la vie des idées.fr

La crise de la réplicabilité

par Aurélien Allard

Certaines sciences sociales et bio-médicales sont confrontées, depuis vingt ans, à l'impossibilité de répliquer les expériences sur lesquelles se fondait jusque-là leur notoriété. Pour préoccupante qu'elle soit, cette défaillance offre l'occasion de réviser en profondeur leur méthode scientifique.

Les sciences sociales et bio-médicales connaissent actuellement une des crises les plus importantes de leur histoire. Depuis le début des années 2000, et tout particulièrement depuis 2011, il est devenu de plus en plus visible que de nombreuses expériences, aussi bien en psychologie qu'en économie ou en études animales, ne sont pas réplicables et ne correspondent pas à une véritable connaissance scientifique. De façon de plus en plus importante, les méthodes actuellement employées dans ces disciplines sont critiquées pour leur manque de rigueur, et un nombre croissant de chercheurs appellent à un renouvellement radical des pratiques scientifiques dans ces domaines.

De nombreuses collaborations scientifiques internationales visant à répliquer des expériences passées ont été mises en œuvre ces dernières années. Répliquer une expérience ou une enquête consiste à tenter de la reproduire à l'identique, afin de voir dans quelle mesure des résultats similaires peuvent être obtenus. De 2012 à 2015, une collaboration internationale de grande ampleur a tenté d'estimer la réplicabilité d'expériences et d'enquêtes de psychologie sociale et cognitive. Sur les 100 expériences que les chercheurs du *Reproducibility Project : Psychology* ont tenté de reproduire, seules 39 ont donné des résultats conformes à ceux de l'expérience originale. Ces résultats ont suscité un large écho, aussi bien parmi la communauté de chercheurs que dans la presse grand public anglo-américaine. En économie expérimentale, un projet similaire, mais conduit à une échelle moindre, n'est parvenu à répliquer les résultats que de 11 des 18 expériences. Une entreprise comparable est en cours en biologie du cancer, après que des chercheurs travaillant pour des entreprises privées (*Bayer* et *Amgen*) ont alerté la communauté scientifique, entre 2011 et 2012, sur le fait que la grande majorité des expériences publiées dans les disciplines bio-médicales n'étaient pas reproductibles. Ces

derniers disent n'être parvenus à répliquer qu'une toute petite minorité (entre 11 et 25 %) d'expériences[1].

Si ces résultats mettent en évidence certains problèmes déterminants pour les sciences sociales et bio-médicales, ils offrent également l'occasion d'une amélioration radicale des pratiques de recherche dans ces domaines. L'existence de collaborations internationales révèle tout à la fois l'ampleur des problèmes de réplicabilité et la volonté de mouvements réformateurs d'améliorer les pratiques. La réflexion méthodologique et les projets d'amélioration des pratiques ont connu un essor sans précédent ces dernières années. Certains commentateurs ont ainsi considéré que les années 2010 présidaient à une véritable renaissance de la psychologie comme discipline scientifique.

Reproduire et répliquer

On pourrait croire que les questions de réplicabilité s'appliquent avant tout aux disciplines expérimentales. La question générale de la possibilité de reproduire une observation est cependant fondamentale pour toute discipline à prétention scientifique, que celle-ci soit qualitative ou quantitative, expérimentale ou non. Pour comprendre les enjeux liés aux questions de réplicabilité, il faut faire la distinction entre reproduire analytiquement une observation (*reproduce* en anglais) et la répliquer (*replicate*)[2].

Reproduire analytiquement une observation consiste à chercher à reproduire ses résultats à partir des matériaux originellement utilisés par l'expérimentateur. Par exemple, dans le cas d'une recherche qualitative basée sur des entretiens, comme en anthropologie ou en sociologie, la reproduction des résultats peut consister dans le fait de consulter des enregistrements effectués par l'anthropologue, et de vérifier si les participants de l'enquête ont bien dit ce que l'anthropologue déclare qu'ils ont dit dans sa recherche. Dans le cas d'une recherche quantitative, il est possible de se servir du même tableau de données utilisé par quelqu'un, d'effectuer les mêmes analyses statistiques, et de vérifier que les résultats concordent. La reproduction analytique des résultats est cruciale à la fois pour prévenir le problème de la fraude scientifique, et pour repérer les erreurs honnêtes qui peuvent conduire les scientifiques à reporter des résultats en contradiction avec ce qu'ils ont trouvé.

