Mesure et évaluation en éducation

Le score de propension : un guide méthodologique pour les recherches expérimentales et quasi expérimentales en éducation

Aurélie Lecocq, Mehdi Ammi et Élodie Bellarbre

Volume 37, numéro 2, 2014

Date de réception : 8 mars 2013
Date de réception de la version finale : 11 novembre 2013
Date d'acceptation : 12 mars 2014

URI : https://id.erudit.org/iderudit/1035914ar
DOI : https://doi.org/10.7202/1035914ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Lecocq, A., Ammi, M. & Bellarbre, É. (2014). Le score de propension : un guide méthodologique pour les recherches expérimentales et quasi expérimentales en éducation. *Mesure et évaluation en éducation*, *37*(2), 69–100. https://doi.org/10.7202/1035914ar

Résumé de l'article

La méthode du score de propension devient de plus en plus populaire pour estimer les effets causaux d'un programme d'intervention. Si les applications empiriques de cette méthode sont encore rares dans les recherches en éducation, des exemples de son utilisation se trouvent aisément dans d'autres disciplines. Cependant, sa mise en place soulève plusieurs questions. L'objectif de cet article est de fournir des éléments de réponses guidant le chercheur et l'évaluateur du domaine de l'éducation pour l'estimation et l'utilisation du score de propension. Les différentes étapes de son application sont présentées pas à pas : évaluation du biais de sélection, construction du score de propension et mesure de sa qualité, et choix des stratégies d'utilisation du score dans l'estimation des effets d'un traitement. Les questions méthodologiques soulevées sont discutées à chaque étape. Pour faciliter la compréhension, un exemple d'une expérimentation en maternelle illustre la méthode.

Le score de propension : un guide méthodologique pour les recherches expérimentales et quasi expérimentales en éducation

Aurélie Lecocq
Université d'Ottawa

Mehdi Ammi
Carleton University

Élodie Bellarbre
Université de Bourgogne

Mots clés : score de propension, appariement sur score de propension, pondération inverse sur les probabilités d'être traité, doubles différences, biais de sélection.

La méthode du score de propension devient de plus en plus populaire pour estimer les effets causaux d'un programme d'intervention. Si les applications empiriques de cette méthode sont encore rares dans les recherches en éducation, des exemples de son utilisation se trouvent aisément dans d'autres disciplines. Cependant, sa mise en place soulève plusieurs questions. L'objectif de cet article est de fournir des éléments de réponses guidant le chercheur et l'évaluateur du domaine de l'éducation pour l'estimation et l'utilisation du score de propension. Les différentes étapes de son application sont présentées pas à pas : évaluation du biais de sélection, construction du score de propension et mesure de sa qualité, et choix des stratégies d'utilisation du score dans l'estimation des effets d'un traitement. Les questions méthodologiques soulevées sont discutées à chaque étape. Pour faciliter la compréhension, un exemple d'une expérimentation en maternelle illustre la méthode.

Keywords: propensity score, propensity score matching, inverse probability of treatment weighting, difference in differences, selection bias.

Propensity score methods become increasingly popular for estimating the causal effects of an intervention. If empirical applications of this method are still unusual in education researches, examples can be found easily in other fields. However, its implementation raises several questions. The aim of this paper is to provide guidance to educational researchers regarding the estimation and the utilization of the propensity score. The various stages of its application are presented step by step:

evaluation of the selection bias, construction of the propensity score and measurement of its quality, and decision on the strategies to use in the estimation of treatment effects. Methodological issues are discussed at each stage. To facilitate the understanding, an example of an experiment in kindergarten illustrates the method.

Palavras-chave: pontuação de propensão, aparelhamento da pontuação de propensão, ponderação inversa sobre as probabilidades de ser tratado, diferenças duplas, viés de seleção.

O método da pontuação de propensão está tornar-se cada vez mais popular para estimar os efeitos causais de um programa de intervenção. Embora as aplicações empíricas deste método ainda sejam raras na investigação em educação, exemplos do seu uso podem ser facilmente encontrados noutras disciplinas. No entanto, a sua implementação levanta diversas questões. O objetivo deste artigo é fornecer algumas respostas para orientar o investigador e o avaliador do domínio da educação para a estimativa e uso da pontuação de propensão. As diferentes etapas da sua aplicação são apresentadas passo a passo: avaliação do viés de seleção, construção da pontuação de propensão e medição da sua qualidade e escolha das estratégias para o uso da pontuação na estimativa dos efeitos do tratamento. As questões metodológicas são discutidas em cada etapa. Para facilitar o entendimento, um exemplo de uma experiência no jardim-de-infância ilustra o método.

Note des auteurs : La correspondance liée à cet article peut être adressée à Aurélie Lecocq à l'adresse courriel suivante : [aurelie.lecocq@oghma.ca].

Les données utilisées pour illustrer la méthode sont issues d'un projet financé par le Haut-Commissariat à la jeunesse (HCJ) intitulé *Stimuler les capacités cognitives pour éviter l'échec scolaire*. L'expérimentation musicale a été rendue possible grâce à l'expertise de Bruno Suchaut.

Les méthodes d'évaluation des programmes et des politiques éducatives reposent essentiellement sur la comparaison d'individus bénéficiant de l'intervention, soit le groupe traité ou expérimental, avec des individus n'en bénéficiant pas, soit le groupe témoin ou de contrôle. Deux types de méthode peuvent être utilisés selon que les individus bénéficient ou non de l'intervention d'après un critère aléatoire.

L'expérimentation sociale randomisée se fonde sur une sélection aléatoire des participants et sur une allocation aléatoire du traitement[1], c'est-à-dire une répartition aléatoire des individus dans le groupe témoin ou dans le groupe traité (Altman & Bland, 1999; Banerjee & Duflo, 2009). Les comportements des individus sont alors mesurés et comparés selon les conditions de traitement afin de tester l'hypothèse à vérifier ou, plus globalement, de répondre à la question des effets du traitement sur leurs comportements. Outre le facteur à tester, les dispositifs extérieurs au traitement expérimental doivent être rigoureusement identiques pour les deux groupes. Cette méthode de recherche se développe fortement en sciences de l'éducation et permet d'évaluer les effets d'un programme d'intervention sur divers aspects scolaires : diminuer le sentiment dépressif en milieu scolaire (Marcotte & Baron, 1993), gérer le stress (Dumont, Leclerc, Massé, Potvin, & McKinnon, 2009), lutter contre le décrochage scolaire (Fortin, Royer, Potvin, Marcotte, & Yergeau, 2004 ; Fortin, Marcotte, Potvin, Royer, & Joly, 2006) ou encore améliorer la relation enseignant-élève (Murray & Malmgren, 2005). Les résultats de ces études sont à l'origine de stratégies de transfert des connaissances qui ont permis de développer de nombreux outils (dépistage, programme d'intervention, guide de bonnes pratiques), comme *Pare-Chocs*, un programme d'intervention auprès d'adolescents dépressifs (Marcotte, 2006) ; *Funambule : pour une gestion équilibrée du stress* (Dumont et al., 2009) ; le *Logiciel de dépistage du décrochage scolaire* (Fortin & Potvin, 2007) ; *Trait d'union,* un programme de prévention du décrochage scolaire (Fortin, 2012) ou encore le *QES*, disponible en version web, qui est un ensemble d'outils d'évaluation permettant de tracer le portrait d'une école quant à son environnement socio-éducatif (Janosz, 2012).

En plus de cette méthode, l'évaluation des politiques ou des programmes éducatifs se base également sur des devis quasi expérimentaux. Le groupe témoin y est généralement constitué de manière ad hoc et, surtout, contrairement aux études expérimentales, les individus ne sont pas randomisés. La réalisation d'une intervention dans certaines écoles sélectionnées selon des critères géographiques, mais pas dans d'autres qui servent alors de groupe témoin, comme la mise en place de la politique des zones d'éducation prioritaire (ZEP), est un exemple de politique menant à ce type de devis.

Qu'il s'agisse d'expérimentation sociale ou d'étude à devis quasi expérimental, l'évaluation repose sur une hypothèse forte selon laquelle les groupes traité et témoin sont suffisamment comparables pour permettre une mesure des effets nets de l'intervention. Autrement dit, il suffirait de comparer la différence des résultats entre les deux groupes pour obtenir une mesure non biaisée des effets du traitement[2]. Néanmoins, pour un ensemble de raisons qui sera présenté au cours de cet article, les caractéristiques des individus des groupes traité et témoin peuvent ne pas être comparables, ce qui donne lieu à un biais de sélection (Heckman, Ichimura, & Todd, 1998). Avec un tel biais, un effet du traitement peut être détecté là où il n'y en a pas ou, inversement, un effet réel peut être amoindri (Wood et al., 2008).

Les méthodes statistiques et économétriques d'évaluation ont permis des progrès considérables dans la gestion du biais de sélection au cours de la dernière décennie (Imbens & Wooldridge, 2009). Parmi l'ensemble des techniques disponibles, la méthode du score de propension devient de plus en plus populaire dans de nombreuses disciplines, mais demeure encore peu utilisée dans les recherches en éducation. Si elle est d'un grand intérêt pour le domaine de l'éducation, la mise en place de cette méthode soulève de nombreuses questions.

