Quelle est la fiabilité des méta-analyses relatives à la psychologie et à des disciplines apparentées ? Une investigation de la qualité d'un échantillon de méta-analyses publiées sur PsycINFO en 2016

Auteur : Ajamieh, Sara
Promoteur(s) : Tirelli, Ezio; Bruyere, Olivier
Faculté : þÿ F a c u l t é d e P s y c h o l o g i e , L o g o p é d i e e t S c i e n c e s d e l E d u c a t i o n
Diplôme : Master en sciences psychologiques, à finalité spécialisée en neuroscience cognitive et comportement
Année académique : 2017-2018
URI/URL : http://hdl.handle.net/2268.2/5804

Quelle est la fiabilité des méta-analyses relatives à la psychologie et à des disciplines apparentées ? Une investigation de la qualité d'un échantillon de méta-analyses publiées sur PsycINFO en 2016

Sara Ajamieh

Promoteur : Pr Ezio Tirelli
Co-Promoteur : Pr Olivier Bruyère
Lecteurs : Pr Etienne Quertemont et Dr Jessica Simon

Mémoire réalisé en vue de l'obtention du grade de Master en sciences psychologiques

Faculté de psychologie, logopédie et sciences de l'éducation
Année académique 2017-2018

Préambule

Aussi loin que ma mémoire peut remonter, j'ai toujours été intéressée par les résultats et par la méthode de la science. Au fil du temps, et au cours d'un processus qui a commencé son œuvre peu avant mon entrée à l'université, j'ai été de plus en plus captivée par la seconde, et par toutes les menaces qui peuvent lui être adressée. Ainsi, lorsque j'ai appris il y a un peu plus d'une année qu'un doctorat de méta-recherche était en cours, j'ai sauté sur l'occasion pour tenter de réaliser un mémoire dans ce domaine.

C'est Monsieur Ezio Tirelli qui m'a aiguillé vers le projet en question, et j'ai finalement été accueillie par Monsieur Olivier Bruyère et Mademoiselle Victoria Leclercq qui ont accepté que je collabore à ce projet.

Je leur suis à tous très reconnaissante, ainsi qu'à Monsieur Etienne Quertemont et Madame Jessica Simon qui ont pris le temps de lire et de critiquer ce mémoire.

Table des matières

1) Introduction

La méthode scientifique a été développée au fil des derniers siècles par et pour l'ambition d'une description rationnelle, c'est-à-dire argumentée par les faits, des effets de la nature. Les acteurs du monde scientifique peuvent aujourd'hui revendiquer l'efficacité de cette méthode qui permet l'étude de la réalité observable au plus proche de ses véritables fondements.

Nous faisons tous les jours l'expérience de la puissance de la science – y a-t-il un seul aspect de notre quotidien qui n'ait pas été concerné et transformé par la marche des connaissances ? Ces évolutions tiennent à la construction et à l'application d'une méthode hypothético-déductive qui décrit plusieurs temps à la construction du savoir : une hypothèse est formulée à propos d'un phénomène particulier, une expérience est menée et les résultats supportent – ou ne supportent pas – l'hypothèse de départ. Le résultat, quel qu'il soit, permet de construire, de modifier, et parfois même de rejeter la théorie dont l'hypothèse est issue, ce qui soulève de nouvelles questions et de nouvelles études, et amène à une compréhension de plus en plus fine de la théorie sous-jacente (Chambers, 2017).

Malheureusement, la culture de la production scientifique, telle qu'elle est aujourd'hui, ne permet pas toujours d'honorer pleinement la promesse contenue dans cette méthode, et donc de construire des théories fiables. Des formes de pratiques de recherche discutables s'immiscent à plusieurs de ses niveaux – elles sont filles d'un contexte régit par une pression à la publication de résultats à la fois « originaux » et « significatifs », dont la réalité est maintenant bien reconnue par l'ensemble du monde académique. Cette pression incite en effet des chercheurs à arranger, plus ou moins consciemment, les résultats de leurs recherches afin d'obtenir un matériel « publiable » selon les critères de nombreux journaux scientifiques. Cet état de fait, avec d'autres problèmes méthodologiques courants, a des conséquences assez importantes pour occuper aujourd'hui une partie de la communauté scientifique à l'étude et, tant que possible, à la résolution de ces pratiques.

Le projet dont ce mémoire est issu s'inscrit directement dans le travail de cette communauté ; il constitue en fait une excellente opportunité pour fournir une présentation articulée de ces problématiques. Quoi de mieux, en effet, qu'une investigation de la qualité de méta-analyses

pour brosser une *synthèse* de cette « méta-recherche » ! L'évaluation de la qualité des méta-analyses, quel que soit le domaine considéré, est en effet indissociable de l'évaluation plus globale de la qualité de la littérature, pour au moins deux raisons. Premièrement, les études sujettes à des pratiques de recherche discutables et à des problèmes méthodologiques déforment la qualité des effets publiés ; ceux-ci sont par conséquent susceptibles d'influencer fallacieusement les résultats des méta-analyses lorsque celles-ci les incorporent. Ensuite, les méta-analyses sont elles-mêmes vulnérables à l'exercice de ces pratiques. Il reste ainsi important de cerner, dans un premier temps, la nature et l'ampleur des problèmes qui mettent en péril la fiabilité de la littérature scientifique, en mettant l'accent sur la situation en psychologie, pour nous intéresser ensuite à la qualité des méta-analyses et à quelques facteurs censés y être relatifs.

Mais plantons d'abord un tant soit peu le décor…

La pression à la publication et la quête de la nouveauté

Le mois de septembre 2011 a été certainement été riche en émotions pour la communauté des physiciens. Deux cents scientifiques de l'expérience OPERA avaient alors annoncé que des particules subatomiques, appelées neutrinos, pouvaient voyager à une vitesse plus grande que celle de la lumière. Si cette découverte était avérée, c'était l'ensemble de la physique einsteinienne, qui n'admet pas de tels phénomènes, qui allait devoir être remise en question. Des essais de réplication furent aussitôt réalisés, mais aucune de ces expériences n'a réussi à reproduire ce curieux résultat. Au mois de juillet 2012, l'équipe originale est revenue sur ses conclusions : ces résultats avaient été induits… par un GPS mal branché (Bronner, 2013).

Cette anecdote est intéressante en ce qu'elle met en évidence que les effets d'une pression à la publication peuvent se manifester dans n'importe quel corps de métier dédié à la délivrance d'informations. Ici, en l'occurrence, c'est à la fois les physiciens à l'origine de cette « découverte » et les journalistes qui l'on relayée qui ont fait les frais des effets de la concurrence dans leurs domaines respectifs ; les seconds n'ont en effet pas manqué de propager largement l'information dans la presse, à l'heure où sortir des titres fracassants est plus que jamais une façon de subsister à l'époque d'internet et du flux titanesque des

informations qui y circule. Les premiers ont, pour leur part, fait preuve d'une énorme imprudence au vu des conséquences théoriques de leur proposition – Mais pouvaient-ils, avant la découverte de leur erreur, ne fut-ce que considérer passer à côté de la renommée consubstantielle à ce genre de découverte ? Bronner (2013) s'est intéressé à ces dérives, et a proposé une modélisation qui lie la pression issue d'une forme de concurrence à la fiabilité de l'information diffusée dans un corps de métier donné (figure 1) : si la concurrence a des effets positifs sur la fiabilité de l'information lorsqu'elle s'exerce de façon raisonnable (les effets du manque de concurrence peuvent être exemplifiés à l'extrême, dans le domaine journalistique, par la tutelle qu'imposent les régimes dictatoriaux sur la diffusion des informations), cette fiabilité tend à décroitre lorsque cette concurrence devient effrénée.

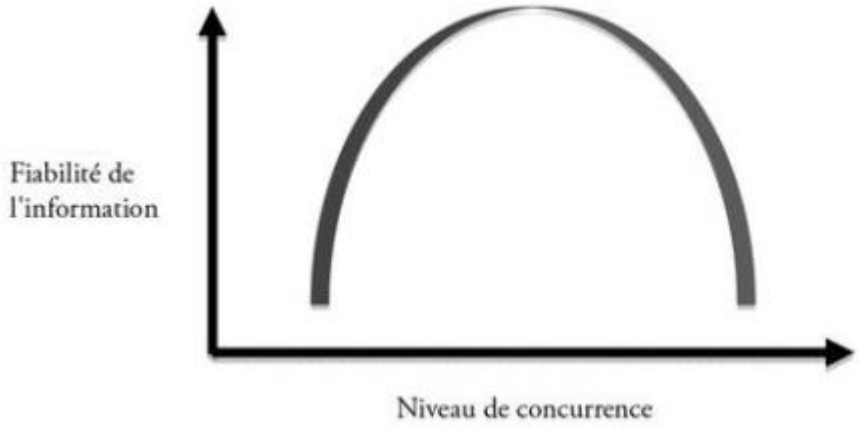


Figure 1 : lien entre la fiabilité de l'information et le niveau de concurrence. Tiré de Bronner (2013).

Dans le domaine scientifique, les avatars de la concurrence à la publication se manifestent clairement dans la politique de beaucoup de journaux, qui acceptent préférentiellement des résultats « originaux » et statistiquement significatifs. Par exemple, les conditions du journal *Brain* préviennent les auteurs que certains articles sont « rejetés sans revue par les pairs pour manque de nouveauté » (Oxford University Press, 2018a) celles du journal *Cerebral Cortex* précisent que « l'acception finale des articles ne dépend pas uniquement du mérite technique, mais d'un jugement subjectif de la nouveauté » (Oxford University Press, 2018b), et *Psychological Inquiry* souligne que les « manuscripts (…) devraient présenter des théories nouvelles, vastes, provocantes » (Taylor and Francis Online, 2018). Cette concurrence est, de façon beaucoup plus large, visible à travers l'existence d'indices bibliométriques censés

refléter l'impact d'un journal ou d'un chercheur (comme le facteur d'impact ou l'index H), ou dans les avantages dont les chercheurs les plus prolifiques bénéficient en termes de carrière. Lorsque cette concurrence est adressée par des sondages, les résultats font ressortir la conscience d'un problème, voire d'un mal-être. Par exemple, un sondage réalisé chez 315 scientifiques travaillant dans le milieu médical, et issus de 5 universités flamandes, a révélé que 72% d'entre eux estimaient la pression à la publication comme étant trop excessive ; à une autre question, 52% des sondés ressentaient leur évaluation par leurs collègues (sur base de leurs publications) comme un phénomène stressant. Cependant, seuls 52% des sondés pensaient que la pression à la publication était délétère pour la science (Tijdink, Verbeke, & Smulders, 2014). Un autre sondage réalisé chez 140 stagiaires américains, travaillant également dans le milieu biomédical, a révélé qu'un cinquième d'entre eux percevaient une pression à publier des résultats incertains ; un tiers rapportait se sentir obligé de supporter l'hypothèse d'un supérieur même quand les données n'allaient pas dans le sens de celle-ci, et près de la moitié rapportait avoir connaissance d'un supérieur qui requérait aux membres de l'équipe la publication d'un article à haut impact pour compléter leur formation dans leurs laboratoires (Begley, Buchan, & Dirnagl, 2015). De tels résultats sont certainement généraux à la pratique de la publication scientifique, telle qu'elle est entretenue aujourd'hui.

