas le moindre risque encouru par une expérimentation randomisée. La qualité des agences ou opérateurs en charge de la mise en œuvre de l'expérimentation est par essence variable. Ces opérateurs sont en général choisis à l'issue d'une procédure d'appel d'offres, pilotée par un comité d'usagers, de praticiens et de personnes qualifiées. Rien ne garantit que la proposition effectivement choisie émane d'un groupe d'experts de qualité suffisante. Somme toute, le nombre de statisticiens de valeur consacrant leur temps et leur énergie à la pratique des expérimentations est assez limité. Il est donc possible que les équipes répondant aux appels d'offres soient parfois de qualité moyenne et ne contiennent que rarement des chercheurs présentant des garanties de qualité suffisantes. Plus la pratique des expérimentations, et notamment des expérimentations à grande échelle, se diffusera, et plus le problème de qualité des évaluateurs se posera. Certains répondront qu'il suffit de former de plus en plus de jeunes chercheurs de valeur à cette méthode d'évaluation. Mais l'expérimentation n'est pas la seule technique statistique nécessaire à notre compréhension des phénomènes économiques, politiques et sociaux. Elle n'est pas non plus la plus exigeante, bien que les difficultés statistiques qu'elle soulève soient réelles, et trop souvent ignorées par les économistes et sociologues qui y consacrent tout ou partie de leur carrière.

Conclusion

Les expérimentations randomisées sont aujourd'hui un dispositif d'évaluation de plus en plus utilisé et de mieux en mieux maîtrisé. Les gains que l'on peut en attendre dans un domaine aussi important que celui des politiques éducatives sont substantiels. Pour autant, leur mise en œuvre peut encore être améliorée. Parmi les progrès les plus notables devant être enregistrés, figure certainement la meilleure connaissance des problèmes méthodologiques qui leur sont inhérents. Cet article en a dressé la liste. Il appartient maintenant aux chercheurs d'œuvrer à la résolution de ces problèmes et de proposer des protocoles statistiques qui permettent d'obtenir des résultats fiables, et surtout des interprétations analytiques allant au-delà du simple constat « cela fonctionne/cela ne fonctionne pas ». Pour le dire autrement, pour ceux qui veulent persuader du bien-fondé de la démarche expérimentale en sciences sociales, et tout particulièrement dans le domaine de l'éducation, l'impératif est aujourd'hui, selon moi, d'abandonner l'effervescence praticienne pour atteindre la maturité scientifique. À défaut de cette transition, beaucoup d'expérimentations randomisées risquent de livrer des résultats peu convaincants et insuffisamment articulés. ■

BIBLIOGRAPHIE

[1] Burtless G., 1995, The Case for Randomized Field Trials in Economic and Policy Research, *Journal of Economic Perspectives*, vol. 9, n° 2, p. 63-84.

[2] Ding W., and Lehrer S. F., 2010, Estimating Treatment Effects from Contaminated Multiperiod Education Experiments: The Dynamic Impacts of Class Size Reductions. *The Review of Economics and Statistics*, vol. 92, n° 1, p. 31-42.

[3] Hausman J., and Wise D., 1979, Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment. *Econometrica*, Vol. 47, n° 2, p. 455-473.

[4] Heckman J., 1992, Randomization and Social Policy Evaluation, dans *Evaluating Welfare and Training Programs*, édité par C. Manski et I. Garkinkel. Cambridge, Massachusetts, Harvard University Press, p. 201-230.

[5] Heckman J., 1997, Instrumental Variables: A Study of Implicit Behavioral Assumptions in One Widely Used Estimator. *The Journal of Human Resources*, vol. 32, n° 3, p.441-461.

[6] Heckman J. and Smith J., 1995, Assessing the Case for Social Experiments, *Journal of Economic Perspectives*. vol. 9, n° 2, p. 85-110.

[7] Heckman J., Smith J., and C. Taber, 1998, Accounting for Dropouts in Evaluation of Social Programs. *The Review of Economics and Statistics*, vol. 80, n° 1, p. 1-14.

[8] List J., Sadoff S. and Wagner M., 2010, *So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design*, National Bureau of Economic Research, Working Paper 15701, Cambridge, Massachusetts.

[9] Manski C., 1995. *Identification Problems in the Social Sciences*. Harvard University Press, 172 pages.

[10] Manski C., 1997, The Mixing Problem in Programme Evaluation. *The Review of Economic Studies*, vol. 64, n° 4, p. 537-554.