L'objectif de cet article est de fournir un guide pour le chercheur et l'évaluateur du domaine de l'éducation pour l'estimation et l'utilisation du score de propension dans le cas le plus fréquent : deux groupes (traité ou non) et au moins deux périodes de temps (avant et après). L'article est organisé comme suit. Une première section présente le score de propension et justifie son utilisation. Y sont expliqués les deux hypothèses fondamentales sur lesquelles le score de propension repose et son intérêt selon le type de méthode d'évaluation de programme utilisé. La deuxième sec-

tion est consacrée à l'estimation du score de propension et fournit de l'information relative au choix des variables à inclure dans le modèle, à l'évaluation de la qualité du score et à la délimitation du support commun. Une troisième section examine trois méthodes utilisant le score de propension : l'appariement, la pondération inverse sur la probabilité d'être traité et l'analyse en doubles différences sur score de propension ; les avantages et les limites de chacune des trois méthodes seront présentés. Enfin, la dernière section est consacrée à un exemple d'application dans le domaine de l'éducation qui illustre la méthode. Il s'appuie sur les données d'une expérimentation ayant pour objectif d'accroître les capacités cognitives des élèves de maternelle.

Le score de propension

Définition et hypothèses

Initialement introduit par Rosenbaum et Rubin en 1983 dans un article intitulé *The central role of the propensity score in observational studies for causal effects*, le score de propension, connu sous le terme anglais *propensity score*, désigne la probabilité d'être exposé à un traitement, selon un ensemble de caractéristiques observables. Cette méthode, largement utilisée dans les recherches quasi expérimentales en économie (voir Lechner, 2002, pour un exemple sur les politiques du marché du travail) ou en épidémiologie, se révèle particulièrement prometteuse pour les futures recherches en éducation (Rubin, Stuart, & Zanutto, 2004).

Le score de propension peut être calculé de deux manières : par régressions logistiques et par les arbres de classification et de régression (connus sous le terme anglais c*lassification and regression tree analysis* – CART). La régression logistique, qui prédit la probabilité d'occurrence d'un événement, est la technique la plus communément utilisée pour estimer les scores de propension.

Sur le plan statistique, le score de propension $e(x_i)$ estime, pour chaque individu i, la probabilité conditionnelle P de recevoir le traitement étudié z, étant donné ses caractéristiques initiales (x_i) :

$$e(x_i) = \mathrm{P}\,(z_i = 1 | \, x_i)$$

où :

$z_i = 1$ pour traitement

$z_i = 0$ pour contrôle

x_i = l'ensemble des covariables observées pour le $i^{ème}$ sujet.

Le score de propension est une probabilité ; sa valeur est comprise entre 0 et 1. Il s'appuie sur deux hypothèses fondamentales : (1) l'hypothèse d'indépendance conditionnelle à des caractéristiques observables (*conditional independance assumption* – CIA) et (2) l'hypothèse de la condition de support commun (*overlap*).

L'hypothèse de CIA signifie que le biais de sélection peut être contrôlé s'il existe un ensemble de variables observables pour lesquelles une indépendance d'affectation au traitement peut être vérifiée (Brodaty, Crépon, & Fougère, 2007). Cette hypothèse est à la base des différentes méthodes d'appariement. Or, ces méthodes ne présentent d'intérêt que si l'on dispose de suffisamment de variables de conditionnement pour rendre compte de l'hétérogénéité des individus. Cependant, conditionner sur un ensemble étendu de covariables pose un problème de dimensionnalité, car il est difficile de trouver deux individus (traité et non traité) comparables, dès lors que le nombre de variables de conditionnement est élevé. Afin de résoudre ce problème, Rosenbaum et Rubin (1983) proposent de ramener le nombre de variables de conditionnement à une seule et unique variable, qui serait un résumé univarié de l'ensemble des covariables. Cette variable est précisément le score de propension[3].

La seconde hypothèse, celle de la condition de support commun, se rapporte au support de la distribution du score de propension. Cette hypothèse permet de s'assurer que les individus avec un même ensemble de covariables peuvent être à la fois traités ou non traités, ou, autrement dit, que les individus de chaque groupe d'analyse se ressemblent suffisamment pour que la comparaison ait un sens. L'absence de support commun entraînerait ce que Rosenbaum et Rubin (1983) appellent un facteur de confusion structurelle (*structural confounding*) et interdirait toute conclusion quant à l'effet causal d'un traitement. Pour prévenir ce problème, le score de propension ne doit être utilisé que dans la zone de support commun, c'est-à-dire la zone commune de la distribution des scores de propension des groupes traité et témoin.

Intérêt du score de propension : études expérimentales et quasi expérimentales

Le score de propension peut être mobilisé pour la recherche et l'évaluation en éducation de deux manières, soit comme outil de détection et de correction du biais de sélection. Il sert à juger l'efficacité de la randomisation dans les expérimentations sociales et la comparabilité des groupes dans les études quasi expérimentales. Dans les études randomisées, les individus ont en moyenne une chance sur deux de faire partie du groupe témoin. La randomisation doit permettre, du moins en théorie, d'obtenir deux groupes identiques en moyenne, c'est-à-dire que les caractéristiques initiales des individus du groupe traité et du groupe témoin sont semblables, tant du point de vue des caractéristiques observables qu'inobservables.

Cependant, malgré la randomisation, certains facteurs peuvent introduire un biais de sélection (Berger, 2005; Berger & Exner, 1999), comme les préférences et l'autosélection des acteurs ou un nombre inadéquat de participants (Altman & Doré, 1990). En effet, contrairement au domaine clinique, où il est relativement aisé de s'assurer d'une affectation aléatoire du traitement[4], les caractéristiques inhérentes aux interventions dans le domaine de l'éducation, comme dans d'autres domaines sociaux, rendent la tâche délicate. Prenons l'exemple d'un programme destiné aux enseignants visant à accroître les performances scolaires de leurs élèves. Un évaluateur consciencieux peut tirer au sort des enseignants et les affecter à un groupe traité ou témoin. Cependant, il est possible que seuls les enseignants les plus motivés et les plus expérimentés acceptent de participer au programme expérimental, les autres se trouvant de facto dans le groupe témoin. Or, les élèves fréquentant la classe d'enseignants expérimentés sont susceptibles de présenter des performances scolaires supérieures aux élèves du groupe témoin, ce qui peut amener à surestimer les effets du programme. Les réalités du terrain et les décisions individuelles peuvent réduire l'efficacité des efforts initiaux de randomisation, si bien que, selon ses caractéristiques, chaque individu n'aura plus nécessairement une chance sur deux d'être traité. La distribution du score de propension fournit alors un critère de jugement de l'efficacité de la randomisation et de la nécessité de traiter le biais de sélection : lorsque la randomisation a permis de rendre les individus des groupes traité et témoin comparables sur un ensemble de caractéristiques, alors la répartition du score de propension

de chacun des deux groupes doit être similaire. À l'inverse, des valeurs de score de propension élevées pour les individus du groupe exposé et des valeurs faibles pour les individus du groupe témoin indiquent la présence d'un biais de recrutement qui doit être traité.

Bien que les recherches expérimentales, au sens contrôlé et randomisé, soient considérées comme la méthode de référence, leur application peut être impossible pour différentes raisons, notamment des raisons d'ordre éthique. Par exemple, il ne serait pas éthiquement acceptable de séparer un échantillon en deux groupes, puis de demander à un groupe d'élèves de commencer à consommer du cannabis pour mesurer les effets de cette consommation sur leurs capacités cognitives. Dans ce cas de figure, ce sont les individus consommant déjà régulièrement cette substance qui seront assignés au groupe traité (Harrison, Gruber, Hudson, Huestis, & Yurgelun-Todd, 2002).

Les études quasi expérimentales représentent une option intéressante lorsque la randomisation n'est pas possible, mais, aussi et surtout, parce qu'elles tendent à être davantage représentatives de ce qui se passe dans le monde réel. Une expérimentation randomisée étant contrôlée, les individus y agissent donc dans un environnement qui n'est pas totalement naturel, ce qui peut les amener à modifier leurs comportements d'une manière différente de ce qu'ils feraient sans un tel contrôle.

Avec un devis quasi expérimental, les individus présentent vraisemblablement des différences au stade initial de l'intervention (Grimes & Schulz, 2002; Steiner, Cook, Shadish, & Clark, 2010). Il est pour ainsi dire naturel de ne pas obtenir une équirépartition des chances de traitement, les individus n'ayant pas été affectés aléatoirement au groupe traité ou témoin. Cela est d'autant plus vrai si la participation au traitement est volontaire, par exemple avec les programmes de soutien scolaire à domicile.

La distribution du score de propension offre un critère de comparabilité des groupes dans les études quasi expérimentales. La comparaison de la distribution des scores dans chacun des groupes permet de s'assurer que les individus traités et témoins sont suffisamment semblables pour que la comparaison ait du sens. C'est seulement dans ce cas que le score de propension peut être utilisé pour ensuite corriger le biais de sélection.

