LES MÉTHODES DE *TESTING* PERMETTENT-ELLES D'IDENTIFIER ET DE MESURER L'AMPLEUR DES DISCRIMINATIONS ?

Romain Aeberhardt*, **Denis Fougère**** et **Roland Rathelot*****

Parmi les méthodes utilisées pour détecter la présence de phénomènes discriminatoires, les expériences de terrain de type *testing* font aujourd'hui l'objet d'un intérêt tout particulier. La mesure de ces phénomènes à l'aide des sources statistiques usuelles soulève en effet un certain nombre de difficultés auxquelles le *testing* est présumé apporter une réponse simple et efficace. Il était donc tout naturel qu'*Économie et Statistique* ouvre ses colonnes à cette approche et l'article de Yannick L'Horty, Emmanuel Duguet, Loïc du Parquet, Pascale Petit et Florent Sari est une très bonne occasion de le faire. Comme c'est souvent le cas pour ce type de travail, les résultats obtenus sont relatifs à un terrain particulier, mais ils sont intéressants par la tentative de croiser différentes sources potentielles de discrimination et d'analyser leur interaction. Et leur article est surtout une introduction concrète très utile à la méthodologie du *testing*, avec une présentation détaillée du protocole suivi et de ses justifications.

Pour autant, il importe de rappeler que cette méthodologie n'est pas sans limites, et que ses résultats doivent donc être considérés avec précaution. Le présent commentaire ne va pas reprendre point par point les différents aspects du travail de L'Horty et de ses co-auteurs, mais il va s'efforcer de rappeler les principaux éléments du débat dont les méthodes de *testing* sont actuellement l'objet.

Le *testing* : la seule méthode permettant de détecter la discrimination ?

Les expériences de *testing* sont-elles les seules méthodes empiriques permettant de mettre en évidence l'existence de discrimination ? Dans la littérature économique, deux autres types de procédures statistiques sont utilisées pour détecter la présence de discriminations à l'encontre d'un groupe de personnes (ce groupe pouvant être défini par son genre, son âge, son origine géographique ou nationale, son apparence physique, ou son orientation sexuelle) et sur un marché particulier : marché du travail, du logement, ou encore du crédit bancaire.

Le premier grand type de méthode consiste à construire un modèle de fonctionnement d'un marché (marché du logement, du travail, etc.) caractérisé par l'existence de comportements discriminatoires et d'estimer, en utilisant des données non-expérimentales, les paramètres de ce modèle. Parmi ces paramètres, certains sont relatifs à la discrimination. Cette méthode a deux avantages : elle s'appuie sur des hypothèses clairement énoncées et la validation du modèle est fondée sur l'analyse statistique de grands échantillons constitués de données non-expérimentales. Dans le cas malencontreux, mais probable, où le modèle est par trop réducteur, ce type d'approche peut fournir des résultats fortement biaisés.

Une autre méthode consiste à comparer les situations auxquelles font face le groupe de personnes potentiellement discriminées et le groupe de celles qui ne le sont pas. Pour être acceptable, une telle comparaison doit évidemment tenir compte des différences de caractéristiques entre les deux groupes.

Par exemple, que reste-t-il de l'écart de taux d'emploi entre les Français dont les parents sont immigrés et les Français de parents nés en France, une fois que l'on a tenu compte du fait que les premiers sont plus jeunes et moins diplômés ? Cette méthode, qui est facile à mettre en œuvre et qui utilise comme la précédente des données non-expérimentales, a un inconvénient majeur : l'écart non expliqué par les différences de caractéristiques observables, telles que l'âge, le niveau d'éducation, la commune ou le quartier de résidence, etc., ne peut être imputé avec certitude à de la discrimination. On ne peut jamais exclure totalement que les écarts inexpliqués soient au moins partiellement dus à des différences qui n'ont pas pu être prises en compte dans l'analyse statistique.