Si nous suivons la proposition de Bronner (2013), cette pression devrait avoir, d'une façon ou d'une autre, un impact négatif sur la fiabilité des résultats de la recherche, mais qu'en est-il vraiment ? Cette question est du ressort d'un domaine de recherche qui est actuellement en plein développement : la méta-recherche. Cette discipline est définie comme l'étude de la recherche en elle-même, aussi bien en ce qui concerne ses méthodes, la publication de ses résultats, sa reproductibilité, ses évaluations et ses incitations, avec l'objectif de la comprendre et de l'améliorer (Ioannidis, 2018). L'émergence de cette discipline a tenu à l'identification de traditions de recherche et de biais qui peuvent être plus ou moins répandus selon les domaines ; il peut s'agir par exemple du type de design ou de statistiques utilisées, de pratiques de recherche discutables plus ou moins répandues, de la consistance et de la qualité des effets publiés, de l'accent placé sur la formation méthodologique et statistique des chercheurs ou encore de la culture de la réplication. La méta-recherche peut permettre d'améliorer la qualité de l'entreprise scientifique et de ses résultats, et à fortiori les conséquences très pratiques de la recherche de mauvaise qualité ; dans certains champs, ce

sont notamment des sommes d'argent colossales qui sont gâchées dans la conduction d'études fondamentalement peu fiables (Ioannidis, 2018).

Des études de la méta-recherche ont ainsi été menées pour adresser la qualité de la littérature dans différents domaines de la science, dont la psychologie – et les résultats sont globalement interpellants. Les données rapportées permettent déjà de classer les problématiques soulevées par ces études en 3 grands phénomènes, eux-mêmes à la base d'autres dysfonctionnements qui affectent la qualité des connaissances : le biais de publication, les pratiques de recherche discutables et le manque de puissance.

Le biais de publication

Le biais de publication est la tendance à la publication de résultats statistiquement significatifs au détriment de ceux non significatifs, et a été largement objectivé depuis la publication princeps de Rosenthal (1979) qui le désignait sous l'expression de « file drawer problem ». Par exemple, en psychologie, Kühberger, Fritz, et Scherndl (2014) ont mis en évidence un biais de publication sur base de l'analyse d'un échantillon de 341 articles publiés sur PsycINFO en 2007, pour lesquels les valeurs de la statistique Z (dérivées des p-valeurs rapportées dans les articles) étaient trois fois plus fréquentes juste au-dessus du score Z correspondant au seuil classique de significativité ($Z = 1,96$, $p = 0,05$), par rapport à la fréquence des valeurs se trouvant juste en dessous de ce seuil (figure 2).

Il existe d'autres preuves du biais de publication. En ce qui concerne les sciences sociales de façon générale, Franco, Malhotra, et Simonovits (2014) ont pu récupérer les projets de financements du programme « Time-sharing Experiments in the Social Sciences » de la National Science Fondation, qui concernaient des études de sondages portant sur différents sujets. Les auteurs ont examiné les résultats statistiques de 221 études publiées entre 2002 et 2012, qui avaient été publiées ou non publiées. Les résultats ont révélé que 64,6% des études rapportant des résultats nuls n'ont même pas été rédigées ! A titre de comparaison, seules 12.2% des études rapportant des résultats « mitigés » n'ont pas été rédigées, ainsi que 4.4% des études rapportant des résultats « forts » (figure 3). Les chercheurs qui ont obtenu des résultats nuls et qui n'ont pas rédigé leur manuscrits justifiaient fréquemment leur

comportement par le fait qu'ils pensaient que leurs résultats n'avaient aucune chance d'être publiés.

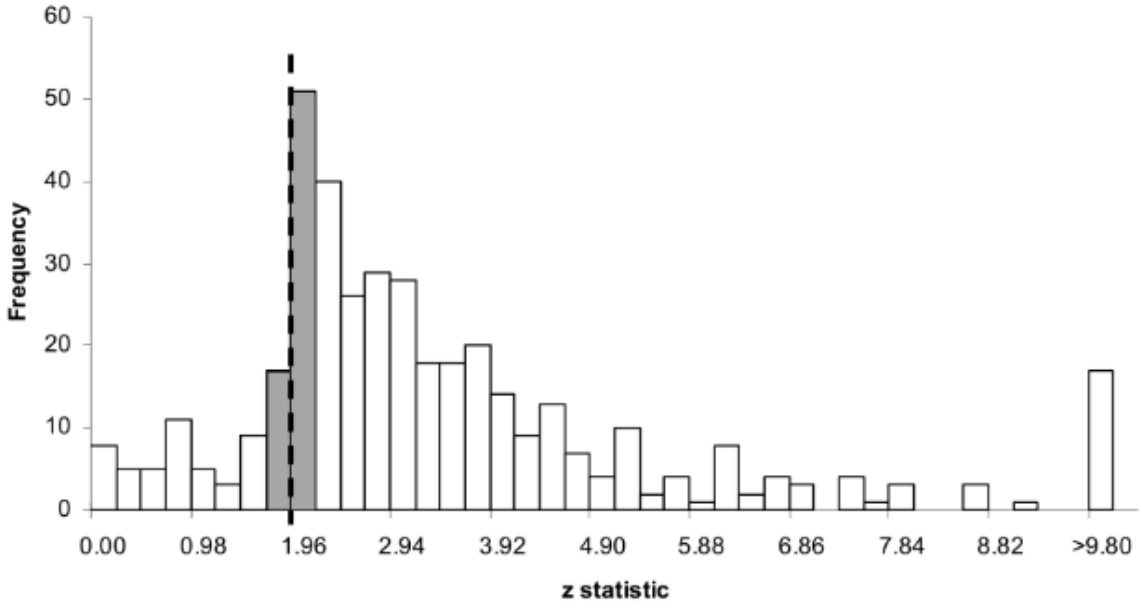


Figure 2 : Fréquence des statistiques Z, dérivées des p-valeurs de 341 articles en psychologie publiés en 2007. La largeur des colonnes correspond à un intervalle Z de 0.245 (12.5% de 1.96). La ligne pointillée correspond au seuil de significativité p = .05 (Z = 1.96). Tiré de Kühberger, Fritz, & Scherndl, (2014).

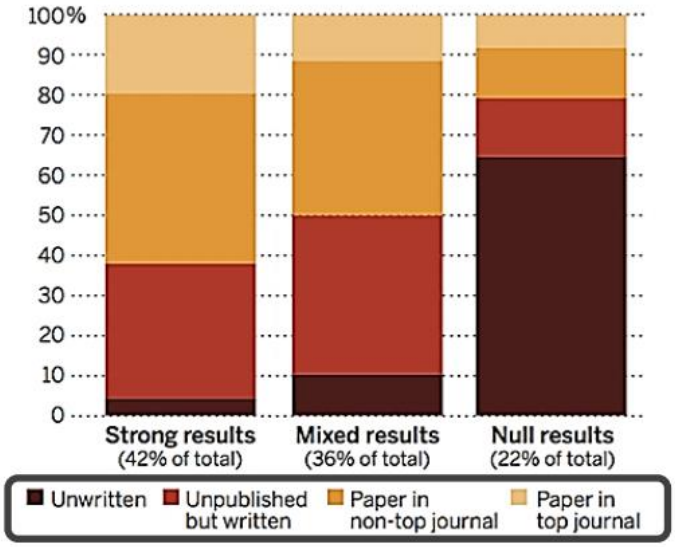


Figure 3 : La majorité des résultats « nuls » en sciences sociales ne sont même pas rédigés. Adapté de Franco, Malhotra, & Simonovits (2014).

On peut encore mettre en évidence un biais de publication en examinant leur nombre de résultats « positifs » publiés dans un domaine : Fanelli (2010) a récupéré 2434 articles représentant 20 disciplines, et analysé la proportion de résultats positifs rapportés dans chacune d'elles. Les résultats ont révélé que les sciences de l'espace avaient le plus faible taux de résultats positifs (70,2%) ; la psychiatrie et la psychologie, considérés ici en un seul domaine, occupaient pour leur part la tête du classement avec 91,5% de résultats positifs (figure 4) !

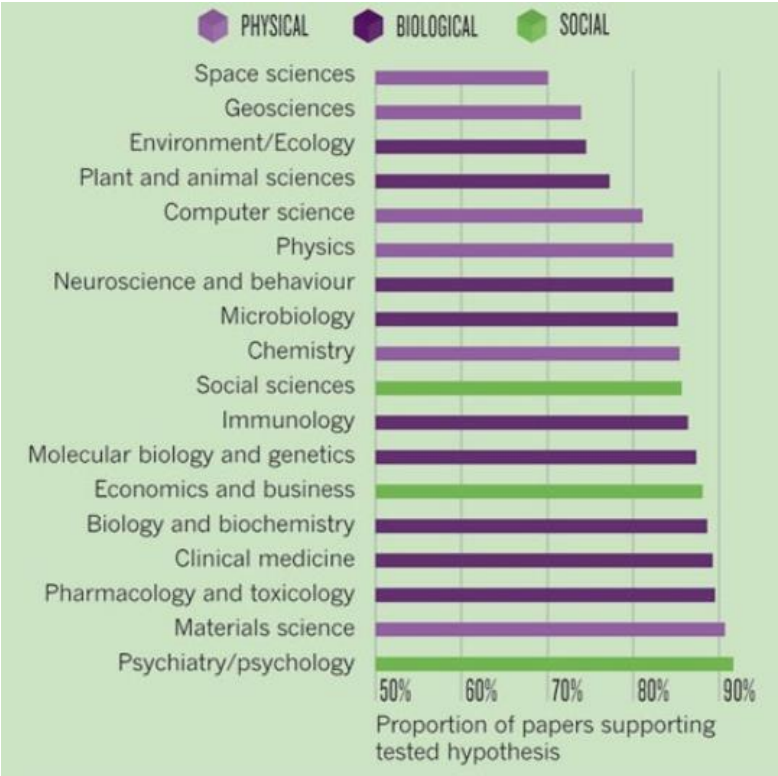


Figure 4 : Proportion des résultats « positifs » selon différentes disciplines scientifiques. Adapté de Fanelli (2010) et tiré de Yong (2012).

Le biais de publication déforme évidemment la représentativité et la fiabilité des résultats de la littérature, mais il n'est pas le seul à jouer son rôle délétère. Nous allons adresser à présent un grand problème de la littérature en psychologie : le manque de puissance.

Le manque de puissance

La puissance fait référence à la probabilité qu'un test statistique a de rejeter correctement l'hypothèse nulle quand celle-ci est fausse (Ellis, 2010). La qualité d'une investigation scientifique dépend donc, au-delà de la qualité de sa méthodologie, d'une réflexion effectuée en amont du projet sur cette problématique. Concrètement, il s'agira d'abord de poser une hypothèse sur la taille d'effet en jeu sur base de la littérature préexistante, et de choisir la puissance de l'expérience, que l'on fixe classiquement à 80%. Ces deux éléments permettent de déterminer les effectifs aptes à détecter la taille d'effet postulée, à un niveau alpha donné (classiquement égal à 0,05). Avec une puissance de 80%, on est donc en position de trouver un effet significatif lors de 80 tests sur 100.