En synthèse, le score de propension permet d'abord de détecter l'existence de différences préexistantes à la mise en œuvre de l'intervention. En cas de différences, il permet ensuite de corriger le biais de sélection et de

déterminer les effets nets d'un traitement sous réserve des hypothèses de CIA et de support commun (Winship & Mare, 1992). Il nécessite néanmoins de répondre à plusieurs questions d'ordre méthodologique, questions qui vont maintenant être présentées.

Estimer le score de propension : construction et évaluation de la qualité

L'effet du traitement se déduit des différences entre les groupes (témoin et expérimental) et les périodes d'évaluation (prétest et post-test). Pour être en mesure de déterminer les effets du traitement, il est donc primordial que les deux groupes soient les plus comparables possible avant la mise en place du traitement expérimental. C'est à cette période que le score de propension est estimé.

Le score de propension s'estime en plusieurs étapes, étapes qui seront décrites pas à pas. La démarche n'est pas purement linéaire et des allers-retours entre ces étapes s'avèrent souvent nécessaires. Dans un premier temps, il est essentiel de déterminer les facteurs expliquant l'affectation au traitement. Le deuxième temps est consacré à l'estimation du score de propension. Dans un troisième temps, il faut évaluer la qualité du score de propension estimé et, dans un quatrième temps, déterminer la zone de support commun. Ce n'est qu'une fois toutes ces étapes effectuées que le score de propension peut être utilisé pour mesurer l'effet du traitement.

L'affectation au traitement

Différentes méthodes classiques peuvent être utilisées pour décrire l'affectation au traitement, telles que des tests de comparaison des moyennes (test de Student ou analyse de la variance [ANOVA]), des tests de comparaison de la distribution (khi deux) ou une série de régressions logistiques univariées et multivariées. L'objectif de cette étape préalable est d'identifier les variables qui pourront potentiellement être intégrées dans la construction du score.

Bien évidemment, si aucune différence ne préexiste, il n'est pas utile de recourir au score de propension[5]. De même, si les différences initiales sont minimes et sur un nombre très restreint de variables, ou sur une ou des variables jugées non pertinentes dans l'analyse, un jugement avisé sur les avantages et les inconvénients devra précéder la décision de recourir au score de propension.

L'estimation du score de propension

L'estimation du score de propension se fait classiquement par l'intermédiaire d'une régression logistique[6], le score de propension étant une probabilité. À ce stade, le choix des variables à introduire dans le modèle est crucial, car, selon les variables choisies pour l'estimation, le score de propension estimé peut être très différent.

Plusieurs stratégies peuvent être adoptées pour choisir les variables à intégrer dans l'estimation du score et il n'existe pas de guide clair sur la façon de procéder. De manière générale, il est nécessaire d'introduire dans le modèle toutes les variables qui ont un impact significatif à la fois sur les probabilités d'appartenir au groupe traité et sur les variables de résultat pour satisfaire de manière crédible à l'hypothèse de CIA (Caliendo & Kopeinig, 2008; Smith & Todd, 2005). Les résultats obtenus dans la première étape servent de guide, puisque toutes les variables ayant un effet significatif sur la probabilité d'être traité y auront été identifiées.

Ce choix des variables pour l'estimation du score de propension représente la première difficulté et amène des arbitrages. Plus le nombre de variables introduites pour estimer le score de propension sera élevé, meilleure sera la description de la probabilité de traitement. Cependant, les scores des individus traités et non traités risquent de se dissocier, ce qui réduit la zone de support commun (Augurzky & Schmidt, 2001; Bryson, Dorsett, & Purdon, 2002). L'objectif est avant tout d'obtenir un score de propension qui permet d'équilibrer les groupes. En outre, ce choix implique un second arbitrage entre consistance et efficience : omettre des variables importantes peut introduire un biais dans l'estimation des effets du traitement (Heckman, Ichimura, & Todd, 1997), tandis que surcontrôler peut accroître la variance des estimateurs (Bryson et al., 2002).

Choisir les variables pour le score de propension n'est pas chose facile et la littérature fournit peu d'indications à ce sujet. Plusieurs auteurs proposent différentes méthodes, mais sans véritable consensus (Caliendo & Kopeinig, 2008; Dehejia & Wahba, 2002). Trois éléments peuvent néanmoins guider le chercheur. Tout d'abord, seules les variables qui influencent à la fois la décision de participation et la variable d'intérêt doivent être incluses. Ensuite, il est nécessaire d'inclure uniquement les variables qui ne sont pas affectées par la participation ; pour ce faire, les variables doivent être mesurées avant la participation au programme. Enfin, les données pour les participants et les non-participants doivent provenir des

mêmes sources (par ex., le même questionnaire). Aussi, cette étape nécessite de nombreuses itérations et, bien souvent, il faut estimer une dizaine de scores de propension, voire plus (en utilisant des variables différentes, en enlevant une variable, en en ajoutant une autre, en créant des termes d'interaction), avant de trouver un score de propension qui satisfait à la propriété d'équilibre. À cette étape décisive, les analyses préalables et la connaissance des théories et des résultats d'études empiriques de la discipline permettent de guider le choix des variables à inclure dans l'estimation du score de propension.

L'évaluation de la qualité du score de propension

Une fois le score de propension estimé, il est nécessaire d'évaluer sa qualité. Un bon score de propension est un outil d'équilibrage : il doit permettre d'équilibrer la distribution des variables choisies dans les deux groupes. Pour chacune des variables intégrées dans la construction du score, il faut comparer leur distribution par strate de scores de propension. Des exemples de stratification sur scores de propension sont proposés par Rosenbaum et Rubin (1983, 1984), Rubin (1997) et Dehejia et Wahba (1999).

Avec un score de propension de qualité, les différences significatives entre les groupes traité et témoin à la période initiale ne devraient plus subsister, et ce, pour chacune des variables utilisées dans la construction du score. Des tests classiques de comparaison de deux distributions sont utilisés pour évaluer la significativité des différences. Si des différences sont détectées, il faut d'abord examiner le nombre de variables et les strates pour lesquelles celles-ci persistent. Lorsque seul un petit nombre de variables est concerné et que, pour celles-ci, seules quelques strates sont touchées, la stratégie consiste à retenir une sous-classification de la strate et à vérifier si la différence persiste[7] (Rosenbaum & Rubin, 1984). Si elle persiste ou si le nombre de variables concernées par des différences est important, il est nécessaire de refaire une estimation du score de propension en ajoutant une ou plusieurs variables, en retirant des variables dont la significativité disparaît, en créant des termes d'interaction ou en créant des termes d'ordre supérieur pour les variables. Ces étapes sont répétées jusqu'à l'obtention d'un score de propension équilibré ou, à défaut, qui minimise le déséquilibre avec les données disponibles.

Le support commun

Le support commun est la zone de superposition des deux groupes sur l'ensemble des valeurs du score de propension (Heckman, LaLonde, & Smith, 1999). Le support commun du score de propension permet de s'assurer qu'il est possible, pour chaque individu du groupe traité, de trouver au moins un participant du groupe témoin ayant les mêmes caractéristiques (score de propension) initiales (Bryson et al., 2002). L'utilisation du score de propension n'est adéquate que pour les individus situés dans cette zone. Il est possible de déterminer graphiquement la zone de support commun par une analyse visuelle de la distribution du score de propension des deux groupes (Lechner, 2002). Il n'est pas nécessaire de mettre en place des stratégies complexes pour déterminer la zone de support commun. Néanmoins, deux principales méthodes existent : la première se base sur la comparaison des minima et maxima des scores de propension des deux groupes et la seconde, sur la comparaison de la distribution (*trimming*).

La méthode du minimum-maximum (Dehejia & Wahba, 1999) consiste à conserver l'ensemble des individus traités et non traités, à l'exception des individus pour lesquels il n'existe pas de contrefactuel, c'est-à-dire les individus dont le score de propension est inférieur au minimum ou supérieur au maximum du score des individus de l'autre groupe.

La méthode de comparaison de la distribution (*trimming*) de Smith et Todd (2005) consiste à exclure les individus non traités pour lesquels la proportion de contrefactuels potentiels, c'est-à-dire les individus traités dont le score de propension est très proche de celui des individus non traités considérés, est la plus faible.

Le choix de l'une ou l'autre méthode et celui des seuils de sensibilité dans la méthode de comparaison de la distribution sont à la discrétion de l'analyste. Il est néanmoins nécessaire de mentionner que, lorsque la répartition du score de propension tend vers la symétrie, la méthode de comparaison de la distribution permet de maximiser la zone de support commun en excluant peu d'individus[8].

Évaluer les effets d'un traitement avec le score de propension

Le score de propension désormais estimé, sa qualité confirmée et la zone de support commun déterminée, il est possible de commencer l'étape qui motive initialement le recours à ce score, soit l'évaluation de l'im-

pact du traitement. Trois méthodes seront présentées. Les deux premières sont des applications directes du score : l'appariement sur score de propension (Rosenbaum & Rubin, 1983, 1985a, 1985b) et la pondération inverse sur les probabilités d'être traité (Hirano, Imbens, & Ridder, 2003). La dernière méthode consiste à combiner le score de propension avec une analyse en doubles différences (Heckman et al., 1997).