Le *testing* serait de ce fait la seule méthode permettant de détecter directement l'existence de

* *Dares et Crest.*
** *CNRS, Crest, LIEPP (Sciences PO Paris), CEPR et IZA.*
*** *Crest et Insee.*

comportements discriminatoires. Néanmoins, comme le rappellent Duguet *et al.* (2009), procéder à un *testing* est long et coûteux puisqu'une telle procédure oblige à concevoir et mettre sur pied une expérimentation et ensuite à recueillir des données spécifiques. En outre, les résultats d'une telle expérience n'ont souvent qu'une portée limitée ; il n'est pas possible de les extrapoler au-delà du contexte particulier de moment, de lieu, de marché, etc., au sein duquel s'est déroulée l'expérience. Une troisième limite, inhérente aux expériences de *testing* reposant par exemple sur l'envoi de CV, est de se limiter à la première étape de la sélection des candidats, en ce cas à la convocation à un rendez-vous d'embauche ou à un entretien. Mais ces limites ne sont pas les seules et plusieurs précautions doivent donc être prises par les chercheurs au moment de la conception du protocole de l'expérience et lors de l'analyse statistique de ses résultats.

Ce que l'on veut tester n'est pas toujours ce que l'on teste

Lorsque la caractéristique qui est source potentielle de discrimination (l'âge, l'origine sociale ou ethnique, le genre, etc.) n'est pas directement identifiable dans le dossier, le CV ou l'apparence de l'acteur qui participe au *testing*, les chercheurs doivent faire des hypothèses implicites sur la manière dont les agents qui choisissent un candidat (employeurs, propriétaires, agences de location, organismes de crédit, etc.) interprètent l'information effectivement transmise. Par exemple, puisqu'il n'est pas courant que les personnes postulant à un emploi mentionnent sur leur CV leur origine ethnique, l'hypothèse souvent faite est que les employeurs déduisent du nom ou du prénom des candidats leur origine nationale ou ethnique. *Stricto sensu*, lorsque l'on construit un *testing* visant à mettre en évidence par le biais de l'envoi de CV la discrimination à l'encontre des immigrés ou de leurs descendants au moment de l'embauche, on mélange souvent deux questions pourtant distinctes : quel type d'information un employeur déduit-il d'un nom ou d'un prénom que les chercheurs considèrent comme « typiquement maghrébin », et quelle est la réaction de l'employeur face à un candidat d'origine maghrébine ? À l'évidence, dans cet exemple, la difficulté provient du fait qu'au vu du nom ou du prénom (voire de l'adresse) du candidat, l'employeur peut prendre une décision qui ne s'apparente pas directement à un comportement discriminatoire. Deux exemples permettent d'illustrer ce problème.

Commentant les résultats obtenus par Bertrand et Mullainathan (2004) à l'issue d'un *testing* par envoi de CV, Fryer et Levitt (2004) ont montré que les prénoms choisis pour signaler aux recruteurs qu'ils ont affaire à des candidats Afro-américains sont généralement des prénoms qui sont non seulement plus souvent portés par des Afro-américains, mais également par des personnes par ailleurs plus désavantagées que la moyenne (ayant une mère plus pauvre, fait moins d'études, etc.). Dans ce cas, les écarts mesurés entre les deux groupes ne distinguent pas le désavantage lié à l'ethnicité de celui lié à une origine sociale plus défavorisée. Cette critique s'applique-t-elle au cas français ? Les Français d'origine maghrébine portant un prénom moins « typé » (Inès, Sofia, etc.) ont-ils des origines sociales, géographiques, etc., différentes de ceux dont le prénom est plus caractéristique (Aicha, Fatima, etc.) ? Pour répondre à cette question, il serait donc utile, avant de procéder à un *testing*, d'examiner la corrélation empirique entre la fréquence des prénoms et l'origine sociale ou géographique des personnes du groupe considéré.

Le deuxième exemple, extrait d'un article d'Heckman et Siegelman (1993), est celui d'un protocole où des acteurs participent à des entretiens d'embauche pour tester l'existence éventuelle d'une discrimination ethnique à l'embauche. Sauf cas exceptionnel, les acteurs sont peu nombreux et jouent chacun un rôle bien déterminé : ainsi, un acteur Noir ne peut pas prendre la place d'un acteur Blanc. Même si ces acteurs sont censés être les plus comparables possibles (à l'exception de la caractéristique qui fait l'objet du test), on ne peut jamais complètement exclure qu'il subsiste de légères différences comportementales (par exemple, en termes de dynamisme, d'élocution, d'enthousiasme, etc.) qui favorisent l'un ou l'autre des acteurs au moment de l'entretien. Puisque les acteurs ne peuvent pas changer de rôle, ces différences sont systématiques, et de ce fait, en leur présence, il est difficile de repérer l'existence d'éventuels comportements discriminatoires. L'estimation reposant sur la comparaison des moyennes de résultats (ici, le nombre de propositions d'embauche) est donc biaisée ; en outre, le signe et

l'amplitude de ce biais sont difficilement prévisibles. La seule solution consiste ici à multiplier le nombre d'acteurs, ce qui peut augmenter significativement le coût de l'expérience.