Les notions de taille d'effet, de puissance statistique et d'effectifs sont en relation de telle manière qu'à une puissance donnée, les grands échantillons détectent davantage de tailles d'effet que les échantillons plus petits (Lipsey, 1990). Plus précisément, pour la même puissance, un grand échantillon sera capable de détecter de plus petites tailles d'effet que les échantillons plus modestes ; inversement, les petits échantillons ne pourront détecter que les plus grandes tailles d'effet (figure 5). Le corolaire pratique de cet état de fait est qu'il vaut mieux prendre le temps d'estimer la taille d'effet que l'on désire mettre en évidence dans une expérience ; si celle-ci se révèle être modeste et que l'on désire atteindre une puissance raisonnable, il faudra mobiliser de grands effectifs pour maximiser les chances que le test statistique ressorte significatif. Ce dernier point est à la base d'une troisième grande problématique touchant à la fiabilité de la littérature scientifique, qui représente certainement le plus grand problème méthodologique dans le domaine de la psychologie : beaucoup d'expériences ne sont pas en réalité assez puissantes pour pouvoir mettre en évidence les effets qu'elles rapportent dans la littérature.

Cette problématique a été soulevée depuis des décennies (e.g. Cohen, 1962) et reste largement d'actualité, comme en attestent de nombreuses études touchant à différents domaines de la psychologie. Par exemple, Fraley et Vazire (2014), ont évalué la puissance des études publiées dans différents journaux à haut impact relatifs à la psychologie sociale et à la

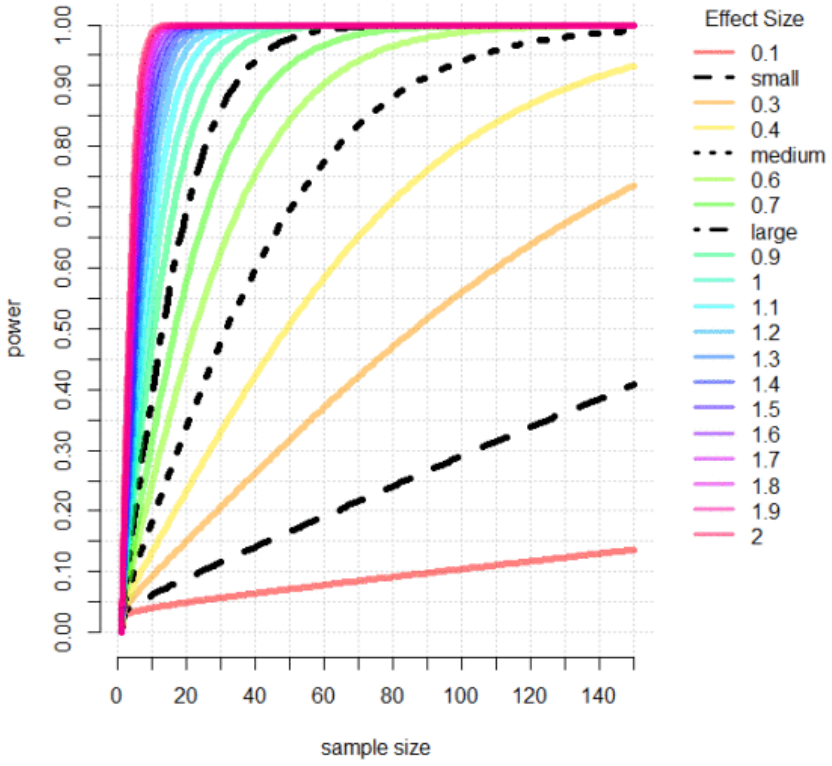


Figure 5 : Lien entre la puissance, la taille de l'échantillon et les tailles d'effet. Les tailles d'effet sont ici exprimées en d de Cohen, et sont indiqués en pointillés noirs les cut-offs correspondant à de petits, de moyens et de grands effets. Ces données sont relatives à un test t comparant deux moyennes.
On peut voir que les petits échantillons ne détectent que de grands effets. Par exemple, à 20 sujets par groupes, le test ne détecte que des tailles d'effet supérieures à d = 0.9 à une puissance de 80%. A titre de comparaison, avec 100 sujets par groupes, le test détecte des tailles d'effets supérieures à d = 0.4 à une puissance de 80%. Tiré de Grapov (2013).

psychologie de la personnalité, à partir des tailles d'échantillons typiquement rapportées par chacun de ces journaux et pour différentes tailles d'effet. Les résultats ont montré que les études de ces journaux ne sont pas assez puissantes pour détecter adéquatement une taille d'effet moyenne (r = .20, ou d = 41), alors même qu'il s'agit de l'ampleur de la taille d'effet typique dans ces domaines selon une méta-analyse compilant plus de 25.000 études (Richard, Bond, & Stokes-Zoota, 2003) ; la taille d'effet en question rapportée dans cette étude est égale à r = .21, ou à d = .43). Par exemple, la puissance des études de quatre des six journaux étudiés n'excédait pas 50% pour la détection de cette taille d'effet, ce qui signifie que si l'hypothèse nulle est fausse et que la taille d'effet populationnelle est égale à d = .41, les études publiées dans ces journaux ne détectent pas mieux cet effet qu'un test basé sur un lancer de pièce « pile ou face » ! La puissance estimée d'autres champs de recherche est encore plus basse : La puissance médiane des études en neuroscience a été estimée à 21%, ; celles utilisant plus

spécifiquement des modèles animaux a été estimée à 18% pour les études utilisant un labyrinthe aquatique, et à 31% pour celles incluant un labyrinthe radial dans des études évaluant la différence de performance entre des rats mâles et femelles à ces tâches ; et celles de neuroimagerie structurelle et volumétrique à 8% (Button et al., 2013). On peut encore porter en exemple les études de stimulation transcrânienne à courant direct, dans lesquelles la puissance a été estimée à 14% pour les études portant sur l'effet de la stimulation sur des processus cognitifs au sens large du terme, et à 5% pour les études portant spécifiquement sur la mémoire de travail (un résultat qui aurait été retrouvé sur base de données générées de façon aléatoire) (Medina & Cason, 2017).

Quelles sont les conséquences du manque de puissance dans un domaine de recherche? Une première réponse est qu'une littérature à la fois affectée par un biais de publication et comportant des études trop peu puissantes rapporte une plus grande proportion de faux positif (Fraley & Vazire, 2014), et plus globalement des résultats dont la représentativité, la robustesse et la magnitude sont déformés de leur véritable signification. Ce dernier point est parfaitement illustré par le phénomène dit du « winner's curse » (Young, Ioannidis, & Al-ubaydli, 2008), et mérite que l'on s'y attarde. Nous savons que les petites études ne peuvent détecter que de grandes tailles d'effet. Dans un contexte dans lequel une étude « publiable » doit dépasser un seuil de signification statistique, les effets publiés de telles études ont ainsi tendance à biaiser la littérature, car leur magnitude est exagérée par rapport celle des effets réels (Button et al., 2013). Des simulations montrent ainsi que pour différents niveaux de puissance, les tailles d'effet « capturées » par le seuil de significativité surestiment les tailles d'effet réelles, et que ce phénomène s'accentue d'autant plus que la puissance est faible (Maiväli, 2015 ; figure 6). Le winner's curse affecte ainsi les tailles d'effet méta-analytiques lorsqu'elles sont calculées uniquement sur base de la littérature publiée.

Enfin, le manque de puissance peut impacter directement la réplicabilité de la recherche : la tentative de réplication d'une étude trop peu puissante, peu importe la véracité des résultats qu'elle rapporte, risque d'échouer si le réplicateur utilise le même design et la même taille d'échantillon que l'étude originale, rendant le résultat de la réplication potentiellement ambigu (Fraley & Vazire, 2014).

Le biais de publication et le manque de puissance ne sont malheureusement pas les seuls à devoir être pris en compte pour juger de la qualité de la littérature publiée. Les chercheurs peuvent en effet être motivés, plus ou moins consciemment, à orienter les résultats de leurs travaux de façon à être publiés plus facilement dans un contexte régit par une compétition à la publication, ce qui déforme encore davantage la fiabilité de la recherche. Nous allons maintenant traiter de ces comportements.

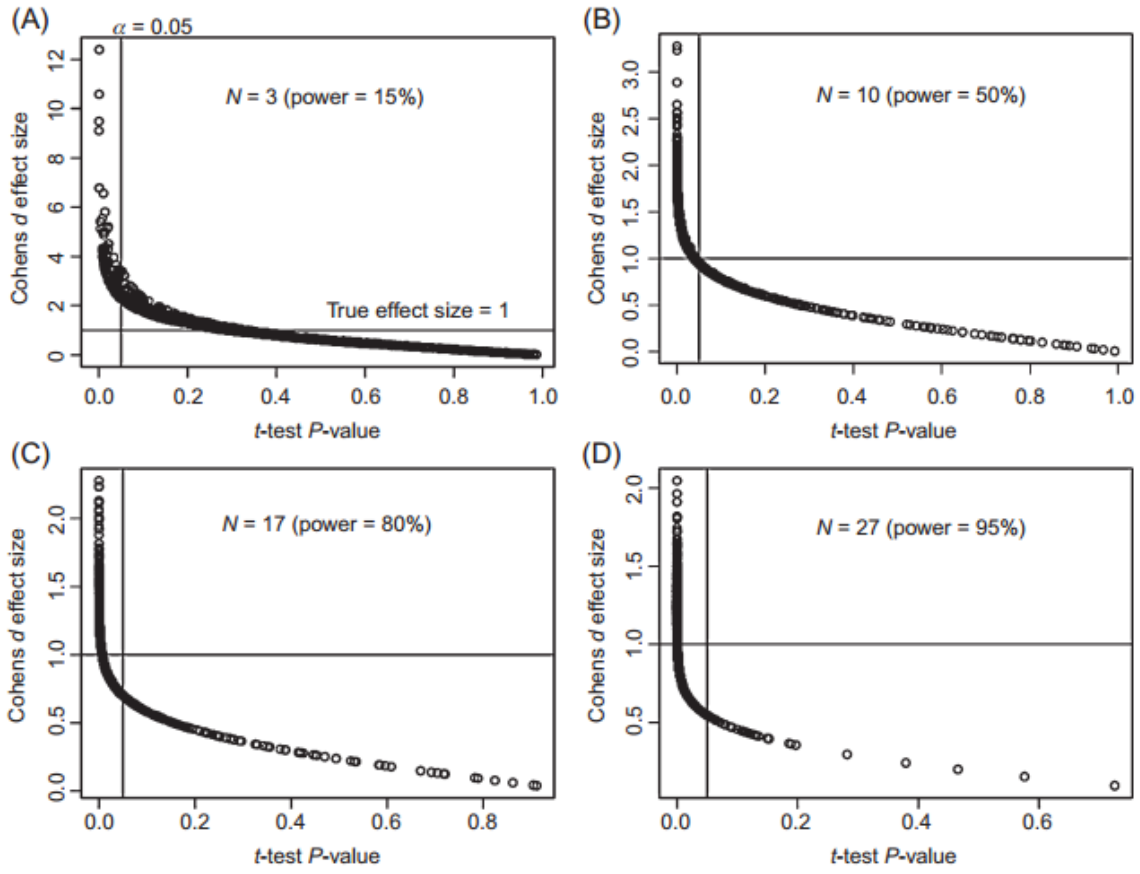


Figure 6 : Distribution des tailles d'effet (axes verticaux) selon la p-valeur (axes horizontaux) et la puissance (une par graphique). La figure montre que l'adoption d'un seuil de sélection (ici, alpha = 0.05) peut mener à l'inflation des tailles d'effet : c'est le "winner's curse". Deux populations dont les effets sont normalement distribués ont été générées (moyenne = 0, SD = 1 et moyenne = 1, SD = 1). De chaque population ont été extraites 1000 échantillons de quatre tailles différentes (qui correspondent ainsi à chacun des quatre graphiques), et qui ont été comparées entre elles avec un test t. La taille d'effet véritable (1.0) et le niveau de significativité (0.05) sont représentés respectivement avec des lignes horizontales et verticales, et chaque point correspond à une comparaison. Plus la puissance est faible, plus la magnitude du winner's curse augmente, c'est-à-dire que la surestimation des tailles d'effets capturées par le seuil de significativité devient de plus en plus importante. Tiré de Maiväli (2015).