L'appariement

L'appariement sur score de propension, connu sous le terme anglais *propensity score matching* (PSM), réfère à l'appariement d'individus des groupes traité et de contrôle possédant des valeurs de score de propension proches ou similaires, et écarte les individus non appariés. Les méthodes d'appariement tentent de jumeler chaque individu traité avec un ou plusieurs individus non traités dont les caractéristiques observables sont les plus proches possible. L'objectif de l'appariement est de construire un groupe témoin comparable au groupe traité afin de permettre une estimation non biaisée de l'effet du traitement sur les individus traités, en contrôlant le biais de sélection (Abadie & Imbens, 2005; Caliendo & Kopeinig, 2008; Dehejia, 2005; Dehejia & Wahba, 2002; Imbens, 2004; Smith & Todd, 2001). Il existe différents estimateurs d'appariement, dont les principaux sont présentés ci-dessous.

Le plus proche voisin. Il s'agit de la méthode d'appariement la plus utilisée. Un participant du groupe traité est apparié avec un participant du groupe témoin sur la base du plus proche score de propension. L'appariement peut être réalisé avec ou sans remise. Dans la méthode sans remise, les individus non traités ne sont utilisés qu'une seule fois, un individu traité étant apparié avec un seul individu non traité. Dans la méthode avec remise, les individus peuvent être utilisés plus d'une fois. Cette technique est privilégiée lorsque la distribution du score de propension entre les deux groupes est très différente (Smith & Todd, 2005). L'appariement sur le plus proche voisin n'est toutefois pas toujours performant, notamment lorsque le voisin le plus proche se trouve à une distance éloignée de l'individu traité à apparier. Des individus très différents peuvent alors être jumelés. D'un point de vue pratique, il convient d'ordonner aléatoirement les données avant d'apparier, puisque le pairage s'effectue par ordre.

La stratification. Pour éviter de jumeler deux individus trop distants, une seconde technique d'appariement peut être utilisée, soit la stratification, aussi connue sous le terme anglais *subclassification* (Rosenbaum & Rubin, 1983). Avec cette méthode, les scores de propension sont classés par intervalles, nommés les strates, et les individus traités et non traités sont ensuite appariés au sein de strates. Le nombre de strates peut être librement choisi, mais Cochran et Chambers (1965) puis Imbens (2004) montrent que l'utilisation de cinq strates suffit à contrôler 95% du biais. Cette méthode comporte néanmoins un inconvénient : l'utilisation de strates n'est pas toujours adéquate, car les individus les plus proches des points de césure pourraient être classés dans une strate voisine. Ainsi, certains individus ne sont pas appariés alors que la distance les séparant est faible, parce qu'ils se situent dans une strate différente. Généralement, cette méthode génère plus d'individus non appariés que la méthode du plus proche voisin. Il en résulte une diminution de la taille de l'échantillon, une perte de puissance statistique et un risque accru d'erreur de type II. Cependant, l'appariement par stratification est souvent de meilleure qualité que celui par plus proche voisin, puisque les individus jumelés se ressemblent plus.

Le caliper. Un participant du groupe témoin est apparié avec un participant du groupe traité sur la base du plus proche score de propension, sous réserve d'une certaine distance maximale, le *caliper*. Les individus traités pour lesquels le plus proche voisin non traité n'appartient pas à la région définie sont exclus de l'analyse. Le *caliper* est à la discrétion de l'analyste ; il n'existe donc pas de méthode permettant de déterminer le niveau « raisonnable » de tolérance à choisir. Plus le *caliper* sera petit, plus les individus appariés seront semblables, mais plus le nombre d'individus non appariés augmentera. Il est généralement recommandé de tester différentes versions du *caliper* et de mener des analyses de sensibilité selon les différentes tailles de *caliper*.

Plusieurs auteurs ont comparé les méthodes d'appariement et les effets de celles-ci sur la réduction du biais de sélection (Baser, 2006 ; Caliendo & Kopeinig, 2008). Ces études tendent à montrer que l'appariement avec *caliper*, plutôt que sans, permet une meilleure réduction du biais de sélection.

La qualité de l'appariement

La dernière étape consiste à tester la réduction du biais de sélection à la suite de l'appariement[9]. Ce test s'effectue avec des analyses descriptives, des comparaisons des moyennes (Rosenbaum & Rubin, 1985a, 1985b), des comparaisons de distribution ou des tests de stratification (Dehejia & Wahba, 1999, 2002). Il suffit de comparer les individus selon les groupes, avant et après appariement, sur plusieurs variables. Si l'appariement est de qualité, alors les différences initiales entre les individus des groupes traité et témoin ne devraient plus être présentes après appariement sur score de propension. Cependant, il est possible que ce ne soit pas le cas. Il convient alors de réaliser un nouvel appariement en changeant de méthode (plus proche voisin, stratification ou *caliper*) ou de seuil (nombre de strates, taille du *caliper*) jusqu'à ce qu'il n'y ait plus de différence.

Une fois la qualité d'appariement avérée, les individus non appariés peuvent être exclus de la base de données pour créer un échantillon dans lequel les différences initiales entre individus traités et non traités sont réduites, voire éliminées. Les effets du traitement sont ensuite estimés sur cette nouvelle base avec des méthodes traditionnelles (ANOVA ou régression).

La pondération inverse sur les probabilités d'être traité

La deuxième méthode reposant directement sur le score de propension est la pondération inverse sur les probabilités d'être traité, plus connue sous le terme anglais d'*inverse probability of treatment weighting* (IPTW) et notamment popularisée en économie par Hirano et al. (2003). Cette méthode est principalement utilisée dans le champ de l'épidémiologie dans le cadre de modèles structurels marginaux (Robins, Hernán, & Brumback, 2000) et est, dans l'ensemble, moins utilisée que l'appariement sur score de propension.

Avec l'IPTW, chaque individu reçoit une pondération inverse de la probabilité d'avoir reçu le traitement, probabilité ici estimée par le score de propension. Plus spécifiquement, pour les individus traités, la pondération est l'inverse du score de propension, c'est-à-dire $(PS)^{-1}$, tandis que, pour les individus du groupe témoin, la pondération est $(1\text{-}PS)^{-1}$.

Le principe est similaire à la pondération sur données d'enquête. Pour ce type de données, les poids sont utilisés pour rendre les observations issues de l'échantillon enquêté similaires à la population dont elle est issue, en appliquant par exemple un poids plus élevé aux femmes si ces dernières sont sous-représentées dans l'enquête.

La pondération inverse réduit le poids de ceux qui avaient de fortes chances de recevoir le traitement actuellement reçu d'après leurs caractéristiques observables, et augmente le poids de ceux qui avaient peu de chance de recevoir le traitement effectivement reçu, toujours selon ces caractéristiques observables. Ce faisant, les groupes traité et témoin sont rendus comparables, car ils auraient eu les mêmes chances d'être traités si le traitement n'avait pas été effectivement réalisé.

La méthode de l'IPTW implique donc de créer une nouvelle variable, qui est ensuite intégrée dans les analyses traditionnelles, par exemple dans des régressions. Cette méthode présente l'avantage, par rapport à l'appariement, de conserver l'ensemble de l'échantillon pour l'analyse, sous réserve de respect du support commun, là où l'appariement conduit à exclure les individus qui ne peuvent être jumelés. C'est en contexte dynamique que cette méthode présente ses principales limites, notamment lorsque les individus peuvent influencer le traitement au cours du temps (Bjerk, 2009). De plus, la méthode de l'IPTW ne doit pas être utilisée si, à chaque période de temps, il existe une covariable telle que, pour une certaine valeur de cette covariable, tous les individus sont systématiquement traités ou non traités[10] (Robins et al., 2000).

L'analyse en doubles différences sur score de propension

Cette dernière technique consiste à combiner le score de propension avec une autre méthode, l'analyse en doubles différences. Il s'agit d'une méthode d'analyse de panel principalement utilisée en économie et qui trouve désormais des applications dans le champ de l'éducation. Bénabou, Kramarz et Prost (2004) l'utilisent notamment pour évaluer l'efficacité de la politique des ZEP adoptée en France en 1981, là où une simple comparaison des élèves en ZEP et hors ZEP après l'introduction de la politique n'aurait pas permis de contrôler le biais de sélection. Les établissements classés ZEP scolarisent en effet une proportion plus élevée d'élèves ayant des difficultés d'ordre scolaire et sociale que les autres établissements scolaires.

L'utilisation de la méthode des doubles différences nécessite que les individus soient observés sur deux temps de mesure (avant et après le traitement) et soient répartis en deux groupes (témoin et traité). Ce type de devis d'études étant celui considéré dans cet article[11], il est intéressant d'aborder cette méthode. Les effets d'un traitement expérimental peuvent être estimés à partir d'une double différence. La première correspond aux différences temporelles (prétraitement et post-traitement) pour chacun des groupes (traité ou témoin). La seconde est la différence de ces différences. Le modèle de doubles différences (Imbens & Wooldridge, 2009) s'estime de la manière suivante :

$$\alpha_{did} = \left(\overline{Y_{t1}^{T}} - \overline{Y_{t0}^{T}}\right) - \left(\overline{Y_{t1}^{C}} - \overline{Y_{t0}^{C}}\right)$$

où :

$_t$ = le temps de mesure, avec $_{t0}$ pour les mesures initiales et $_{t1}$ pour les mesures en post-traitement

$\bar{Y}$ = les moyennes de la variable d'intérêt selon les groupes, ^{T} pour le groupe traité et ^{C} pour le groupe de contrôle.