---

[1] Sur ces divers projets de réplication, voir respectivement Open Science Collaboration, « Estimating the Reproducibility of Psychological Science », *Science*, 349, 2015 ; Colin Camerer et al., « Evaluating Replicability of Laboratory Experiments in Economics », *Science*, 2016 ; C. Glenn Begley et Lee M. Ellis, « Drug Development: Raise Standards for Preclinical Cancer Research », *Nature*, 483, 2012 ; Florian Prinz, Thomas Schlange, & Khusru Asadullah, « Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets? », *Nature Reviews Drug Discovery*, 10, 2011, 712.

[2] Il existe en anglais un certain flottement dans la distinction entre *replicability* et *reproducibility*, différents auteurs employant les deux termes dans des sens parfois contradictoires. La perspective adoptée ici est cependant largement répandue. Voir Roger D. Peng, Francesca Dominici, Scott L. Zeger, « Reproducible Epidemiologic Research », *American Journal of Epidemiology*, vol.163, n° 9, 2006, p. 783-789.

Même si la possibilité de reproduire les résultats d'une recherche est une question importante, les problèmes évoqués dans cet article concernent essentiellement la question de la réplicabilité. Ce dernier concept s'applique à toute situation où un chercheur veut généraliser au delà des cas qu'il a observés. Répliquer une observation consiste alors à utiliser les mêmes méthodes que celles employées par l'auteur original, et à tenter de voir dans quelle mesure les nouvelles observations sont compatibles avec celles originalement reportées. En anthropologie, on peut penser à la célèbre controverse autour de l'échec de réplication du terrain de Margaret Mead par Derek Freeman à Samoa. Alors que Mead avait décrit Samoa comme une société à la sexualité libre, Freeman avait mis en avant son caractère puritain. Mead avait avant tout fondé son travail de terrain sur des entretiens avec deux informateurs, et avait cherché à généraliser ses données à la sexualité de la population de Samoa tout entière. Pour répliquer des observations quantitatives, et toujours dans un cadre non expérimental, il est possible d'effectuer une nouvelle enquête suivant les mêmes méthodes, afin de mesurer jusqu'à quel point les résultats sont généralisables. Par exemple, Fabien Jobard et Sophie Nevanen ont observé que le groupe ethnique n'avait pas d'impact sur les jugements de condamnation pénale dans un Tribunal de grande instance de la région parisienne ; ils ont ensuite généralisé leurs résultats, les interprétant comme l'indication d'une absence de discrimination sur l'ensemble de la France[3]. Répliquer ces résultats pourrait prendre la forme d'une autre enquête dans plusieurs autres tribunaux français, afin de voir s'il est possible d'obtenir des résultats concordants.

Crise de confiance

Si les problèmes de réplicabilité touchent l'ensemble des sciences sociales et biomédicales, la psychologie sociale et cognitive a été au cœur de la crise, et des mouvements de réforme.

Au cours des années 2000, la psychologie sociale a connu une période faste en Grande-Bretagne et en Amérique du Nord, marquée par la publication de nombreux best-sellers éveillant l'intérêt d'un large public[4]. À partir de 2011, cependant, une série de scandales a révélé certains problèmes profonds affectant la discipline.