Les pratiques de recherche discutables

Les pratiques de recherche discutables regroupent une série de comportements dont les effets contribuent également à déformer la réalité des découvertes. Des exemples célèbres incluent des chercheurs comme William Summerlin, qui a affirmé avoir mis au point une technique de greffe de tissus entre espèces qui n'entraînait aucun rejet (Broad & Wade, 1994), Diederik Stapel, qui a inventé et manipulé des données relevant de 58 publications dans le domaine de la psychologie sociale (Retraction Watch, 2015), ou Haruko Obokata, qui a prétendu avoir inventé une méthode révolutionnaire pour créer des cellules souches pluripotentes à partir de cellules somatiques différenciées (Hartoupian, 2016).

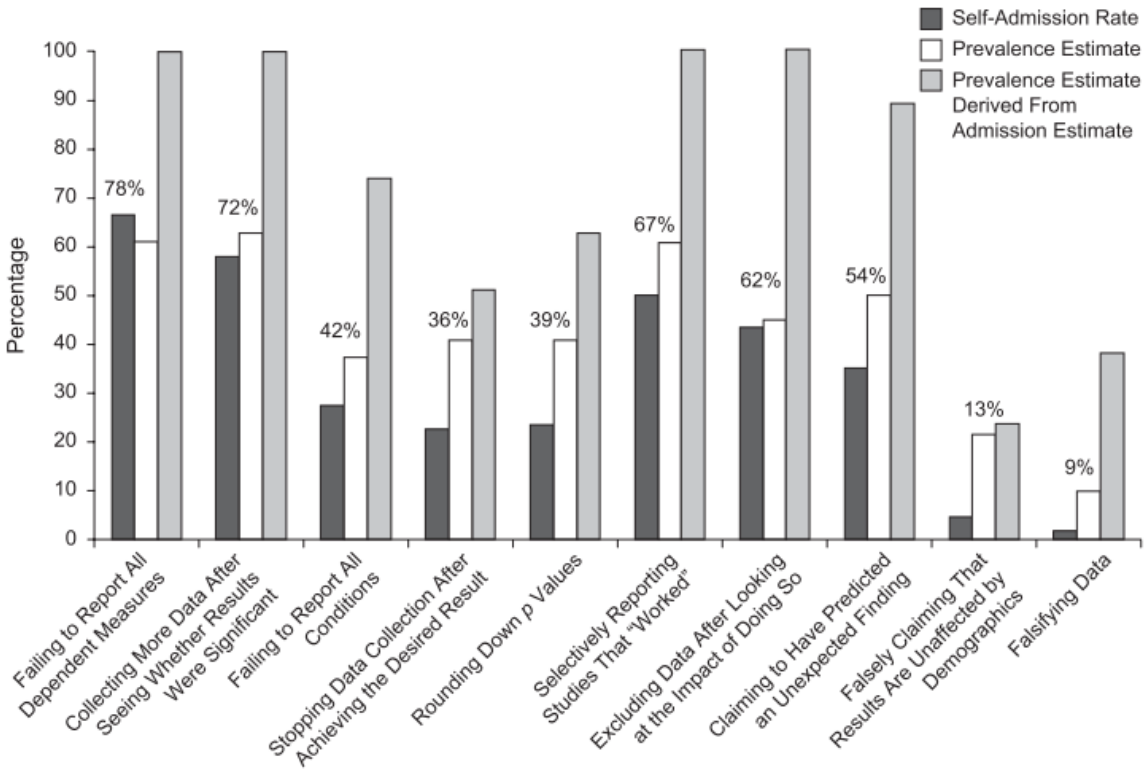
Ces exemples relèvent de la pure fraude et sont largement médiatisés lors de leur découverte. Ce genre d'évènement est cependant assez rare par rapport à d'autres formes d'inconduites bien plus répandues, dont les existences sont largement tributaires de la pression que subissent les chercheurs à publier des résultats significatifs et originaux (Chambers, 2017). On peut par exemple citer le « p-hacking », qui se peut se manifester lorsque les chercheurs collectent des données jusqu'à ce que le test statistique ressorte significatif, ou lorsque que plusieurs analyses sont menées et que seules sont rapportées celles qui soutiennent les hypothèses de l'investigation ; le « HARKing », qui consiste à présenter une hypothèse générée à partir de l'examen des résultats comme ayant été formulée à priori ; ou le manque de partage des données collectées, qui empêche l'exercice optimal de la réplication et de la détection des inconduites (Chambers, 2017). Corvol et Maisonneuve (2016) y incluent également les méconnaissances méthodologiques, telles que les erreurs statistiques, des méthodes faibles ou inappropriées, l'usage d'échantillons trop petit ou encore le manque de recherche documentaire avant la conduction de la recherche… La liste est longue et peut varier selon les auteurs qui se sont penchés sur la question.

Quel est la prévalence de ces inconduites en psychologie ? John, Loewenstein, et Prelec (2012) ont mené un sondage à l'attention de 2155 psychologues américains, dans lequel dix inconduites scientifiques étaient présentées. Pour chacune d'elles, les répondants devaient indiquer s'ils l'avaient eux-mêmes commise, estimer la prévalence de ces inconduites et le taux d'admission de ces inconduites chez leurs collègues. De plus, les répondants étaient

invités à exprimer leur degré de doute par rapport à l'intégrité de la recherche menée par leurs collègues issus d'autres institutions, de leur collègues issus de leur propre institution, des étudiants, de leur collaborateurs et d'eux-mêmes, sur une échelle allant de 1 à 4 (1 = jamais, 2 = une ou deux fois, 3 = occasionnellement, 4 = souvent). Les chercheurs ont prévenu la moitié des participants que leurs réponses seraient vérifiées par un algorithme, et qu'une donation de charité serait accordée à l'institution du choix du répondant si elles s'avéraient correctes. Les autres participants n'étaient pas incités à rapporter la vérité.

Les résultats de cette enquête sont réellement interpellant : un pourcentage élevé de psychologues ont admis avoir réalisé des inconduites, et cela encore plus lorsque les répondants étaient incités à dire la vérité (particulièrement pour les pratiques perçues comme étant les moins défendables ; c'était par exemple le cas pour un item qui portait sur la falsification de données). Dans cette dernière condition, certaines inconduites atteignaient plus de 50% d'admissions (pas de rapport de toutes les mesures dépendantes, 66,5% ; collecte de données après avoir vérifié si les résultats étaient significatifs, 58% ; publier sélectivement les études qui ont « marché », 50%), et les taux d'admissions pour les autres formes d'inconduites restaient anormalement élevées (figure 7a). Enfin, approximativement 35% des répondants exprimaient des doutes par rapport à l'intégrité de leurs propres recherches, et étaient encore une fois davantage critiques relativement aux recherches menées par des collègues travaillant dans d'autres institutions, par rapport à celles menées par leurs collaborateurs ou par eux-mêmes (figure 7b).

La grande majorité de ces inconduites résultent de l'envie de produire des résultats significatifs et publiables, et déforment encore davantage la fiabilité de la littérature scientifique. Considérés ensemble, le biais de publication, le manque de puissance et ces inconduites posent également la question de l'efficacité des méta-analyses, surtout lorsque ces dernières se basent exclusivement sur la littérature publiée ; comme nous allons le voir à présent, la réalisation d'une méta-analyse comprend en plus de nombreuses étapes et est loin d'être un travail aisé.



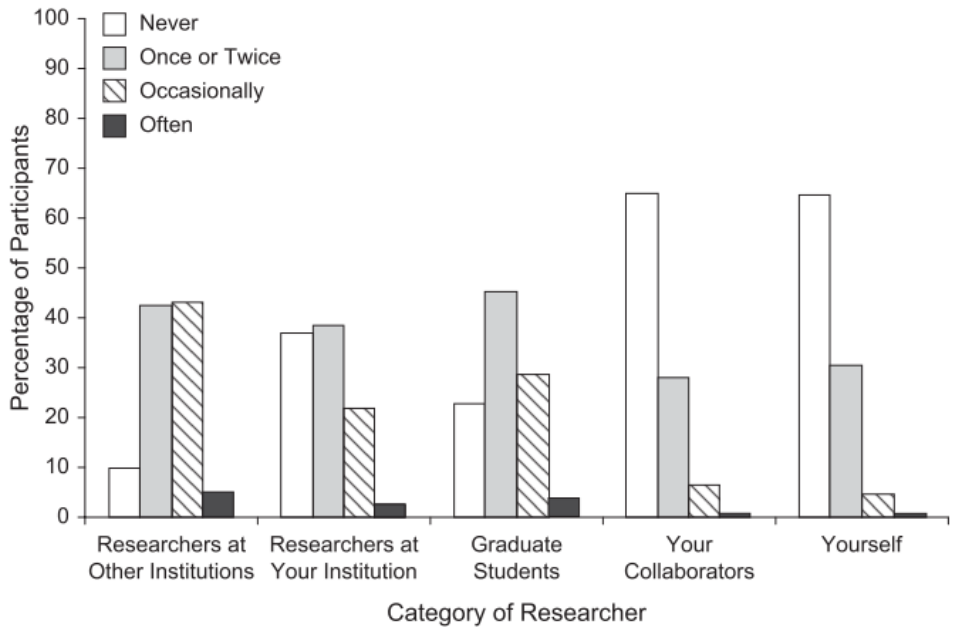


Figure 7 a) Pourcentages des auto-admissions (« self-admission rate »), des estimations de prévalence chez des collègues (« prevalence estimate »), et des estimations d'admission chez des collègues (« prevalence estimate derived from admission estimate ») par rapport à dix inconduites scientifiques dans la communauté des psychologues, dans une condition dans laquelle les répondants étaient incités à rapporter la vérité. Les pourcentages indiqués sur la figure correspondent aux moyennes géométriques des trois pourcentages ; b) Distribution des réponses concernant l'intégrité du travail de plusieurs catégories de chercheurs. Il est à souligner la différence des réponses entre les deux dernières catégories « soi-même » et « collaborateurs », par rapport aux autres catégories. Tiré de John, Loewenstein, & Prelec, (2012).

La méta-analyse

La méta-analyse est la synthèse statistique des résultats issus d'une série d'études (Borenstein, Hedges, Higgins, & Rothstein, 2009). Elle est la partie quantifiée de la revue systématique, et est considérée avec elle comme l'investigation scientifique dont les résultats sont les plus fiables (Howick et al., 2018). La méta-analyse a effectivement de grands avantages à son actif. La combinaison des études permet d'établir un portrait général des résultats qui concerne une question particulière, avec plus de précision et de puissance que les investigations individuelles. Si les résultats viennent à ne pas être consistants entre les études, l'investigation méta-analytique permet parfois de tester le rôle de certaines variables relatives, par exemple, à la méthodologie ou aux participants des études, et d'éclaircir leur influence sur les résultats (par exemple, par des analyses dites en « sous-groupe » ; voir infra).