L'estimateur fournit une mesure nette du changement de la variable d'intérêt induit par l'appartenance au groupe. Le changement n'est pas uniquement le fruit du traitement. En effet, de nombreux facteurs autres que l'intervention pourraient avoir causé ou accentué le changement observé dans la variable d'intérêt. Or, cette dynamique naturelle pourrait affecter également le groupe témoin. La méthode des doubles différences repose ainsi sur l'hypothèse fondamentale de tendance temporelle commune entre les deux groupes, selon laquelle en l'absence de traitement les changements observés dans les groupes auraient été identiques. La différence restante est alors attribuable à l'impact du traitement sur la variable d'intérêt.

L'hypothèse de tendance commune est moins restrictive que les conditions d'indépendance conditionnelle à des caractéristiques observables et de support commun, qui doivent être toutes deux respectées pour estimer et utiliser le score de propension. L'estimation des effets d'un traitement par doubles différences représente donc une option intéressante à l'utilisation du score de propension, notamment lorsque l'utilisation de ce dernier n'est pas appropriée (petit échantillon, score de propension non équilibré, zone de support commun restreinte ou appariement ne permettant pas de réduire les différences initiales).

C'est néanmoins davantage dans l'esprit d'une complémentarité que d'opposition que cette méthode est présentée ici. En effet, le score de propension permet de contrôler le biais de sélection sur les facteurs observables, tandis que les doubles différences permettent de contrôler ce biais sur les facteurs inobservables, dès lors que l'influence des caractéristiques inobservables sur la variable d'intérêt est considérée comme étant constante dans le temps. Ainsi, l'utilisation combinée de ces deux méthodes permet une meilleure correction du biais de sélection, et l'estimation obtenue de l'effet du traitement sera plus sûrement encore une mesure de causalité. La combinaison peut reposer sur les différentes utilisations du score de propension, soit l'appariement et la pondération inverse, selon la méthode la plus adaptée aux données. Dans le premier cas, l'estimateur de différences est calculé sur les données appariées, tandis qu'il l'est sur des observations pondérées par l'inverse des probabilités d'être traité dans le second cas.

Si la combinaison du score de propension avec la méthode des doubles différences semble à priori permettre une estimation plus nette de l'impact causal, elle se fait au coût d'une augmentation des hypothèses à respecter. L'hypothèse centrale de tendance commune en doubles différences ne peut être testée directement, mais son respect peut être approximé dès lors que plusieurs temps de mesure en prépériode (au moins deux) sont disponibles. Une simple représentation graphique ou des statistiques descriptives relatives à la variable d'intérêt avant le traitement permettent de visualiser la vraisemblance de la tendance commune aux deux groupes. Sans la possibilité d'évaluer si la tendance commune est raisonnable, la combinaison des deux méthodes revient à imposer des restrictions supplémentaires. C'est donc là la principale limite d'une telle combinaison, puisqu'obtenir des données longitudinales avec plusieurs temps de mesure en prépériode est souvent difficile.

Application : impact d'une expérimentation musicale sur les capacités cognitives d'élèves de maternelle

L'objectif de cette section est d'illustrer les différentes étapes de la méthode afin d'en faciliter la compréhension et l'application future. Les différentes étapes et résultats y seront détaillés. Les analyses sont réalisées à l'aide des logiciels Stata et SPSS. Les commandes correspondant à l'utilisation du score de propension seront présentées[12].

Les données

Les données sont issues d'une expérimentation musicale ayant pour objectif d'accroître les capacités cognitives chez des enfants de 5 ans. Cette étude a été menée dans une région française. Il s'agit d'une étude à devis expérimental, c'est-à-dire que les élèves ont été tirés au sort pour faire partie de l'expérimentation. De la même manière, l'affection au groupe traité est randomisée. L'échantillon se compose de 480 élèves, avec 254 élèves dans le groupe témoin et 226 dans le groupe expérimental. De nombreuses capacités cognitives ont été mesurées, mais, à des fins d'exemple, seuls les effets du traitement sur la discrimination visuelle seront décrits. Pour une présentation plus approfondie de l'expérimentation ayant permis d'obtenir ces données, le lecteur peut consulter Lecocq (2012).

Le tableau 1 décrit brièvement l'échantillon. Il se compose à parts sensiblement égales de filles et de garçons : 50,4 % des élèves sont des filles contre 49,6 % de garçons. La répartition par trimestre de naissance est plutôt homogène. Les caractéristiques professionnelles des parents ayant un impact avéré sur les performances scolaires (Lecocq, 2012), elles ont été recueillies. Un trait marquant est la forte proportion de pères ouvriers et de mères inactives ou au chômage dans l'échantillon. En effet, 39,2 % des élèves ont un père ouvrier (qualifié, non qualifié et agricole confondus) et ils sont près de quatre fois moins nombreux à être enfants de cadre. La proportion d'élèves dont le père est inactif ou chômeur est de 5,6 %, proportion s'élevant à 34,7 % pour les mères. La répartition des emplois dans la population française diffère peu de celle des parents d'élèves de l'échantillon (Institut national de la statistique et des études économiques), exception faite de la présence massive des mères au foyer et des pères ouvriers dans l'échantillon, qui peut être expliquée par la nature défavorisée du terrain de l'expérimentation. Il est essentiel de préciser que près d'un tiers des établissements de l'expérimentation sont classés en éducation prioritaire. Les deux tiers restants sont situés en majorité dans des quartiers défavorisés.

Tableau 1
Caractéristiques des élèves de l'échantillon

Caractéristiques de l'élève N (%)		Total (N=480)	Témoin (N=254)	Musique (N=226)
Sexe:	Garçon	238 (49,6)	129 (50)	115 (51,9)
	Fille	242 (50,4)	129 (50)	111 (49,1)
Pays de naissance:	France	430 (94,9)	219 (94,8)	210 (95)
	Etranger	23 (5,1)	12 (5,2)	11 (5)
Trimestre de naissance:	Premier	116 (25.2)	58 (24,5)	58 (25,9)
	Second	116 (25,2)	69 (29,2)	47 (20,9)
	Troisième	110 (23,9)	65 (27,4)	45 (20,1)
	Quatrième	119 (25,7)	45 (18,9)	74 (33,1)
Profession des parents	Père ouvrier	167	63	104
	Mère ouvrière	45	13	32
	Mère cadre	114	77	37
D. visuelle	Pré-test	73,3 (23,6)	76,1 (21,7)	70,1 (25,3)
Moyenne (écart-type)	Post-test	81,4 (20,4)	83,7 (18,7)	78,8 (21,9)

La nécessité de traiter le biais de sélection : l'évaluation de la comparabilité des groupes

La répartition des élèves au sein des deux groupes témoin et traité s'est effectuée de manière aléatoire. La randomisation devait permettre, en principe, d'obtenir deux groupes comparables, tant du point de vue des caractéristiques personnelles que des scores au test initial. Cependant, malgré la randomisation, les élèves des deux groupes ne possèdent pas des caractéristiques initiales similaires. Des analyses préliminaires, soit des tests de comparaison des moyennes, ont montré que les élèves du groupe témoin obtiennent de meilleures performances initiales au test standardisé de discrimination visuelle : les élèves du groupe témoin obtiennent 101,8 et ceux du groupe traité 97,8 ($F = 6,69$; $P = 0,01$). De plus, les résultats de la régression logistique binaire montrent que les probabilités d'appartenir au groupe traité dépendent d'un certain nombre de caractéristiques des élèves, comme le trimestre de naissance et la profession des parents. Cette analyse statistique permet d'obtenir la régression de la variable dichotomique dépendante, ici l'appartenance au groupe *musique*, en fonction de l'ensemble des variables socioéconomiques et scolaires suivantes : le genre de l'élève, son pays de naissance, la langue parlée à la maison, son tri-

mestre de naissance et la profession de ses parents. Dans ce modèle, le R^2 de Nagelkerke explique à hauteur de 23,5 % la variance de la variable dépendante. Le « pourcentage global » indique que le modèle classe correctement les individus dans 68 % des cas.

La plupart des variables introduites dans le modèle n'exercent pas une influence statistiquement significative sur la probabilité d'appartenir ou non au groupe expérimental : le genre, le pays de naissance et la langue parlée à la maison n'affectent pas les « chances » d'appartenir au groupe musique.

Cependant, la régression montre que le tirage aléatoire a classé moins d'élèves des second et troisième trimestres ($B = -0{,}78$ et $-0{,}93$; $p < 0{,}01$) dans le groupe expérimental, le rapport de cote étant respectivement de 0,46 et 0,4. Concernant la profession du père, il apparaît que les élèves dont le père est ouvrier ($B = 0{,}75$; $p < 0{,}01$) ont plus de probabilité d'appartenir au groupe expérimental, le rapport de cote étant de 2,1, équivalant à une augmentation de 110 % de chance par rapport à un élève dont le père est sans activité. Les élèves dont la mère appartient aux catégories cadre ou profession intermédiaire ont une probabilité plus faible d'appartenir au groupe expérimental, contrairement aux élèves dont la mère est ouvrière, qui ont deux fois plus de chance d'appartenir au groupe expérimental.