En 2011, Daryl Bem, professeur à la prestigieuse université de Cornell, aux États-Unis, a publié une série d'expériences dans le *Journal of Personality and Social Psychology*, une

[3] Fabien Jobard et Sophie Névanen, « La couleur du jugement », *Revue française de sociologie*, 2007/2, vol. 48, p. 243-272.
[4] Parmi les nombreux best-sellers des années 2000, on peut penser à l'ouvrage de Carol Dweck, *Mindset (*Ballantine Books, 2007), ou à *Stumbling on Happiness* (Random House Books, 2006) de Daniel Gilbert.

des revues les plus réputées du champ[5]. Cette série d'expériences visait à montrer l'existence de pouvoirs extra-sensoriels : D. Bem a tenté de démontrer que les humains étaient capables de prédire l'avenir[6]. Une expérience visait notamment à prouver que des individus qui ont la possibilité de réviser *après la fin de l'examen* obtiennent de meilleures notes à cet examen que ceux qui ont été privés de cette possibilité. À travers 10 expériences différentes, D. Bem a reporté à chaque fois des résultats conformes à ses prédictions.

Il va sans dire que de tels résultats contredisent toutes les lois physiques connues, et qu'à peu près personne n'y a cru à l'époque. Cependant, ces expériences ont posé un problème radical à la psychologie sociale et cognitive, parce que les méthodes employées par Daryl Bem étaient exactement les mêmes que celles employées dans les recherches jusque-là jugées parfaitement respectables. L'existence de tels résultats était hautement problématique, puisqu'elle prouvait qu'il était possible de « démontrer » n'importe quoi, y compris des conclusions absurdes, en se servant de méthodes expérimentales traditionnellement employées en psychologie. Dans les années qui ont suivi, plusieurs articles ont naturellement échoué à répliquer les résultats de D. Bem.

Si la publication des expériences extra-sensorielles a provoqué un choc particulier dans la communauté scientifique, une série de scandales convergents a conduit ses membres à un examen de conscience beaucoup plus poussé. En 2011 toujours, un professeur réputé de l'université d'Amsterdam, Diederik Stapel, est révoqué pour avoir fabriqué ses propres résultats expérimentaux. Dans les années 2010, plusieurs expériences classiques de psychologie sociale échouent au test de la réplicabilité. L'ensemble de ces problèmes a donné lieu à la naissance d'un large mouvement de réforme, qui aboutit en 2015 à la publication du *Reproducibility Project : Psychology* (ou *RP:P*), évoqué en introduction.

Le *RP:P* suggère donc que la majorité des expériences de psychologie exagèrent très largement l'importance des résultats étudiés. Les effets reportés dans les articles originaux étaient en effet en moyenne deux fois plus importants que ceux obtenus par les réplications. Cela ne veut cependant bien sûr pas dire que l'ensemble de la recherche en psychologie est faussé. En 2014, un projet de reproduction de 13 expériences de psychologie sociale et cognitive, le *Many Labs Project*, a abouti à des résultats contrastés en matière de réplicabilité. Alors que les expériences du *RP:P* avaient été choisies dans le but d'obtenir un échantillon représentatif de la discipline, le *Many Labs Project* avait pris soin de sélectionner des classiques de la psychologie, afin d'étudier dans quelle mesure la reproduction de ces résultats dans de nouveaux laboratoires et sur des populations diverses était possible et aisée. Les tentatives de réplication ont été conduites dans 9 pays différents, sur 10 expériences classiques de la discipline, et 3 plus récentes. Tandis que l'équipe de chercheurs est parvenue à reproduire les

[5] Daryl J. Bem, « Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect », *Journal of Personality and Social Psychology, 100*(3), 2011, p. 407-425.
[6] Même si une telle tentative peut ressembler à un canular, Bem était parfaitement sérieux, et croyait véritablement en l'existence de pouvoir extra-sensoriels.

10 expériences classiques, aucune des trois expériences récentes n'a passé le test de la réplicabilité[7]. Il se peut qu'une majorité d'expériences en psychologie soient inexploitables, mais il existe bien un noyau dur de résultats fiables au sein de la discipline.