La discrimination potentielle n'est pas nécessairement la discrimination réelle

Heckman (1998) met en doute la pertinence de la discrimination mesurée par *testing*. Pour lui, la procédure de *testing* mesure la discrimination potentielle, c'est-à-dire la discrimination qui surviendrait dans le cas hypothétique où le groupe potentiellement discriminé se comporterait comme les chercheurs administrant le *testing* l'imaginent. En réalité, le comportement de ce groupe peut être très différent de celui qui est postulé par les chercheurs. La discrimination réellement subie sur le marché considéré peut de ce fait être d'une ampleur sensiblement différente de celle mesurée par la procédure de *testing*. Pour le dire plus précisément encore, le *testing* donne des informations statistiques sur les comportements de *demande* observés sur le marché considéré, mais pas nécessairement sur la situation prévalant à l'équilibre sur ce marché.

Une expérience de *testing* par CV consiste à envoyer pour chaque offre d'emploi déposée deux types de CV, qu'un seul critère différencie. Parmi les entreprises contactées, certaines sont susceptibles d'adopter un comportement discriminatoire. Cependant, dans certains cas, lorsque les personnes potentiellement discriminées sont relativement peu nombreuses, il se peut qu'elles trouvent suffisamment d'offres parmi les entreprises qui ne discriminent pas et ne soient de ce fait que rarement victimes des comportements des employeurs qui discriminent.

Un autre exemple est fourni par Riach et Rich (2010) qui utilisent le *testing* pour mesurer la discrimination à l'encontre des travailleurs âgés. Dans les entreprises du secteur de l'hôtellerie et de la restauration qui offrent des emplois, les chercheurs envoient des candidatures spontanées de serveurs jeunes et d'autres plus âgés. Leurs résultats indiquent un écart de réponse important en faveur des candidats plus jeunes. Cet écart est en toute vraisemblance la conséquence d'une discrimination à l'encontre des travailleurs les plus âgés. Mais est-il pour autant une mesure exacte du phénomène ? Si l'on fait l'hypothèse que les serveurs plus âgés utilisent plus fréquemment les contacts personnels et moins souvent les annonces et les candidatures spontanées lorsqu'ils cherchent un emploi, alors il se peut que le *testing* conduit par Riach et Rich (2010) ne reflète pas vraiment le comportement de recherche d'emploi des candidats âgés, et que la discrimination à leur encontre soit moindre que celle suggérée par le résultat de leur étude.

Détecter la discrimination ou en mesurer l'ampleur ?

Sous réserve de s'assurer que la variable utilisée pour définir les deux groupes correspond bien à la caractéristique dont on souhaite tester l'influence et que les offres sont représentatives du marché dont on étudie le fonctionnement, le *testing* permet de calculer un écart moyen de la variable d'intérêt entre les deux groupes. Un test de la significativité statistique de cet écart correspond à un test de l'hypothèse nulle : « Le côté *demande* du marché traite de manière identique les individus des deux groupes ». Il s'agit donc d'un test de l'existence d'un processus discriminatoire.

Quid de la taille du coefficient estimé ? Peut-on dire qu'il mesure l'ampleur de la discrimination ? On peut penser à deux obstacles. Premièrement, comme souligné précédemment, la discrimination mise en évidence par le *testing* correspond à une discrimination potentielle et non nécessairement à la discrimination effectivement subie. Deuxièmement, l'écart mesuré entre les deux groupes dépend crucialement de la qualité des candidatures. Ce deuxième point est important car il rend difficile les comparaisons d'une expérience à l'autre. L'écart mesuré entre les deux groupes sera en général différent si les chercheurs envoient des candidatures de plus ou moins bonne qualité. Sous l'hypothèse d'une discrimination relativement homogène, il est probable que l'écart mesuré entre les groupes sera croissant en fonction de la qualité des candidatures.