Ainsi, selon leur trimestre de naissance et la profession de leurs parents, les élèves de l'échantillon n'ont pas tous la même probabilité d'appartenir, ou non, au groupe expérimental. Les élèves du groupe expérimental sont, en moyenne, plus jeunes, mais ils sont également plus nombreux à être enfants d'ouvrier et moins nombreux à être enfants de cadre ou de profession intermédiaire. Or, ces deux facteurs sont connus dans la littérature pour avoir un effet sur les performances aux tests de capacités cognitives (Lecocq, 2012). Le score de propension permettra d'estimer des effets non biaisés du traitement sur les capacités de discrimination visuelle.

L'estimation du score et la vérification de sa qualité

Le score de propension est estimé avec la commande *pscore* du logiciel Stata (Becker & Ichino, 2002). Afin d'estimer le score de propension, plusieurs spécifications ont été testées. Au final, les variables ayant un impact sur les capacités cognitives et dont les valeurs apparaissent différentes en prépériode selon le groupe d'expérimentation sont introduites dans le modèle. Le modèle final intègre le genre, le trimestre de naissance, la profession des parents et le score initial à différents tests cognitifs. Le score de

propension est estimé par une régression logistique (logit). Lorsque la variable expliquée est binaire, comme c'est le cas ici (individu traité = 1 ou non traité = 0), l'estimation peut être réalisée indifféremment par un modèle logit ou probit. L'équilibre du score de propension est vérifié avec une stratification en cinq blocs. Ce faisant, l'équilibre est examiné par quintile de distribution du score de propension. Ce nombre de strates est généralement le minimum pour lequel un équilibre du score doit être trouvé. Enfin, les analyses sont limitées au support commun (*comsup*). Il est préférable d'appliquer cette restriction dès à présent, puisqu'il s'agit de la condition sine qua non pour éviter un biais de confusion structurelle lors de l'estimation des effets du traitement avec le score de propension.

Deux éléments permettent de s'assurer de la qualité du score : une zone de support commun étendue et un score équilibré. La zone de support commun est assez étendue ici, signe d'une bonne comparabilité des groupes musique et témoin, puisque comprise entre 0,1088 et 0,8222. Au deçà et au-delà de ces limites, il n'y a pas de contrefactuels, c'est-à-dire pas d'élèves du groupe opposé ayant un score de propension identique. Le score de propension a été estimé pour tous les élèves, mais les élèves ne figurant pas dans la zone de support commun sont exclus des analyses (6 élèves du groupe témoin exclus) pour un total de 474 élèves. Le score de propension est équilibré, c'est-à-dire que les scores de propension des élèves du groupe témoin et traité sont comparables, pour chaque variable introduite dans le modèle et pour chaque strate. Si cela n'avait pas été le cas, une première étape aurait été d'examiner l'équilibre du score sur de plus petites strates (par ex., d'abord 10, puis 20 strates). En l'absence d'équilibre malgré cette étape supplémentaire, il aurait fallu estimer un nouveau score de propension avec une spécification différente.

L'appariement

La zone de support commun étendue a incité à retenir l'appariement. La pondération par l'inverse de probabilité de traitement aurait également pu être envisagée, mais, dans la mesure où très peu d'individus étaient exclus par l'appariement, cette dernière méthode, plus directe, a été privilégiée. L'appariement sur score de propension est effectué à l'aide de la commande *psmatch2* (Leuven & Sianesi, 2012). La méthode retenue est l'appariement par plus proche voisin, en spécifiant d'apparier les élèves

dont le score de propension se situe dans la zone de support commun. Cette méthode d'appariement, la plus simple, est celle avec laquelle débuter. Si la qualité de l'appariement n'avait pas été satisfaisante, comme il le sera expliqué dans la sous-section suivante, les méthodes d'appariement par stratification et par *caliper* auraient été successivement utilisées. La figure ci-dessous représente la répartition du score de propension selon les groupes, avant et après l'appariement par plus proche voisin.

Avant l'appariement, les élèves du groupe témoin sont nombreux à avoir un score de propension inférieur à 0,5. À l'inverse, les élèves du groupe traité ont majoritairement un score de propension supérieur à 0,5. Le score de propension étant une probabilité d'appartenir au groupe traité compte tenu des caractéristiques choisies, les élèves ayant un score de propension proche de 1 ont bien entendu plus de chance d'appartenir au groupe traité. Cette figure illustre la présence d'un biais de sélection, puisque les élèves du groupe traité sont plus nombreux que les élèves du groupe témoin à avoir un score de propension élevé, tandis que les élèves du groupe témoin sont plus nombreux à avoir un score de propension faible. La seconde figure représente la répartition du score de propension après l'appariement. La répartition est symétrique, ce qui indique que l'appariement a rendu les élèves des deux groupes comparables.

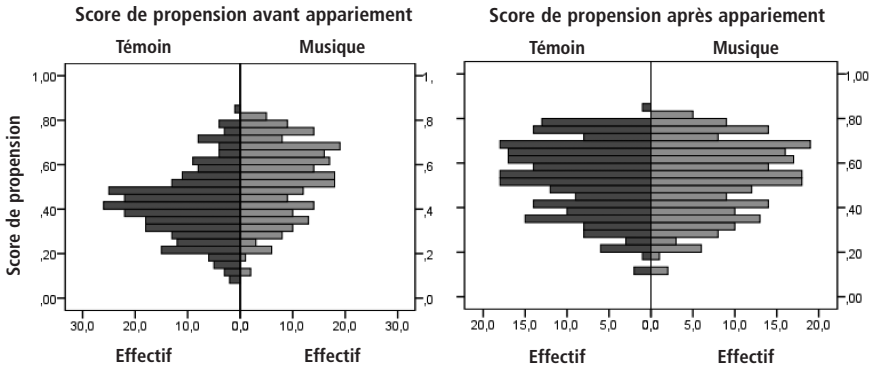

Figure 1. ***Répartition des scores de propension selon les groupes, avant et après appariement sur score de propension***

La qualité de l'appariement : significativité des différences et test de la réduction du biais

L'objectif est de vérifier que l'appariement réduit les biais initialement observés et, donc, de corroborer le diagnostic graphique précédent. La qualité de l'appariement est évaluée par des tests de comparaison des moyennes, ou de pourcentage selon la nature des variables, ainsi que par la réduction du biais. Il s'agit de comparer les moyennes (ou pourcentages) entre les groupes traité et témoin, avant et après l'appariement pour s'assurer que les différences initialement significatives ne le sont plus. La réduction du biais s'obtient de la manière suivante : le biais est d'abord calculé sur l'échantillon non apparié en divisant la différence des moyennes entre individus traités et non traités par l'écart-type commun de l'échantillon[13] ; le biais est de nouveau calculé après appariement. Au numérateur figure la différence des moyennes des individus traités et témoins dans l'échantillon apparié, alors que c'est l'écart-type commun de l'échantillon non apparié qui est utilisé au dénominateur. Enfin, la réduction du biais est obtenue par la différence entre ces deux biais. L'appariement est jugé de qualité s'il permet de réduire les différences initiales sur au moins un des deux critères (significativité des différences et réduction du biais).

Le tableau 2 présente les résultats issus de la commande *pstest* du logiciel Stata. Il montre par exemple que, avant appariement, 46 % des élèves du groupe traité ont un père ouvrier, contre 25 % pour les élèves du groupe témoin. Cette différence est significative ($p > 0{,}000$). Après l'appariement, 46 % des élèves du groupe traité et 46,5 % des élèves du groupe témoin ont un père ouvrier. La différence n'est plus significative. Le biais est réduit de 97 % (il aurait été de 100 % si les deux groupes avaient un pourcentage égal après l'appariement). Au total, le biais est réduit sur toutes les variables considérées, à l'exception du genre, et, pour ces variables, il ne persiste aucune différence significative entre les deux groupes après appariement. Ainsi, la qualité de l'appariement par plus proche voisin sur score de propension est donc satisfaisante pour ces données.