D'autres disciplines en sciences sociales ou bio-médicales connaissent des difficultés similaires. Le cas de la recherche bio-médicale est complexe, car des domaines où les recherches sont menées de façon relativement rigoureuse coexistent avec d'autres où prévaut un manque de rigueur assez net. Pour le dire vite, il existe un contraste gigantesque entre les recherches effectuées sur les êtres humains, généralement d'assez bonne qualité, et celles qui portent sur les animaux, parfois très approximatives. La recherche médicale effectuée sur les êtres humains fait l'objet d'une réglementation relativement stricte qui l'astreint à respecter des exigences scientifiques assez fortes, notamment dans les tests préalables à la mise sur le marché d'un médicament[8]. Cette dernière est préparée par plusieurs phases de tests, qui visent à éliminer les médicaments présentant un danger pour l'homme ou n'ayant pas fait la preuve de leur efficacité. À l'inverse, la recherche effectuée sur des animaux est très peu contrôlée, et les pratiques de recherche y sont souvent médiocres[9].

Cette situation conduit à un gâchis financier très important, étant donné que les médicaments testés sur l'homme le sont sur la base d'études faites sur des animaux. Si une molécule semble avoir démontré des effets prometteurs pour traiter un problème cardiaque chez des souris, un médicament basé sur cette molécule sera probablement développé et testé sur l'homme. Cependant, la recherche initiale, sur des populations animales, ayant été faite sans grande rigueur, les études portant sur l'homme échouent la plupart du temps à démontrer le moindre bénéfice du médicament en question. Ainsi, seulement 5 % des médicaments testés sur l'homme finissent par prouver leur efficacité et par être mis sur le marché[10].

Pourquoi la recherche est-elle aussi peu fiable ?

On peut distinguer 4 facteurs principaux à l'origine du manque de fiabilité des résultats publiés dans les revues de sciences sociales et de recherche bio-médicale : le fait que la culture scientifique accorde une importance exclusive aux résultats novateurs et « positifs », l'excessive flexibilité dans l'analyse statistique des résultats, l'absence de réplication des travaux antérieurs par des laboratoires autres que ceux ayant publié les études princeps, et la faiblesse des échantillons.

---

[7] Richard A. Klein et al., « Investigating Variation in Replicability: A "Many Labs" Replication Project », *Social Psychology, 45* (3), 2014, p. 142-152.
[8] Ceci reste une grande simplification : les problèmes de réplicabilité sont aussi extrêmement prégnants dans les études portant sur la santé humaine, même si la situation est bien meilleure que pour les études animales.
[9] David W. Howells, Emily S. Sena, & Malcolm R. Macleod, « Bringing Rigour to Translational Medicine », *Nature Reviews Neurology*, 10, 2014, p. 37–43.
[10] John Arrowsmith, « A Decade of Change », *Nature Reviews Drug Discovery*, 11, 2012, p. 17–18.

La culture scientifique dans les sciences sociales et bio-médicales accorde une forte valeur aux résultats novateurs et positifs. Les revues scientifiques préféreront publier des articles portant sur un nouveau phénomène, et tout particulièrement si ceux-ci montrent que ce phénomène est porteur d'applications pratiques décisives. La principale conséquence de cet état d'esprit est que les chercheurs sont incités à ne pas publier de résultats « négatifs », c'est-à-dire des résultats n'indiquant aucune différence significative entre divers traitements. Par exemple, dans le domaine médical, on préférera publier un article portant sur un traitement prometteur du cancer qu'un article décrivant comment un nouveau traitement échoue à le guérir. Cela a pour conséquence que seule une partie des expériences sont publiées, ce qui aboutit à ce qu'on appelle un « biais de publication ».

Un tel biais peut conduire à la publication exclusive d'articles démontrant une différence, même là où il n'en existe aucune. Imaginons que 20 laboratoires médicaux cherchent à déterminer si une molécule particulière peut guérir un cancer. Il arrivera nécessairement que certains laboratoires découvriront, par pur hasard, une relation positive entre ce traitement et la cure du cancer (par exemple, parce que les personnes traitées avaient au préalable un cancer moins agressif que les personnes dans le groupe témoin). D'autres découvriront peut-être une relation négative, donnant l'impression que le traitement aggrave la situation. En fin de compte, seul le laboratoire qui aura trouvé un effet fort du traitement sur la guérison du cancer pourra publier son article, parce que seul celui-ci semblera suffisamment intéressant. Les 19 autres laboratoires, à l'inverse, relégueront probablement leur étude au fond d'un tiroir, parce qu'ils savent que, si cette étude ne promet pas de guérir le cancer, elle n'est pas publiable.