La méta-analyse devient indispensable pour les chercheurs au vu de la croissance exponentielle du nombre d'articles scientifiques au cours de ces dernières années. Entre 2008 et 2014, le nombre d'articles inclus dans l'index de citations scientifique de Thomson Reuters (Science Citation Index of Thomson Reuters' Web of Science) a augmenté de 23 %, pour passer de 1 029 471 à 1 270 425 articles (UNESCO, 2015). Le nombre de revues systématiques et de méta-analyses a également fortement augmenté : les articles identifiés comme des revues systématiques ou des méta-analyses sur PubMed n'étaient respectivement que 1024 et 334 en 1991 ; en 2014, ces chiffres sont passés à 28,959 et à 9135, ce qui correspond à un taux de publication de 2728% pour les revues systématiques et de 2635% pour les méta-analyses (Ioannidis, 2016).

Les méta-analyses sont ainsi capitales pour tous les chercheurs attachés à l'obtention de résultats fiables, et une certaine minutie lors de leur réalisation est dès lors indispensable pour parvenir à des résultats au plus proche de la vérité. Les méta-analyses s'effectuent en effet en une série d'étapes.

Après la définition des critères d'inclusion des articles et la construction d'une (ou de plusieurs) équation(s) de recherche utiles à l'examen des bases de données bibliographiques, sur base d'une question dont chacun des aspects aura été circonscrit (sur base du « PICO »

par exemple), il s'agira de construire une stratégie de recherche qui soit la plus exhaustive possible ; cela signifie notamment que des efforts doivent être mobilisés à la recherche de la littérature non publiée. Ce point est important, pour toutes les raisons qui ont été exposées concernant les effets du biais de publication, du manque de puissance et des inconduites scientifiques ; si la recherche est limitée à la littérature publiée, le résultat méta-analytique aura encore plus de chance de représenter davantage la tendance des exagérations des résultats que les véritables effets des études. Pour approcher cette exhaustivité, les chercheurs auront bien entendu à chercher les articles dans des bases de données courantes, comme SCOPUS ou PsycInfo, mais aussi à examiner les références des articles, à contacter les auteurs et les experts de la question de recherche, et à examiner des bases de données spécifiques si la question de recherche le permet (par exemple, ClinicalTrial.gov est la base de donnée qui recense actuellement le plus grand nombre d'essais cliniques, qu'ils soient terminés ou en cours de réalisation). Les documents récupérés, en plus des publications conventionnelles, peuvent donc également être des abstracts de conférence, de thèses de doctorat, voire des mémoires d'étudiants.

Après la récupération des documents vient l'étape de la sélection des études qui se réalise en deux phases : l'exclusion des articles non pertinents sur base des titres et des abstracts, puis une autre phase d'exclusion sur base de la lecture détaillée des articles. A ce niveau, les chercheurs ont à expliquer la raison de l'exclusion de chaque article qui ne répond pas aux critères d'inclusion ; ces justifications devront figurer, d'une façon ou d'une autre, dans le manuscrit final ou dans ses annexes. Pour éviter un maximum d'erreurs dans ce processus, ce travail doit être réalisé par (au moins) deux personnes évaluant chacune les références de façon indépendante, et discutant ensuite de chacun de leurs désaccords pour atteindre un consensus sur les articles à inclure (d'autres alternatives existent à cette procédure, comme l'examen d'un échantillon des articles par les deux auteurs, puis par le classement du reste des références par un seul auteur si le résultat d'un test de fiabilité inter-observateur est élevé pour le classement des articles de l'échantillon).

Après l'atteinte d'un consensus sur les articles à inclure, il s'agira de récupérer, pour chaque article, toutes les données pertinentes à l'analyse ; il s'agit bien sûr des tailles d'effet, mais aussi de toutes les informations utiles à la description des études, comme les caractéristiques

des articles, des participants, les méthodes utilisées, la durée des interventions… Ces données devront figurer dans le manuscrit, et le travail doit, comme pour la sélection des études, être réalisé par au moins deux auteurs indépendants ou selon une méthode alternative.

Ensuite, il faudra évaluer les risques de biais dans les articles. Le « biais », dans le cas présent, est défini comme « une erreur systématique ou une déviation de la vérité dans les résultats ou dans les inférences » (Cochrane, 2018) ; il ne s'agit donc pas de juger de *l'imprécision* des résultats, comme ceux résultants de tailles d'échantillons inadaptées et de la taille des intervalles de confiance. Par exemple, l'outil de référence pour les essais contrôlés randomisés est le « Cochrane risk of bias tool », développé par la fondation Cochrane (Higgins & Green, 2008) qui porte sur des domaines tels que la qualité de la randomisation des participants, de leur allocations aux groupes de l'étude, de la qualité du double aveugle ou encore le rapport sélectif de certains résultats. Là encore, les résultats de cette investigation doivent figurer dans l'article final pour chacun des articles étudiés, et idéalement deux auteurs doivent encore réaliser cette tâche de façon indépendante.

Après cela, les chercheurs peuvent enfin se lancer dans la méta-analyse à proprement parler. Le résultat d'une méta-analyse n'est pas une simple moyenne de tailles d'effet ; un « poids » est accordé à chaque taille d'effet, et il est plus important pour les résultats dont la variance est petite (ce qui signifie que davantage de poids est accordé aux études qui incluent beaucoup de participants, qui présentent plus d'évènements et ont des intervalles de confiance plus étroits). De plus, l'analyse devra être réalisée selon un modèle statistique que les auteurs auront spécifié à l'avance (i.e. un modèle à effets « fixe », si les auteurs estiment que la seule source de variation entre les tailles d'effet est due à l'erreur d'échantillonnage, ou un modèle à effets « aléatoires » si à cette influence s'ajoute des différences dans les caractéristiques des participants, des interventions)... On a enfin l'estimation d'un effet résumé, mais d'autres analyses peuvent encore être réalisées ! Il faudra, par exemple, souvent faire attention à l'hétérogénéité des tailles d'effet, qui correspond à la variation des effets entre les études, et qui peut être évaluée par des tests spécifiques (par un test « Q » qui teste l'hypothèse nulle de l'homogénéité et fournit une p-valeur, et par un test « I^2 » qui donne une taille d'effet). Si l'hétérogénéité s'avère importante, les chercheurs doivent en explorer les causes : il s'agira d'étudier séparément le rôle de certaines variables relatives aux participants

ou aux études, et qui sont susceptibles d'influencer l'effet résumé d'une manière ou d'une autre. Une analyse commune qui permet ce genre d'investigation est l'analyse en « sous-groupes », qui consiste à diviser les différentes tailles d'effets en groupes selon les modalités de la variable dont on veut explorer l'effet particulier (par exemple, on pourrait diviser les études en deux groupes selon qu'elles se situent au-dessus ou en dessous d'un score « cut-off » relatif à leur qualité méthodologique). Une partie importante des résultats doivent concerner l'évaluation du biais de publication, dont les tests les plus courant sont certainement le « funnel plot », qui fournit une représentation graphique du biais de publication, et le « trim and fill » qui permet d'obtenir une nouvelle taille d'effet résumée sur base de la correction d'un éventuel biais de publication. D'autres analyses sont encore possibles (e.g. les méta-régressions, dont le but est similaire aux analyses en sous-groupes mais qui permettent en plus d'explorer les effets de variables continues ; les analyses de sensibilité, dans lesquelles on relance des analyses afin de tester l'impact de différentes décisions prises durant la réalisation de la méta-analyse). Toutes ces analyses doivent apparaitre au mieux dans le protocole publié avant la réalisation de la méta-analyse, et au moins dans la méthode du manuscrit. Les analyses post-hoc doivent bien entendu être déclarées comme telles.

Enfin, la rédaction de l'article, comme toutes les investigations scientifiques, doit être réalisé de manière à ce que la méthode et les résultats soit compréhensibles et reproductibles. La liste PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses ; Moher et al., 2009) est un outil répandu et utile à cet égard.

La qualité des méta-analyses

La réalisation d'une méta-analyse est ainsi une tâche longue, et certaines de ses étapes peuvent être véritablement fastidieuses. Borah, Brown, Capers, et Kaiser (2017) ont utilisé la base de données PROSPERO, qui regroupe les protocoles des revues systématiques et des méta-analyses, pour évaluer le temps moyen mobilisé par les chercheurs pour réaliser une revue systématique (dont la méta-analyse est le résultat quantifié), en se basant sur le temps écoulé entre la publication du protocole et la publication de l'article : le temps moyen estimé pour réaliser une revue systématique était ainsi de 67,3 semaines. Des manquements ou des

biais peuvent de plus être introduit à chacune des étapes, et avoir des conséquences parfois importantes sur le résultat final. Par exemple, Useem et al. (2015) ont examiné dans quelle mesure des méta-analyses issues de la collaboration Cochrane, dont les revues systématiques et les méta-analyses font référence en matière de qualité méthodologique, étaient différentes de revues « non Cochrane » menées sur les mêmes questions relatives au domaine cardio-vasculaire. Concrètement, 80 méta-analyses Cochrane et non Cochrane ont été récupérées et classées en paires selon la similarité de leurs questions et de leurs analyses. Au total, les méta-analyses incluaient 344 études, dont 111 (32,3%) étaient inclues uniquement dans une revue Cochrane, 104 (30,2%) uniquement dans une revue non Cochrane, et 129 (37,5%) étaient inclues dans les deux sources, et cette différence n'étaient pas uniquement expliquée par la séquence de publication pour 47% de 32 paires divergentes (i.e. la publication d'une revue Cochrane avant la publication d'une revue non Cochrane, ou inversement, qui pourrait expliquer les différences dans la mesure ou de nouvelles études peuvent être publiées dans l'intervalle de temps séparant la publication des deux méta-analyses). De plus, sur les 40 paires de méta-analyses, 7 d'entre elles différaient quant à leur conclusion (i.e. l'intervalle de confiance était différent de manière à ce qu'une des méta-analyses d'une paire rapporte un résultat statistiquement significatif, et l'autre non), 3 s'accordaient sur la direction de la taille d'effet mais pas sur sa magnitude, qui était au moins deux fois plus importante pour l'une des revues, et 5 différaient quant à la direction de la taille d'effet. Enfin, les revues non Cochrane rapportaient globalement de plus grandes tailles d'effet, et étaient moins précises (leurs erreurs standard étaient plus grandes) que les revues Cochrane. Un résultat préoccupant est que parmi les paires qui différaient quant à la magnitude de leur taille d'effet, la méta-analyse de la paire qui avait le plus grand effet était citée environ 4 fois plus que l'autre méta-analyse. Qu'est ce qui peut expliquer ces différences ? Useem et al. (2015) proposent qu'elles peuvent essentiellement être due à des différences dans les stratégies de recherche des articles, et/ou à la façon de sélectionner et d'exclure les études, une hypothèse qui est renforcée par le fait que les revues Cochrane et non Cochrane n'incluaient généralement pas les mêmes études. Par exemple, les revues Cochrane et non Cochrane pourraient explorer préférentiellement certaines bases de données, différer sur l'inclusion des articles qui ne sont pas rédigés en anglais…

Ces résultats révèlent la problématique des choix qui peuvent être posés aux différentes étapes des méta-analyses, et plus largement de la qualité de ces investigations pour l'ensemble de la littérature scientifique. La notion de « qualité » est ici polysémique. Il peut d'une part s'agir de la qualité de « reporting », c'est-à-dire de la manière dont les auteurs ont rédigé leur article et qui doit permettre aux lecteurs de comprendre, et éventuellement de reproduire, l'expérience dont il est question. La check-list PRISMA (Moher et al., 2009), dont nous avons déjà touché un mot, contient 27 items qui portent notamment sur la définition de la question de recherche, sur les aspects de la méthode, sur la présentation des résultats ou encore sur la structure de la discussion, et permet ainsi d'aider les chercheurs à rédiger le manuscrit de leurs méta-analyses. Le second sens de « qualité » porte bien entendu sur la qualité intrinsèque, méthodologique des méta-analyses, et ici aussi des check-lists ont été développées afin d'aider les auteurs et les lecteurs à s'en faire une certaine représentation. La plus connue de ces listes est certainement la liste AMSTAR (A MeaSurement Tool to Assess systematic Reviews ; Shea et al., 2007), qui contient 11 items relatifs à la conduction des différentes étapes des méta-analyses. Une nouvelle version de cette liste, AMSTAR 2, a été publiée récemment (Shea et al., 2017) et contient quant à elle 16 items.