Tableau 2
Test de la réduction des biais

Variable	Sample	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|
os0	Unmatched	97.102	102.51	-36.5		-4.00	0.000
	Matched	97.102	95.636	9.9	72.9	0.99	0.325
rythme0	Unmatched	97.134	102.56	-36.6		-4.01	0.000
	Matched	97.134	95.154	13.4	63.5	1.37	0.172
graph0	Unmatched	98.457	101.31	-19.1		-2.08	0.038
	Matched	98.457	96.89	10.5	45.0	1.16	0.245
memoire0	Unmatched	99.372	100.44	-7.1		-0.78	0.438
	Matched	99.372	99.59	-1.5	79.6	-0.15	0.880
fille	Unmatched	.50885	.5	1.8		0.19	0.847
	Matched	.50885	.42035	17.7	-900.0	1.89	0.059
premiertri~e	Unmatched	.25664	.23016	6.2		0.67	0.501
	Matched	.25664	.25221	1.0	83.3	0.11	0.914
secondtrim~e	Unmatched	.28319	.35317	-15.0		-1.64	0.102
	Matched	.28319	.32743	-9.5	36.8	-1.02	0.308
troisiemet~e	Unmatched	.19912	.25794	-14.0		-1.53	0.128
	Matched	.19912	.21681	-4.2	69.9	-0.46	0.644
activitemu~e	Unmatched	.01327	.05952	-24.8		-2.67	0.008
	Matched	.01327	.01327	0.0	100.0	0.00	1.000
pouvrier	Unmatched	.46018	.25	44.9		4.92	0.000
	Matched	.46018	.4646	-0.9	97.9	-0.09	0.925
pcadre	Unmatched	.07522	.10714	-11.1		-1.20	0.229
	Matched	.07522	.07965	-1.5	86.1	-0.18	0.861
mcadreinter	Unmatched	.16372	.30556	-33.9		-3.68	0.000
	Matched	.16372	.19912	-8.5	75.0	-0.98	0.330

L'évaluation des effets du traitement

En l'absence de plusieurs temps de mesure en prépériode, la méthode de doubles différences n'a pas été retenue. Les effets du traitement peuvent alors être évalués indifféremment par des régressions ou des analyses de la variance (ANOVA) à mesures répétées sur l'échantillon apparié. Pour des fins d'illustration de la méthode, les effets de l'entraînement musical sur le score de discrimination visuelle sont estimés à l'aide d'ANOVA à

mesures répétées effectuées à l'aide du logiciel SPSS. La figure suivante présente les résultats de ces analyses effectuées sur la base de données originales, puis sur l'échantillon apparié sur score de propension.

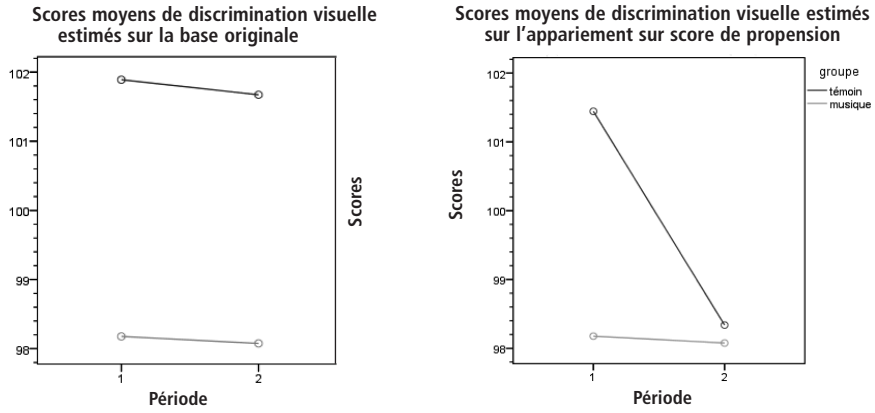

Figure 2. ***Scores moyens de discrimination visuelle selon les groupes, avant et après appariement sur score de propension***

Les résultats diffèrent nettement selon l'utilisation ou non d'une base de données appariées. Sans contrôle du biais de sélection, c'est-à-dire avant appariement, il n'apparaît pas d'effet significatif de l'entraînement musical. Le terme d'interaction groupe et temps n'est pas significatif [F = 0,07; p = 0,936], ce qui amène à conclure de façon indue que l'intervention n'a pas eu d'effet. Cependant, après appariement sur score de propension, il apparaît que les élèves du groupe témoin ont un score de discrimination visuelle qui baisse par rapport à la mesure initiale, tandis qu'il reste stable pour les élèves du groupe musique [F = 3,43; p = 0,65]. Ne pas contrôler le biais de sélection conduit à masquer cet effet protecteur de l'intervention. Les élèves ayant été stimulés par le programme d'entraînement musical ont donc bénéficié d'un effet positif sur la capacité cognitive mesurée, effet positif qui peut être attribué à l'intervention.

Conclusion

Cet article a proposé un guide méthodologique pour les chercheurs et les évaluateurs du domaine de l'éducation désireux de recourir à la méthode du score de propension. Il a présenté l'intérêt du recours à une telle méthode, les différentes étapes nécessaires à son utilisation et les choix que

celle-ci implique, et une illustration sur les données d'une expérimentation en maternelle en présentant les commandes d'un logiciel d'analyse de données répandu (Stata). Les étapes d'utilisation du score de propension sont présentées de manière synthétique dans la figure 3.

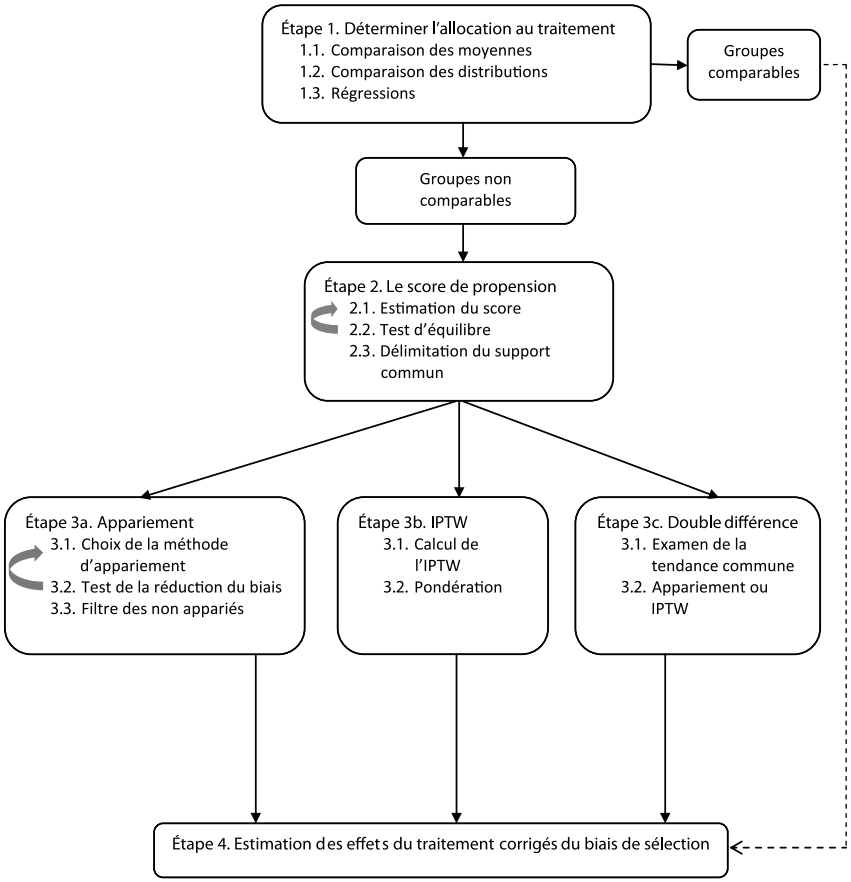

Figure 3. ***Le score de propension étapes par étapes***

IPTW: Inverse Probability of Treatment Weighting; les flèches archées épaisses symbolisent les retours nécessaires lorsque les tests d'équilibre et de réduction du biais ne sont pas satisfaisants.

Ce guide ne se veut pas pour autant une « recette » et souligne les différents arbitrages à réaliser à de nombreuses étapes. En particulier, les décisions relatives aux variables à intégrer dans l'estimation du score de propension nécessitent des arbitrages entre les considérations statistiques (liées à l'accroissement de la variance) et théoriques (pour les variables présupposées influencer le traitement, mais n'apparaissant pas significa-

tives). En l'absence de consensus, la théorie, les résultats des recherches empiriques précédentes et les données sont les meilleurs guides à cette réflexion.

Le choix de la méthode d'utilisation du score de propension soulève également des questions, aucune méthode n'étant à priori meilleure qu'une autre. Elles répondent notamment à des jeux de données différents. La pondération inverse sur les probabilités d'être traité est adaptée aux petits échantillons là où la combinaison avec la méthode des doubles différences est pertinente lorsque plusieurs périodes d'observation en prétraitement sont disponibles. L'appariement reste à ce jour la méthode la plus utilisée, sans doute parce qu'elle est intuitivement la plus aisée à comprendre. Toutefois, les différentes techniques d'appariement des individus présentent chacune des avantages et des inconvénients. Les recherches récentes indiquent que le *caliper* offre les meilleurs résultats en ce qui a trait à la réduction du biais (Baser, 2006), mais au prix d'une exclusion plus importante d'individus.

La méthode du score de propension n'est pas exempte de limites. Premièrement, la méthode est sensible au choix des variables intégrées dans l'estimation du score de propension. Si elle est tout à fait valide, cette objection concerne toute forme d'analyse reposant sur des régressions. Deuxièmement, utilisé seul, le score de propension se concentre sur la sélection par des caractéristiques observables. Ainsi, des caractéristiques inobservables, comme les préférences ou les motivations des individus, peuvent toujours introduire un biais dans les résultats. Ce problème peut néanmoins être corrigé par une combinaison avec les doubles différences. Troisièmement et enfin, pour que les inférences causales réalisées soient robustes, le support commun doit être important. En effet, la méthode n'a pas d'intérêt si tous les individus ayant un score de propension élevé sont dans le groupe traité et si les individus ayant un score de propension bas sont dans le groupe témoin. Par ailleurs, le score de propension n'est d'aucun secours lorsque les groupes comparés sont trop différents[14]. Quelle que soit la méthode d'analyse retenue, il semble légitime de s'interroger sur la qualité du groupe choisi comme comparateur.