La conséquence de telles pratiques est que la recherche publiée dans les revues scientifiques exagère fondamentalement l'efficacité des traitements médicaux, des interventions pédagogiques ou éducatives, ou, dans le domaine psychologique, l'influence de manipulations subtiles sur le comportement humain. La littérature scientifique se retrouve alors parsemée de résultats faux, ou, au minimum, très largement exagérés.

Ce biais de publication est d'autant plus problématique qu'il tend à inciter les chercheurs à « arranger » leurs résultats. Pour maximiser leurs chances de publication, ils peuvent avoir recours à des pratiques de recherche discutables, diminuant la rigueur de la recherche, mais maximisant les chances d'obtenir un résultat positif. Une première pratique discutable consiste à recruter des participants progressivement, en plusieurs étapes, dans le but d'arrêter l'expérience uniquement quand les résultats voulus ont été obtenus. Une deuxième pratique courante consiste à tester plusieurs conditions en même temps, dans le but de comparer ces multiples résultats possibles, avant de se concentrer sur la seule expérience qui a « marché ». Un troisième type de pratiques discutables consiste à effectuer des analyses sur une seule partie du groupe testé. Par exemple, dans la recherche bio-médicale, si des chercheurs obtiennent des résultats qui indiquent qu'un nouveau médicament n'est pas très efficace, ils peuvent chercher à subdiviser leurs analyses pour voir s'il n'existe pas un effet positif sur les femmes, les hommes, les personnes jeunes, âgées, etc. À force de diviser la population en

multiples sous-groupes, il y aura une forte chance que le médicament se trouve, par pur hasard, profiter à un groupe particulier, même s'il n'est en réalité bénéfique pour personne.

Ce manque de fiabilité de la recherche initiale ne serait pas en soi un problème gigantesque si la recherche scientifique se corrigeait au fil du temps. En science, une telle correction peut prendre la forme de tentatives de réplication d'expériences précédentes. Une telle pratique est extrêmement courante en physique, par exemple. Comme nous l'avons vu, répliquer une expérience antérieure est devenue relativement courant en psychologie depuis le début de la crise de la réplicabilité en 2011. Mais cette pratique était, jusque-là, quasiment inexistante : il était extrêmement difficile de parvenir à faire publier une réplication, puisque de tels articles n'étaient pas considérés comme suffisamment novateurs. En conséquence, des études parfaitement erronées ont pu continuer d'être citées pendant des décennies sans qu'aucune correction y soit apportée[11].

Dernier facteur majeur du manque de fiabilité, et non des moindres : la faiblesse des échantillons. Les expériences sont en effet conduites sur un nombre trop faible de sujets. Or plus un échantillon est important, plus grande est la précision dans l'estimation des résultats scientifiques. Dans le cas de la recherche animale, il est ainsi fréquent d'étudier l'efficacité de traitements de problèmes cardiaques sur moins de 10 animaux, groupe témoin inclus. Il est difficile d'imaginer comment des recherches sur un nombre de sujets aussi faible pourraient conduire à des résultats généralisables. Si, en psychologie, les recherches sur un échantillon aussi faible sont rares, la comparaison de groupes d'une vingtaine ou d'une trentaine de personnes a été la norme pendant des années. Certes, des groupes de cette taille peuvent être suffisants pour démontrer l'existence d'effets psychologiques extrêmement forts. L'effet *Stroop*, qui désigne la difficulté qu'éprouvent des sujets à prononcer un mot de couleur (par exemple, rouge) s'il est écrit dans une encre de couleur différente (par exemple, en vert) est un effet massif, extrêmement robuste et facilement démontrable avec un nombre très faible de participants. Mais toute démonstration rigoureuse d'effets plus subtils — et il s'agit probablement de la majorité des effets étudiés en psychologie — est tout simplement impossible avec des effectifs aussi réduits.