L'investigation de la qualité des méta-analyses est à ce jour une entreprise exclusivement biomédicale, et a été réalisée grâce à AMSTAR, AMSTAR 2 et PRISMA dans des champs tels que la chirurgie (Zhang et al., 2016), la chirurgie pédiatrique (Cullis, Gudlaugsdottir, & Andrews, 2017), le psoriasis (Gómez-García et al., 2017), le diabète (Wu et al., 2016), l'orthopédie (Gagnier & Kellam, 2013), la radio-oncologie (Hasan et al., 2017), la gestion des soins des personnes brûlées (Wasiak, Tyack, Ware, Goodwin, & Faggion, 2016) et la dépression (Zhu et al., 2016). Le constat de ces articles est globalement interpellant. Intéressons-nous dans un premier temps aux résultats AMSTAR. Mis à part pour l'item d'AMSTAR relatif au rapport des caractéristiques des études individuelles, toujours respecté à plus de 80% dans ces études, l'adhérence aux autres items est au mieux moyenne, voire toute à fait faible (dans ce qui suit, sont indiqués entre parenthèses respectivement l'adhérence la plus faible et la – ou les – meilleures adhérences) ; si les résultats varient d'une investigation à l'autre, les problèmes les plus prégnants concernent la publication d'un protocole (13% d'adhérence des revues systématiques en orthopédie à cet item d'AMSTAR, jusqu'à 40,9% dans le champ du psoriasis), la recherche de la littérature non publiée (18.8% en chirurgie, 47% en orthopédie),

le rapport de la liste de toutes les études inclues et exclues au moment de la sélection des articles (3.6% en chirurgie, 43% dans le domaine de la brûlure ; le domaine de l'orthopédie s'en sort largement mieux avec 86% des revues systématiques qui adhérent à cet item), l'évaluation du biais de publication (9% en orthopédie, 43,1% en radio-oncologie), et la déclaration des éventuels conflits d'intérêt et des sources de financement (0% pour la gestion des soins des personnes brûlées et 7,5% dans le diabète, 86% en orthopédie). En ce qui concerne AMSTAR 2, seule une étude, au moment de la rédaction de ce mémoire, a été publiée, et concernait l'utilisation de feuilles transparentes et de lentilles de contact colorées pour réduire des difficultés de lecture (4 articles inclus dans cette étude ; Suttle, Lawrenson, & Conway, 2018). A nouveau, les résultats sont assez moyen ; par exemple, aucun des 4 articles ne faisaient référence à un protocole, aucune n'a pleinement mené une stratégie de recherche systématique et 3 n'ont pas évalué adéquatement le risque de biais dans les études.

Cinq de ces articles ont également évalué la qualité du reporting des méta-analyses grâce à PRISMA (Zhang et al., 2016 ; Cullis et al., 2016 ; Gagnier & Kellam, 2013 ; Wasiak et al., 2016 ; Zhu et al., 2016). Les items évaluent globalement le même type de domaines que ceux d'AMSTAR, et rapportent une ampleur similaire dans leurs manquements. Zhang et al. (2016) et Zhu et al. (2016) ont de plus mis en évidence une relation substantielle entre les résultats de PRISMA et d'AMSTAR (respectivement $r^2 = 0.79$ et $r^2 = 0.56$).

La problématique des choix qui peuvent être posés aux étapes des méta-analyses doit également être envisagée par rapport aux motivations relatives à la publication de ces articles. Le cas de la littérature sur les antidépresseurs est très parlant. La quantité des méta-analyses dans ce domaine est absolument colossale : 185 articles ont été produits sur le sujet entre 2007 et 2014 ! 54% de ces études comprenaient des auteurs employés par des sociétés pharmaceutiques, et 79% avaient des liens d'intérêt avec de telles industries. Seules 31% des méta-analyses aboutissaient à un résultat négatif, et celles qui incluaient un auteur employé dans la société à l'origine de la molécule évaluée avaient 22 fois moins de chance d'aboutir à des résultats négatifs que les autres (Ebrahim, Bance, Athale, Malachowski, & Ioannidis, 2016). Ce cas montre combien les conflits d'intérêts peuvent influencer le résultat des méta-analyses qui font alors dans ce cas clairement office d'outils marketing (Ioannidis, 2016). Le problème est bien plus large que les cas relatifs au financement d'études pharmaceutiques : les conflits

d'intérêts incluent également les opinions personnelles, politiques, académiques, idéologiques, voire mêmes religieuses, qui peuvent également jouer un rôle déterminant sur le résultat et la publication d'une étude (Barbour et al., 2008). Autant que possible, ce genre de conflits d'intérêts devrait également être rapporté dans les manuscrits, même s'il s'agit dans les faits d'une entreprise difficile à mettre en œuvre de façon cohérente et standardisée. Des journaux ont néanmoins tenté de favoriser ces déclarations : par exemple, la politique des journaux *Plos* déclare que le rapport des conflits d'intérêts est requis pour la publication d'un article et qu'autrement, le manuscrit peut être immédiatement rejeté, avec une insistance sur le fait que ces conflits peuvent prendre d'autres formes que les intérêts financiers (Plos, 2018).

Indices bibliométriques : garants d'une certaine qualité méthodologique ?

Eugène Garfield a eu l'idée de la création du facteur d'impact en 1955, pour quantifier l'influence d'un article particulier en comptant le nombre de citations que celui-ci recevait (Garfield, 1955). Ce projet s'inscrivait dans l'ambition plus générale d'apporter de la cohérence et de l'efficacité à la science ; avant cela, les scientifiques ne pouvaient en effet pas savoir, de façon systématique, quel article était cité par qui, ou de quantifier le nombre de citations reçues par les articles ou les journaux (Chambers, 2017). Le facteur d'impact est aujourd'hui calculé au niveau des journaux, et non pas des articles. Il s'agit le plus couramment du rapport entre le nombre de citations que le journal a reçu au cours des deux dernières années, tel que rapporté par le *Journal Citation Reports*, sur le nombre total d'articles publiés sur la même période. Par exemple, un facteur d'impact de 1,0 signifie qu'en moyenne, les articles publiés il y a un ou deux ans ont été cités une fois ; un facteur d'impact de 2,5 signifie qu'en moyenne, les articles publiés il y a un ou deux ans ont été cités deux fois et demie (University College Dublin, 2017). Les buts de Garfield avec cette mesure étaient de faciliter la décision à l'abonnement de tel ou tel journal, et d'aider les auteurs à décider de la revue dans laquelle publier leurs articles, sur le principe que les facteurs d'impact les plus élevés reflètent les revues les plus prestigieuses (Garfield, 2006).

Le facteur d'impact a été largement critiqué (e.g. Adler, Ewing, & Taylor, 2008 ; Chambers, 2017 ; Casadevall & Fang, 2015). Par exemple, la mesure simplifie exagérément la réalité de ce qui fait qu'un article ou un chercheur est cité ; pour reprendre la formule d'Adler et al.

(2008), « sa seule utilisation pour juger un journal est comme utiliser uniquement le poids pour juger de la santé d'une personne ». Le fond de la question est de savoir ce qu'une citation signifie réellement, ou formulé autrement, c'est de savoir *pourquoi* les chercheurs citent des articles. Les raisons qui poussent à la citation sont en effet multiples : il peut s'agir de citer une réalisation récente (peu importe sa qualité intrinsèque), d'entretenir un consensus, de relever un aspect méthodologique rédhibitoire, de continuer une discussion scientifique, ou encore parce que les cités ont une certaine relation avec les citants, ou que l'article est issu d'un journal en *open access*, et dont l'accès est plus facile (Adler et al., 2008 ; Chambers, 2017)… Ces raisons n'ont pas nécessairement de lien avec la qualité intrinsèque du travail qui est cité. Un autre problème du facteur d'impact est que la majorité des citations d'un journal est générée par une minorité d'article, un phénomène que Garfield a lui-même a appelé la règle « 80/20 » et qui signifie qu'environ 20% des articles d'un journal expliquent 80% des citations. Ce phénomène est visible au niveau des journaux comme pour la littérature de façon générale : sur les 38 millions d'articles cités entre 1900 et 2005, seuls 0,5% ont été cités plus de 200 fois alors que la moitié n'ont pas été cités une seule fois (Garfield, 2006). Cela remet en question la pertinence du facteur d'impact dans la mesure où la majorité des articles d'un journal ont moins de citations que la valeur de celui-ci ; le facteur d'impact ne représente ainsi adéquatement qu'une minorité des articles publiés (rappelons qu'il s'agit d'une simple *moyenne* du taux de citation). La figure 8 montre un exemple de ce phénomène.

Cela signifie que publier un article dans un journal à haut impact ne garantit pas que l'article soit cité ; inversement, publier dans un journal moins noté n'induit pas nécessairement qu'il soit ignoré. Lozano, Larivière et Gingras (2012) ont en fait montré que depuis les années 1990, la relation entre le taux de citation des articles et le facteur d'impact, qui était déjà faible, diminue, à cause de la publication digitale des articles qui les a libérés du carcan que constituait le format papier : ce dernier était en effet davantage soumis à des considérations relatives au facteur d'impact quand les universités devaient choisir les abonnements aux journaux. Dans le même temps, la proportion d'articles hautement cités en dehors des journaux à haut impact augmente. Ces considérations (parmi bien d'autres, dont l'exposition complète sort du cadre de ce travail ; voir par exemple Adler et al., 2008 ou Casadevall & Fang, 2015) font qu'il est incorrect d'évaluer la qualité des articles pris individuellement, et à fortiori

le travail des scientifiques, sur base du facteur d'impact des journaux dans lesquels ils publient.