Malgré ces limites et en respectant quelques précautions présentées dans cet article, le score de propension est une méthode innovante permettant de mieux évaluer les effets causaux d'une intervention. Les recherches expérimentales et quasi expérimentales en éducation gagneraient à la mobiliser plus souvent.

NOTES

1 Le terme traitement sera utilisé de façon générique pour désigner toute forme d'intervention, de pratique, de politique ou de programme dans le domaine de l'éducation.

2 D'autres stratégies d'analyse pourraient être retenues. Un modèle à effets fixes pourrait être utilisé sur un long panel de données. La technique des variables instrumentales pourrait être utilisée sur des données en coupe transversale et la régression sur discontinuité lorsqu'un « saut » de distribution détermine le traitement (par ex., des élèves reçoivent un soutien scolaire si leurs notes sont inférieures à un seuil fixé au préalable). Se concentrer sur les études expérimentales et quasi expérimentales conduit néanmoins à privilégier et à approfondir la méthode du score de propension. Le lecteur intéressé par ces autres stratégies économétriques d'évaluation de programme peut consulter imbens et wooldridge (2009).

3 Si le score de propension permet de résoudre le problème de dimensionnalité, il n'en demeure pas moins qu'il nécessite toujours un nombre suffisant de covariables.

4 Les essais thérapeutiques, avec procédure en double aveugle où ni le praticien ni le patient ne connaissent la nature du traitement, exemplifient ce type de devis. Cependant, même dans ce cas quasi idéal, les patients acceptant de participer sont déjà autosélectionnés dans la population.

5 Cette situation peut se présenter dans des études expérimentales si la randomisation a parfaitement fonctionné.

6 Les arbres de classification et de régression peuvent également être utilisés, mais cette méthode est très peu courante.

7 Un exemple sera plus parlant : une différence persiste seulement dans le troisième quintile du score de propension pour la variable âge. Il est nécessaire de diviser ce quintile pour examiner les cinquième et sixième déciles du score de propension. Au sein de ces deux strates, la significativité de la différence pour l'âge est de nouveau évaluée.

8 Pour une comparaison détaillée des deux méthodes, le lecteur intéressé peut consulter Caliendo et Kopeinig (2008).

9 Cette étape est à ne pas confondre avec la vérification de la qualité du score de propension, qui aura déjà été faite au préalable, avant d'apparier.

10 Par exemple, si une intervention éducative vise uniquement les garçons et jamais les filles.

11 Le score de propension peut quant à lui être utilisé sur des données en coupe transversale, type de devis qui ne fait toutefois pas partie des recherches présentées ici.

12 Les logiciels R et SAS peuvent également être utilisés pour le score de propension, mais nécessitent davantage de programmation pour y parvenir.

13 Le « biais » tel que défini ici fait donc référence à une forme de résidu standardisé. Le terme « biais » est celui en vigueur dans la littérature du score de propension ; c'est pourquoi il est utilisé ici.

14 Cette situation est peu vraisemblable en devis expérimental, le groupe témoin étant tout de même construit à priori. En revanche, elle l'est en devis quasi expérimental.

RÉFÉRENCES

Abadie, A., & Imbens, G. W. (2005). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*(1), 235-267.

Altman, D. G., & Bland, J. M. (1999). Treatment allocation in controlled trials: Why randomise? *British Medical Journal*, *318*(7192), 1209.

Altman, D. G., & Doré, C. J. (1990). Randomisation and baseline comparisons in clinical trials. *The Lancet*, *335*(8682), 149-153.

Banerjee, V. D., & Duflo, E. (2009). The experimental approach to development economics. *Annual Review of Economics, 1*(1), 151-178.

Baser, O. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health*, *9*(6), 377-385.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal, 2*(4), 358-377.

Bénabou, R., Kramarz, F., & Prost, C. (2004). Zones d'éducation prioritaire : quels moyens pour quels résultats ? *Économie et Statistique*, *380*(1), 3-34.

Berger, V. W. (2005). Quantifying the magnitude of baseline covariate imbalances resulting from selection bias in randomized clinical trials. *Biometrical Journal*, *47*(2), 119-127.

Berger, V. W., & Exner, D. V. (1999). Detecting selection bias in randomized clinical trials. *Controlled Clinical Trials*, *20*(4), 319-327.

Bjerk, D. (2009). How much can we trust causal interpretations of fixed-effects estimators in the context of criminality? *Journal of Quantitative Criminology*, *25*(4), 391-417.

Brodaty, T., Crépon, B., & Fougère, D. (2007). Les méthodes microéconométriques d'évaluation et leurs applications aux politiques actives de l'emploi. *Économie et Prévision, 177*, 91-118.

Bryson, A., Dorsett, R., & Purdon, S. (2002). *The use of propensity score matching in the evaluation of active labour market policies*. Department for Work and Pensions (No. 4) Working paper.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31-72.

Cochran, W. G., & Chambers, S. P. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, *128*(2), 234-266.

Dehejia, R. (2005). Practical propensity score matching: A reply to Smith and Todd. *Journal of Econometrics*, *125*(1), 355-364.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053-1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, *84*(1), 151-161.

Dumont, M., Leclerc, D., Massé, L., Potvin, P., & McKinnon, S. (2009). Programme de gestion du stress des adolescents comme levier de résilience. In N. Nader-Grosbois (Eds.), *Résilience, régulation et qualité de vie : Concepts, évaluation et intervention* (pp. 301-305). Louvain-la-Neuve, Belgique : Presses universitaires de Louvain.

Fortin, L. (2012). *Trait d'union*. Québec, Canada : Centre de transfert pour la réussite éducative du Québec (CTREQ).

Fortin, L., Marcotte, D., Potvin, P., Royer, É., & Joly, J. (2006). Typology of students at risk of dropping out of school: Description by personal, family and school factors. *European Journal of Psychology of Education*, *21*(4), 363-383.

Fortin, L., & Potvin, P. (2007). *Logiciel de dépistage du décrochage scolaire*. Québec, Canada : Centre de transfert pour la réussite éducative du Québec (CTREQ).

Fortin, L., Royer, É., Potvin, P., Marcotte, D., & Yergeau, É. (2004). La prédiction du risque de décrochage scolaire au secondaire : facteurs personnels, familiaux et scolaires. *Revue canadienne des sciences du comportement*, *36*(3), 219-231.

Grimes, D. A., & Schulz, K. F. (2002). Bias and causal associations in observational research. *Lancet*, *359*(9302), 248-252.

Harrison, G. P., Gruber, A. J., Hudson, J. I., Huestis, M. A., & Yurgelun-Todd, D. (2002). Cognitive measures in long-term cannabis users. *The Journal of Clinical Pharmacology*, *42*(11), 41-47.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, *64*(4), 605-654.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, *65*(24), 261-294.

Heckman, J. J., LaLonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. *Handbook of Labor Economics*, *3*, 1865-2097.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161-1189.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*(1), 4-29.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5-86.

Janosz, M. (2012). *QES-Web*. Québec, Canada : Centre de transfert pour la réussite éducative du Québec (CTREQ).

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics*, *84*(2), 205-220.

Lecocq, A. (2012). *Genèse et évolution des compétences des élèves à la fin de l'école maternelle : éléments d'analyse à partir de données de panel et d'une expérimentation musicale* (Thèse de doctorat non-publiée). Université de Bourgogne, Dijon, France.

Leuven, E., & Sianesi, B. (2012). *PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing*. Boston, MA: Boston College Department of Economics.

Marcotte, D. (2006). *Pare-Chocs, programme d'intervention auprès d'adolescents dépressifs.* Québec, Canada : Septembre éditeur.

Marcotte, D., & Baron, P. (1993). L'efficacité d'une stratégie d'intervention émotivo-rationnelle auprès d'adolescents dépressifs du milieu scolaire. *Revue canadienne de counseling et de psychothérapie*, *27*(2), 77-92.

Murray, C., & Malmgren, K. (2005). Implementing a teacher-student relationship program in a high-poverty urban school : Effects on social, emotional, and academic adjustment and lessons learned. *Journal of School Psychology*, *43*(2), 137-152.

Robins, J. M., Hernán, M. Á., & Brumback B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550-560.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985a). The bias due to incomplete matching. *Biometrics*, *41*(1), 103-116.

Rosenbaum, P. R., & Rubin, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, *127*(8), 757-763.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103-116.

Schmidt, C. M., & Augurzky, B. (2001). *The propensity score: A means to an end* (No. 271). Extrait de http://anon-ftp.iza.org/dp271.pdf

Smith, J. A., & Todd, P. E. (2001). Reconciling conflicting evidence on the performance of propensity-score matching methods. *The American Economic Review*, *91*(2), 112-118.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*(1), 305-353.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*(3), 250-267.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology*, *18*, 327-350.

Wood, L., Egger, M., Gluud, L. L., Schulz, K. F., Jüni, P., Altman, D. G., & Sterne, J. A. (2008). Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. *British Medical Journal*, *336*(7644), 601-605.

Date de réception : 8 mars 2013

Date de réception de la version finale : 11 novembre 2013

Date d'acceptation : 12 mars 2014