[11] Parmi les exemples d'expériences non réplicables qui ont eu un énorme impact, on peut par exemple citer l'expérience conduite par John Bargh et ses collègues en 1996, qui a tenté de montrer que présenter des mots associés à la vieillesse conduisait des sujets jeunes à adopter un comportement typique de personnes âgées, à savoir le fait de marcher lentement. L'échec de la tentative de réplication par Doyen et collègues en 2012 a été un des facteurs à l'origine de la crise de la réplicabilité. Voir S. Doyen, O. Klein, C-L. Pichon, A. Cleeremans, « Behavioral Priming : It's All in the Mind, but Whose Mind? », *PLoS ONE*, 2012, 7(1): e29081.

Que faire ? La réforme méthodologique actuelle

Face à l'étendue des problèmes de réplicabilité, la communauté des chercheurs en psychologie a récemment entamé des réformes méthodologiques profondes, d'une ampleur inédite en sciences sociales. Le premier impact, déjà évoqué, est la multiplication de projets de réplications. S'il s'agit d'une réforme cruciale, elle est loin d'être la seule.

Un deuxième courant de réforme a visé à augmenter la précision des mesures expérimentales, notamment par le biais d'une augmentation de la taille échantillons. Même si de nombreuses voix se sont élevées pour apporter leur soutien à ce projet, les progrès ont jusqu'ici été limités. Un des principaux obstacles à la réalisation d'expériences à grande échelle réside bien sûr dans le coût de telles expériences ; peu de laboratoires sont en effet en mesure de faire passer la même expérience psychologique à plus de 1000 participants différents. Pour résoudre cet obstacle, des méthodes de recrutement différentes, sur internet, se sont développées ces dernières années. Ainsi, une grande partie des enquêtes effectuées récemment l'ont été via un service proposé par Amazon, *Amazon Mechanical Turk*, par lequel il est possible de recruter un grand nombre de participants en échange d'une compensation financière. Si le dispositif permet effectivement d'augmenter le nombre de participants, la représentativité de l'échantillon recruté pose question : les participants sont essentiellement indiens ou états-uniens, et, dans le cas des États-Unis, ils sont généralement légèrement plus jeunes, plus éduqués, et plus à gauche que la population états-unienne dans son ensemble[12]. Cependant, toutes les expériences ne peuvent naturellement pas se faire à distance. Une proposition majeure est de développer les collaborations entre laboratoires, afin de procéder à une mutualisation des ressources. C'est un tel développement qui a permis d'améliorer la fiabilité des recherches en génétique : alors que ces travaux étaient d'une qualité médiocre jusqu'au début des années 2000, le développement de collaborations internationales a conduit à une amélioration drastique de la qualité de la recherche en ce domaine. Deux projets de collaboration internationale, en psychologie et en neurosciences, ont ainsi vu le jour au cours de l'année 2017[13].

Un troisième aspect majeur de la réforme méthodologique en cours cherche à remédier aux problèmes engendrés par le manque de rigueur avec lequel sont analysées les données. Comme nous l'avons vu, les chercheurs disposent généralement d'une grande marge de manœuvre dans l'analyse des données, qui rend possible après coup une interprétation favorable des résultats d'une expérience. Pour lutter contre cela, l'idée d'enregistrement

[12] La majorité des articles de psychologie sont publiés à partir de l'étude des participants états-uniens, qu'il s'agisse d'étudiants ou de membres *d'Amazon Mechanical Turk*. Au delà de la représentativité des participants à *Amazon Mechanical Turk* par rapport à la population états-unienne en général, savoir dans quelle mesure il est possible de construire des théories psychologiques universelles à partir de l'étude des habitants d'un seul pays est bien sûr un problème. Les dernières années ont connu de nombreux appels en faveur du développement d'une psychologie interculturelle. Voir J. Henrich, S. Heine, et A. Norenzayan, « The Weirdest People in the World ? », *Behavioral and Brain Sciences*, 2010, 33 (2-3), p. 61-83.