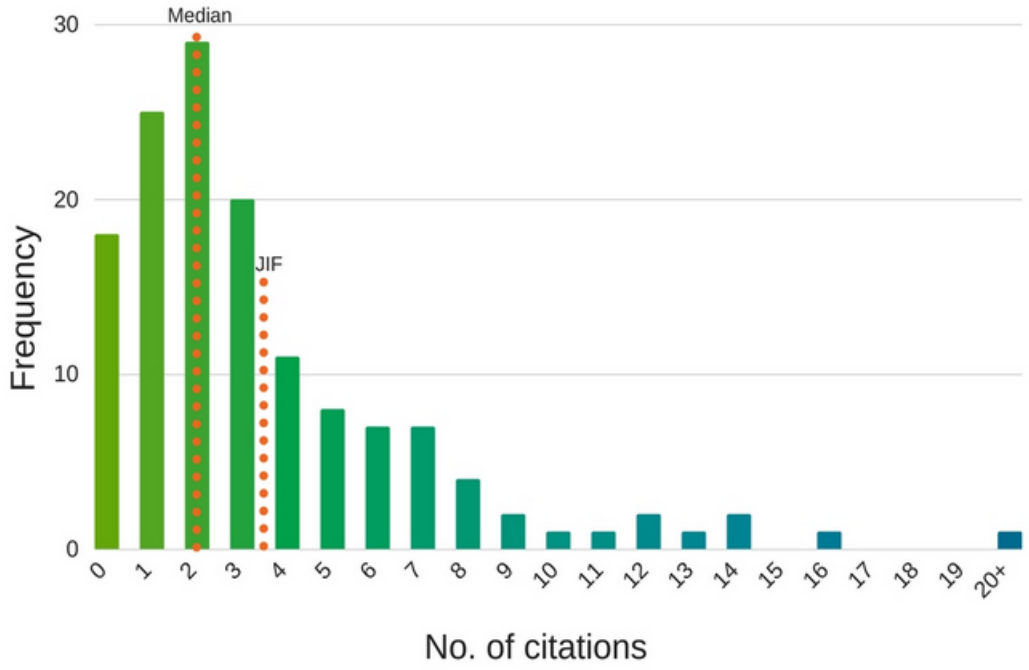


Figure 8 : fréquence des taux de citations pour le journal *Open Biology*. JIF = Journal Impact Factor ; ce dernier est relatif ici à l'année 2016. Tiré de The Royal Society (2018).

Dans les faits, il reste cependant que cette mesure est utilisée pour comparer les articles et les chercheurs. Le facteur d'impact est notamment largement mobilisé pour l'allocation de bourses, de promotions, ou encore pour postuler à un emploi dans une université (Casadevall & Fang, 2015).

On peut, bien entendu, se demander dans quelle mesure le facteur d'impact ne prédit pas tout de même une forme de qualité méthodologique. Par exemple, Brembs, Button, et Munafò, (2013) ne trouvent aucune relation entre la puissance statistique de 650 études en neurosciences et le facteur d'impact des journaux dans lesquels les articles étaient publiés. Ils ont trouvé plus spécifiquement, pour des études évaluant le lien entre un certain génotype et des traits de personnalité relatifs à l'anxiété, une relation entre la mesure dans laquelle une étude surestime la taille d'effet, la taille des effectifs et le facteur d'impact : Plus le facteur d'impact est grand, plus le biais dans l'estimation de la magnitude des tailles d'effet est grande, et plus les effectifs sont petits. De plus, Fang, Steen, et Casadevall (2012) ont trouvé

une corrélation positive entre la proportion d'articles rétractés d'un journal et son facteur d'impact (r^2 = 0.09), la majorité de ces rétractions résultants de cas de fraudes. Pour la littérature touchant à la médecine préclinique, il semblerait que beaucoup d'études publiées dans des journaux à haut impact ne soient pas reproductibles (Begley & Ellis, 2012). Ces résultats peuvent certainement être en partie expliqués par la pression que les scientifiques ont à publier dans ces journaux, et de toutes les conséquences qu'entraine ce genre de culture et dont nous avons déjà parlé. Au niveau des méta-analyses, sept des neuf articles traités lors de la présentation de la qualité des méta-analyses ont aussi examiné l'influence du facteur d'impact sur celle-ci (Gagnier & Kellam, 2013 ; Wasiak et al., 2016 ; Wu et al., 2016 ; Zhu et al., 2016 ; Cullis et al., 2017 ; Gómez-García et al., 2017; Hasan et al., 2017). Seule l'investigation de Gómez-García et al., (2017) sur le psoriasis a clairement montré une relation entre le facteur d'impact calculé sur 5 ans et la qualité des méta-analyses (OR = 1,34 : 95% CI 1,02–1,40) mais au vu des données les auteurs recommandent tout de même de rester prudent, même pour les revues publiées dans des journaux à haut impact.

Dans un esprit analogue de fournir une évaluation « objective » de la productivité et de l'impact des chercheurs, Jorge E. Hirsch a proposé un autre indice bibliométrique qui est aussi couramment utilisé aujourd'hui : l'index « H » (Hirsch, 2005). Concrètement, un scientifique possède un index H de 20 si celui-ci possède 20 articles qui ont été cités au moins 20 fois chacun (Chambers, 2017). L'index H a été imaginé explicitement comme une façon de classer les chercheurs : dans le domaine de la physique, Hirsch a suggéré qu'un H de 12 pourrait permettre de considérer un poste de chercheur dans une université américaine renommée. Une valeur de 18 pourrait conduire à un poste de professeur, une valeur entre 15-20 pourrait mener à devenir membre de *l'American Physical Society*, et une valeur supérieure à 45 pourrait permettre l'admission à la *National Academy of Sciences* aux Etats-Unis (Society for Science & the Public, 2005). Comme le facteur d'impact, l'index H est également concerné par le problème du *sens* des citations, et par le fait que généralement seule une petite partie des travaux d'un chercheur est cité abondamment ; ce qui fait que deux scientifiques qui ont des taux de citations très différents peuvent avoir le même index H. De plus, il a également l'inconvénient d'être corrélé à la durée de la carrière académique, et il dépend aussi du champ de recherche dans lequel l'article est publié est cité (Bornmann & Daniel, 2009). A nouveau, deux des neufs articles présentés lors de l'exposition de la qualité des méta-analyses ont

examiné l'effet du facteur d'impact (Cullis et al., 2017 ; Gómez-García et al., 2017). Aucune de ces études n'a montré un effet clair de l'index H sur la qualité des méta-analyses.

Buts du mémoire

Nous avons vu que les investigations qui se sont attachées à évaluer la qualité des méta-analyses pour différents champs de recherche rapportent des résultats au mieux moyens, voire faibles lorsqu'ils sont évalués avec les outils PRISMA, AMSTAR et AMSTAR 2. Ce genre d'entreprise n'a cependant pas été réalisée spécifiquement en psychologie et dans des disciplines apparentées. On peut poser l'hypothèse que la qualité des méta-analyses ne devrait pas non plus être optimale dans ces domaines. Nous explorerons la qualité de méta-analyses publiées sur PsycINFO grâce à l'utilisation de PRISMA, d'AMSTAR et d'AMSTAR 2.

Nous explorerons également l'impact, sur les scores AMSTAR et AMSTAR 2, du facteur d'impact, de l'index H et de l'expérience du premier auteur en matière de production de méta-analyses. Ces trois variables ont été choisies car on peut supposer que le lecteur courant les associe naturellement avec la qualité des résultats scientifiques. La problématique du facteur d'impact et de l'index H ont déjà été discutés ; L'expérience du premier auteur est ici opérationnalisée par le nombre de méta-analyses qu'il a réalisé. Deux des neufs articles présentés lors de la discussion de la qualité des méta-analyses ont évalué l'impact de cette variable (Zhang et al., 2016 ; Zhu et al., 2016). Seule l'étude de Zhang et al. (2016), relative aux méta-analyses portant sur la chirurgie, a montré un effet de cette expérience sur la qualité méthodologique (Odds ratio = 2.41, 95CI = 1.20 ; 4.84). Nous verrons également si ces trois variables sont liées à l'utilisation, pour la rédaction des méta-analyses, de la grille PRISMA par les auteurs.

Nous verrons enfin comment les scores à PRISMA, AMSTAR et AMSTAR 2 varient selon que les auteurs rapportent avoir utilisé PRISMA ou non ; des tests s'attacheront à évaluer cette influence potentielle au niveau de chacun des items des check-lists.

2) Méthode

Ce mémoire est issu d'un projet plus vaste dont le protocole a été publié sur Open Science Framework (OSF) (Leclercq et al., 2018).

Critères d'inclusion des méta-analyses

Toutes les méta-analyses concernant des problématiques humaines et publiées sur PsycInfo en 2016 ont été recherchées. Seuls les articles rédigés en anglais ont été considérés pour une inclusion, étant donné les ressources et l'expertise limitées des auteurs du projet. Les critères d'inclusion et d'exclusion détaillés se trouvent dans le tableau 1.

Inclusion criteria

- *Meta-analysis*
- *Articles published in the PsycINFO-database*
- *Published between 01.01.2016 to 31.12.2016*
- *English*

Exclusion criteria

- *Overview, review*
- *Meta-synthesis*
- *Qualitative meta-analysis*
- *Umbrella review*
- *Meta-analysis of meta-analyses*
- *Systematic review without meta-analysis*
- *Protocol of meta-analysis*
- *Network meta-analysis*
- *Activation likelihood Estimation Meta-analysis (ALE MA)*
- *Signed differential mapping meta-analysis (SMD MA)*
- *Voxel wise meta-analysis*
- *Individual patient data meta-analysis (IPD MA)*
- *Genetic association study (GWAS), genetic study*
- *Multi-level meta-analysis*
- *Update*
- *Letter, comment, abstract, chapter, erratum, dissertation or editorial journal*

Tableau 1 : critères d'inclusion et d'exclusion des articles. Adapté de Leclercq et al.

Recherche de la littérature et sélection des méta-analyses

L'équation de recherche utilisée sur PsycInfo est fournie dans le tableau 2.

1	meta analysis.md. (15886)
2	meta analysis/ (3940)
3	meta analys*.mp. (24573)
4	data pooling*.mp. (50)
5	2 or 3 or 4 (24599)
6	5 not 1 (10725)
7	1 or 6 (26611)
8	limit 7 to (English and human and yr="2016") (2159)

Tableau 2 : Equation de recherche pour PsycInfo. Les chiffres entre parenthèses représentent le nombre de résultats obtenus à chaque niveau de l'équation.

La recherche a été réalisée en janvier 2017 et a abouti à 2159 articles correspondant aux critères d'inclusion. Une fois les articles pertinents identifiés, il était prévu de sélectionner aléatoirement environ 200 méta-analyses de cette sélection. Pour ce faire, les références des articles ont été indexées dans un fichier Excel, assignées à un nombre et classées par ordre croissant. Deux personnes, dont l'auteur du mémoire, ont classé séquentiellement les méta-analyses soit dans un groupe « PRISMA » (i.e. les auteurs ont affirmé avoir utilisé la liste PRISMA pour la rédaction de leur méta-analyse), soit dans un groupe « non PRISMA » (i.e. les auteurs n'ont pas rapporté avoir utilisé la liste PRISMA) jusqu'à ce que chaque groupe contienne au moins 100 méta-analyses (ce qui a abouti au final au classement de 107 méta-analyses dans le groupe non PRISMA, et 100 dans le groupe PRISMA).

Une deuxième phase de sélection, propre à ce mémoire, a été menée afin de restreindre strictement le thème des méta-analyses aux sujets relevant de la psychologie et de disciplines apparentées, car il s'est avéré qu'un certain nombre des 207 articles sélectionnés adressaient des thématiques biomédicales non pertinentes par rapport à la psychologie et au comportement. Les disciplines en question comprennent la psychiatrie, la neurologie, l'épidémiologie ou encore les soins infirmiers ; les auteurs de ces disciplines investiguent parfois des thématiques et utilisent des méthodes qui permettent de répondre à des questions tout à fait pertinentes par rapport à la psychologie et au comportement. En

pratique, il était ainsi difficilement justifiable d'exclure des méta-analyses simplement sur base de l'appartenance à tel ou tel journal, ou relativement aux différents services dans lesquels travaillaient les auteurs. Pour réaliser cette sélection, l'auteur du mémoire a classé les méta-analyses en deux catégories sur base d'un jugement des titres et des abstracts des articles ; ceux qui ne laissaient pas entendre l'évaluation d'une variable dépendante comportementale et/ou psychologique ont été exclus des analyses, ce qui représente 37 études en tout et réduit l'échantillon de ce mémoire à 170 méta-analyses (90 dans le groupe non PRISMA, et 80 dans le groupe PRISMA). La figure 9 résume le processus de sélection des méta-analyses. Les références des méta-analyses inclues et exclues se trouvent en annexe.