Idéalement, on souhaiterait que les candidatures soient représentatives des candidatures existantes, c'est-à-dire que soit utilisé un mélange de candidatures de bonne, moyenne et mauvaise

qualité. En pratique, lors d'un *testing* sur la discrimination à l'embauche, par exemple, le nombre d'offres est très grand, mais le nombre de CV est faible. Il est souvent peu crédible que les CV utilisés soient représentatifs de la totalité des CV envoyés par les candidats. Là encore, un moyen de résoudre cette difficulté serait d'augmenter le nombre de CV envoyés et de s'assurer de leur représentativité, par exemple en examinant des CV réellement envoyés.

Une fois ce point résolu, on obtiendrait un estimateur crédible de la discrimination *potentielle* sur le marché examiné. Parvenir à mesurer la discrimination réelle, c'est-à-dire être capable de savoir quelle part des écarts empiriques (de taux d'emploi, par exemple) entre deux populations est attribuable à de la discrimination, n'est pas possible sans hypothèse supplémentaire sur le fonctionnement du marché considéré.

Que peuvent apporter les expérimentations contrôlées au *testing* ?

Un *testing* est un cas particulier d'expérimentation contrôlée. Deux pratiques, courantes dans la littérature utilisant l'expérimentation contrôlée pour évaluer l'impact de politiques publiques, pourraient être adaptées au *testing*.

Premièrement, un calcul de puissance [1] permet de calibrer *ex ante* le protocole expérimental. Prenons l'exemple d'un *testing* utilisant des CV pour tester la discrimination à l'embauche selon le groupe ethnique. Préalablement au lancement de l'opération, les chercheurs doivent commencer par choisir leurs populations d'intérêt (ici, les groupes ethniques) et le marché sur lequel ils souhaitent tester la présence de discrimination (par exemple, les techniciens informatiques sortant du système universitaire). Ils énoncent ainsi la question à laquelle ils souhaitent répondre et spécifient l'hypothèse nulle correspondante (ici, il n'existe pas de différence de traitement entre les deux groupes). Ils doivent également faire des hypothèses concernant l'espérance mathématique de la variable d'intérêt dans le groupe potentiellement non discriminé (par exemple, le taux de retour est en moyenne strictement positif pour ces personnes). Ces hypothèses leur permettent ensuite de calculer, compte tenu du plan d'expérience, le nombre d'offres à laquelle il faut envoyer les deux types de CV pour détecter avec une certaine puissance (souvent fixée à 80 %) un écart d'une certaine valeur entre les groupes. Plus l'écart postulé est élevé, plus l'hypothèse nulle est facile à rejeter et plus le nombre de CV devant être envoyés 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. Des calculs de puissance sont présentés par exemple dans l'article de Duflo *et al.* (2008).

De plus en plus souvent, les chercheurs s'efforcent de définir et d'annoncer à l'avance les hypothèses qui vont être testées, tout en restreignant leur nombre de façon à maximiser la puissance statistique des procédures de test. Si tel n'est pas le cas, la tentation peut être grande de chercher à faire apparaître *a posteriori* des résultats significatifs sur certaines sous-populations ou dans des dimensions qui n'étaient pas *a priori* prévues par l'expérimentation. Avec des tests d'un niveau de 10 %, on finit par trouver des effets significatifs dans 10 % des cas, même si les effets réels sont nuls. Par exemple, l'un des objectifs de l'article est de comparer l'effet du lieu de résidence sur l'accès à un entretien d'embauche. Les auteurs concluent peut-être de manière trop hâtive à une hétérogénéité de cet effet en fonction du sexe et de l'origine nationale des candidats à partir de sous-échantillons de faible taille. Ainsi, leurs estimations, trop imprécises, ne leur permettent pas d'affirmer que le lieu de résidence a un effet différent pour les femmes et les hommes d'origine maghrébine vivant à Enghien ou à Villiers-le-Bel (cf. le tableau 4 de l'article).