[13] Dalmet Singh Chawla, « A New "Accelerator" Aims to Bring Big Science to Psychology », *Science*, 2017.

préalable gagne du terrain. Cette méthode contraint le chercheur à annoncer en amont de son expérience les hypothèses qu'il veut tester, ainsi que la manière dont il compte analyser ses résultats. La publication préparatoire est ensuite chargée sur un site internet qui gèle le document, empêchant toute modification ultérieure. Après la réalisation de l'expérience, le chercheur est ainsi obligé de laisser parler les résultats, sans pouvoir procéder à une analyse discutable confirmant *in fine* sa théorie. L'enregistrement préalable est une pratique courante dans les études médicales appliquées à l'homme ; sa généralisation aux sciences sociales et aux études animales représenterait un très net progrès[14].

Dans la même logique, un nombre croissant de revues scientifiques propose aux chercheurs de soumettre leurs articles avant d'avoir conduit leurs expériences. Cette nouvelle méthode de publication, dite des « rapports enregistrés » (*registered reports*), permet aux relecteurs anonymes d'évaluer l'expérience sur la seule base de sa méthode, indépendamment du caractère positif ou non des résultats obtenus. De nombreuses revues de psychologie sociale et cognitive proposent désormais ce mode de publication.

Ces mouvements de réforme ne constituent qu'une partie des propositions avancées pour améliorer la fiabilité des publications scientifiques. De nouvelles organisations, comme le *Center for Open Science* ou la *Society for the Improvement of Psychological Science*, ont été créées pour accélérer le mouvement. De tels appels à la réforme ont bien sûr pu susciter une certaine réticence, voire une certaine opposition, au sein de la communauté scientifique. Certaines voix se sont élevées pour défendre la manière traditionnelle de pratiquer la psychologie et les sciences sociales quantitatives. Ces partisans du *statu quo* défendent par exemple l'idée que la flexibilité dans l'analyse des résultats promeut la créativité des chercheurs, qui risquerait d'être étouffée par le surcroît de rigueur exigé par les réformateurs. D'autres chercheurs ont également insisté sur le fait que l'argent et le temps consacrés à effectuer des réplications d'expériences antérieures seraient mieux employés dans des recherches portant sur de nouveaux domaines. Si de telles critiques ne sont évidemment pas absurdes, elles restent, cependant, difficiles à accepter dans un contexte où la recherche en sciences sociales peine souvent à produire des résultats crédibles.

Pour aller plus loin

Ouvrages généraux sur la question (psychologie et sciences bio-médicales)

- Chris Chambers, *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*, Princeton University Press, 2017.

[14] Notons cependant que la pratique de l'enregistrement préalable n'est pas toujours idéale. D'une part, de nombreuses prédictions sont souvent floues, ce qui laisse une certaine marge de manœuvre dans l'analyse des résultats. D'autre part, il arrive que des chercheurs mentent par rapport à leurs prédictions, afin de maximiser leurs chances de publier l'article en question. Dans ce dernier cas, l'existence de l'enregistrement préalable permet cependant de prouver que l'auteur de l'article a effectué subrepticement des analyses en contradiction avec ses propres prédictions.

- Richard Harris, *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*, Basic Books, 2017.

Articles et ressources statistiques

- Blog de Daniël Lakens, influent et pédagogique sur les questions de réplicabilité (http://daniellakens.blogspot.ca/).
- Leif D. Nelson, Joseph Simmons et Uri Simonsohn, « Psychology's Renaissance », *Annual Review of Psychology*, 69, 2018.
- Joseph P. Simmons, Leif D. Nelson et Uri Simonsohn, « False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant », *Psychological Science*, 22, 2011, p. 1359–66.

Publié dans laviedesidees.fr, le 20 mars 2018.