1. La puissance statistique est la probabilité que l'hypothèse nulle (ici, l'absence de différence de traitement entre les deux groupes considérés) soit rejetée et que le testing *ne puisse donc pas permettre de repérer l'association réellement existante entre la caractéristique sociodémographique considérée (par exemple, l'âge, le genre, l'origine sociale ou nationale, etc.) et la variable de résultat (par exemple, l'accès à un entretien d'embauche). 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'étude, 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'étude puisse mettre en évidence l'effet recherché.*

Choisir le plan d'expérience le plus adapté

L'expérience de *testing* la plus élémentaire est celle qui consiste à comparer la situation de deux groupes. Même dans ce cas simple, deux plans d'expérience sont envisageables. Il est tout d'abord possible d'envoyer une et une seule candidature à chaque offre d'emploi disponible, en appariant de manière aléatoire les candidatures aux offres déposées. Toutefois, plus souvent, les candidatures sont groupées par paires, une candidature de chaque groupe étant adressée de ce fait à une même offre. Laquelle de ces deux possibilités correspond au meilleur plan d'expérience ? Tout dépend de l'hétérogénéité des offres d'emploi. Dans le cas où celles-ci sont de natures très différentes, le second type de plan d'expérience permet de gagner de la puissance statistique. Sur ce point, les travaux relatifs à la mise en place des plans d'expérience dans le domaine de la bio-statistique peuvent être avantageusement mobilisés (voir, par exemple, Montgomery, 2008).

Pour qu'un *testing* soit valide, il est nécessaire qu'il ne soit pas détecté, c'est-à-dire que les candidatures envoyées par les chercheurs soient traitées par les employeurs contactés de la même manière que les candidatures adressées par de « vrais » candidats. Nous ne connaissons pas d'étude ayant plus particulièrement examiné ce type d'éventualité et les biais possibles qu'elle introduit dans les résultats. En outre, ce risque de détection est amplifié lorsque les chercheurs envoient un assez grand nombre de candidatures à la même offre. Il en est ainsi lorsqu'ils souhaitent tester simultanément plusieurs hypothèses à l'aide d'une même expérience de *testing,* et qu'ils sont pour cela obligés de construire plus de deux groupes de candidats. C'est le cas dans la présente étude qui utilise douze groupes de candidats pour tester l'existence concomitante de plusieurs types de discrimination, en fonction de l'origine ethnique, du genre et du lieu de résidence.

Si l'on connaissait, au moins grossièrement, la manière dont la probabilité de détection des « fausses » candidatures augmente avec le nombre de candidatures associées à la même offre d'emploi, il serait possible d'en déduire un plan d'expérience optimal. Un plus grand nombre de candidatures pour chaque offre augmente certes la puissance des tests statistiques mais elle accroît aussi la probabilité de détection des « fausses » candidatures par les employeurs. En présence d'un tel arbitrage, quel est le nombre optimal de candidatures qui doit être adressé à chaque offre ? Cette question est aujourd'hui largement sans réponse.

BIBLIOGRAPHIE

Bertrand M. et Mullainathan S. (2004), « Are Emily and Greg More Employable Than Lakisha and Jamal ? A Field Experiment on Labor Market Discrimination », *American Economic Review*, vol. 94, nº 4, pp. 991-1013.

Duflo E., Glennester R. et Kremer M. (2008), « Using Randomization in Development Economics Research : A Toolkit », dans *Handbook of Development Economics*, édité par T. Paul Schultz et John A. Strauss, Elsevier.

Duguet E., L'Horty Y. et Petit P. (2009), « L'apport du testing à la mesure des discriminations », *Connaissance de l'Emploi*, Centre d'Études de l'Emploi, nº 68.

Heckman J.J. (1998), « Detecting Discrimination », *Journal of Economic Perspectives*, vol. 12, nº 2, pp. 101-116.

Heckman J.J. et Siegelman P. (1993), « The Urban Institute Audit Studies : Their Methods and Findings », dans *Clear and Convincing Evidence : Measurement of Discrimination in America*, édité par M. Fix et R. Struyk, Washington DC, The Urban Institute Press.

Fryer R.G. Jr. et Levitt S.D. (2004), « The Causes and Consequences of Distinctively Black Names », *Quarterly Journal of Economics*, vol. 119, nº 3, pp. 767-805.

Montgomery D. C. (2008), *Design and Analysis of Experiments*, Wiley.

Riach P.A. et Rich J. (2002), « Field Experiments of Discrimination in the Market Place », *Economic Journal*, vol. 112, pp. 480-518.

Riach P.A. et Rich J. (2010), « An Experimental Investigation of Age Discrimination in the English Labor Market », *Annales d'Économie et de Statistique*, nº 99/100, pp. 169-186.