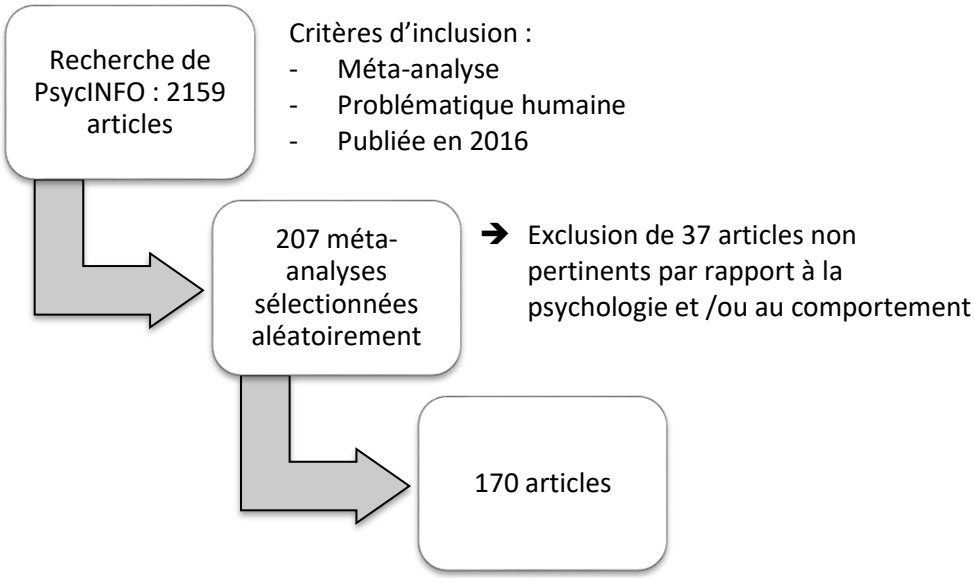


Figure 9 : Résumé du processus de sélection des méta-analyses pour ce mémoire.

Définition et extraction des données

Deux personnes, dont l'auteur du mémoire, ont lu et scoré les 207 méta-analyses sélectionnées des 2159 résultats de la recherche de PsycInfo, de façon indépendante sur PRISMA, AMSTAR et AMSTAR 2 (les check-lists sont disponibles en annexe). PRISMA évalue la qualité de reporting des revues systématiques et des méta-analyses, et chaque item peut être scoré comme étant « validé » ou « non validé », ces réponses ayant été respectivement codées en « 1 » et « 0 ». Le score total de PRISMA est ainsi l'addition des items codés « 1 », et peut être égal à un maximum de 27.

Les items d'AMSTAR ont été codés, dans le cadre de ce mémoire, exactement de la même manière, et le score maximum de cette liste est de 11. Cette logique est aussi similaire pour AMSTAR 2, à la différence que certains items de cette liste, en plus de pouvoir être « validés » ou « non validés », peuvent être « partiellement validés » selon des conditions exposées dans la liste ; les réponses ont ainsi été respectivement codées en « 1 », « 0 » et « 0.5 ». Le score total d'AMSTAR 2 est aussi l'addition de ces scores, et peut être au maximum égal à 16.

Les facteurs d'impact ont été récupérés sur les sites des journaux concernés. Pour un petit nombre de journaux, cette information n'était pas disponible, et a donc mené à quelques données manquantes (c'était le cas pour 6 des 170 méta-analyses de ce mémoire). Les observations pour lesquelles le facteur d'impact était manquant ont été exclues de certaines analyses du mémoire (voir infra, dans la section « analyses statistiques »). Les index H des premiers auteurs ont été récupérés sur SCOPUS. L'expérience des premiers auteurs correspond au nombre total de revues systématiques pour lesquelles ces derniers ont été cités en tant qu'auteur. Le nombre de revues systématiques a été choisi plutôt que le nombre de méta-analyses, car les premières peuvent inclure les secondes sans que cela ne soit rapporté dans le titre de l'article. Cette donnée à également été extraite à partir de SCOPUS. Ces trois variables sont une sélection de nombreuses autres variables dont les rôles sur la qualité des méta-analyses vont être explorés dans le projet dont ce mémoire est issu (la liste complète de ces variables se trouve en annexe). Le facteur d'impact, l'index H et l'expérience de l'auteur ont été sélectionnés dans ce travail car on peut supposer que beaucoup de chercheurs sont, ou pourraient être tentés de les lier subjectivement à la qualité des méta-analyses. Les deux auteurs ont atteint un consensus quant à ce travail d'extraction.

Taille d'effet minimalement détectable

Dans le cadre de ce travail, une analyse de puissance n'est pas aisée étant donné le manque d'informations sur les tailles d'effets en jeu dans la littérature préexistante, notamment au niveau de la différence entre les groupes PRISMA et non PRISMA, et aussi au niveau du nombre de tests effectués au total (voir infra, dans la description des analyses statistiques). La taille d'échantillon choisie par les auteurs du projet dont est issu ce mémoire (au moins 100 méta-analyses par groupe) peut détecter une taille d'effet moyenne (d = 0.46065, calculée via

G*Power version 3.1.9.2) pour un test t pour groupes indépendants (bilatéral), avec une probabilité alpha de 0.05 et une puissance de 90% (afin de comparer les groupes PRISMA et non PRISMA). Suivant la même logique, on peut évaluer la taille d'effet minimalement détectable par l'échantillon plus réduit de ce mémoire : en l'occurrence et pour les mêmes circonstances, il permet également de détecter une taille d'effet moyenne (d = 0.50096), à un niveau alpha de 0.05 et une puissance de 90%.

Statistiques descriptives et inférentielles

Des graphiques représentent, pour les scores AMSTAR, AMSTAR 2 et PRISMA, la proportion des items qui ont été validés ou non validés.

Les méta-analyses ont aussi été classées selon les critères d'AMSTAR 2. Les auteurs de cette liste ont en effet proposé un système de classement des articles selon leur adhérence à certains des items (Shea et al., 2017). Ils ont suggéré que sept items sont particulièrement importants quant à la qualité d'une revue systématique et d'une méta-analyse : l'enregistrement d'un protocole avant le commencement de la revue (item 2), l'exhaustivité de la recherche de la littérature (item 4), les justifications apportées quant à l'exclusion de certaines études (item 7), l'évaluation des risques de biais dans les études (item 9), l'adéquation des méthodes de combinaison des données (item 11), la prise en compte des risques de biais dans les études lors de la discussion des résultats (item 13), et l'évaluation de la présence et de l'impact du biais de publication (item 15). Selon cette classification, un article est de *haute* fiabilité lorsqu'il ne contient pas, ou s'il contient un manquement à un item « non critique », de fiabilité *modérée* lorsqu'il contient plus d'un manquement à des items non critiques, de fiabilité *faible* lorsqu'il possède un manquement à un item critique avec ou sans manquement à des items non critiques, et de fiabilité *sévèrement faible* lorsqu'il contient plus d'un manquement critique avec ou sans manquements non critiques. Un diagramme circulaire a été généré pour représenter le classement des articles selon ces critères.

Effets du facteur d'impact, de l'index H et de l'expérience du premier auteur

L'impact de ces variables sur la propension des auteurs à utiliser PRISMA, ainsi que sur l'adhérence à AMSTAR et à AMSTAR 2, a été évaluée avec des régressions logistiques multivariées. Pour ce faire, les médianes des scores AMSTAR et AMSTAR 2 ont servi de cut-off pour former à chaque fois deux groupes de méta-analyses, l'une représentant les scores les plus bas et l'autre les scores les plus élevés de ces listes. En ce qui concerne PRISMA, la variable dépendante était déjà dichotomique (groupe PRISMA vs groupe non PRISMA).

Des odds ratio et leurs intervalles de confiance à 95% ont été rapportés comme taille d'effet. L'odds ratio est une mesure qui permet de comparer les chances qu'un événement ou qu'un résultat se produise dans un groupe par rapport à la chance que celui-ci a de se produire dans un autre groupe (Ellis, 2010), et s'applique traditionnellement à des tables de contingence. Pour illustrer la manière dont celui-ci se calcule pour des données binaires, reprenons un exemple issu de Simon (2001) qui montre à quel point les hommes étaient plus susceptibles de mourir que les femmes lors du naufrage du Titanic :

	Survie	*Décès*	*Total*
Femmes	308	154	462
Hommes	142	709	851

L'odds ratio de ces données se calcule comme suit :

- « Chance » de mourir dans le groupe des femmes : 154/308 = 0.5
- « Chance » de mourir dans le groupe des hommes : 709/142 = 4.99

Pour le groupe des hommes, l'odds ratio est ainsi égal à 4.99/0.5 = 9.98. Cela signifie que les hommes avaient environ 10 fois plus de « chances » de mourir lors du naufrage du Titanic que les femmes. Si l'odds ratio avait été égal à 1, cela aurait signifié qu'il n'y avait pas de différence entre les hommes et les femmes quant aux « chances » de décès.

Une régression logistique fournit également des odds ratio pour chacune des variables prise en compte dans le modèle statistique, mais leur calcul est différent. Dans ce cas, ils sont obtenus par l'exponentiation de leurs coefficients de régression. L'interprétation de cet odds ratio est légèrement différente et se fait en termes d'augmentation en unités de la variable

explicative (les valeurs des autres variables entrées dans le modèle restant constantes). Concrètement, dans une régression logistique évaluant l'effet du facteur d'impact sur les scores AMSTAR, un odds ratio de 2 aurait signifié qu'à chaque augmentation d'une unité du facteur d'impact, la « chance » d'obtenir un score reflétant une bonne qualité méthodologique (i.e. qui se trouverait dans le groupe représentant les meilleurs scores, tel que défini en début de section) augmenterait d'un facteur de 2.

Impact de l'adoption de PRISMA sur la qualité de reporting et méthodologique des méta-analyses

Les groupes PRISMA et non PRISMA ont été comparés, afin de voir si l'adoption de PRISMA a une influence sur l'adhérence aux scores PRISMA, AMSTAR et AMSTAR 2. Deux approches ont été adoptées pour ce faire : l'une est relative aux scores globaux des listes, l'autre porte sur chacun des items pris individuellement.

Pour les scores globaux, il s'agissait simplement de comparer les groupes PRISMA et non PRISMA, soit avec un test t pour échantillons indépendants, soit avec un test de Wilcoxon. La normalité des distributions des scores AMSTAR, AMSTAR 2 et PRISMA ont été testées avec le test de Shapiro-Wilk. En l'occurrence, les scores de ces listes ne sont pas distribués normalement (PRISMA : W = 0.967099, p = 0.0005 ; AMSTAR : W = 0.966918, p = 0.0004 ; AMSTAR 2 : W = 0.981851, p = 0.0255). La comparaison des groupes s'est donc faite avec le test de Wilcoxon, dans lequel la taille d'effet se calcule selon la formule $r = Z/\sqrt{N}$ (ou N représente l'ensemble de l'échantillon), et s'interprète comme un coefficient de corrélation de Pearson.

Les résultats des tests de Wilcoxon ont été produits à titre indicatifs, mais l'usage des scores globaux des listes n'est pas très informatif, voire même déconseillée (e.g. Shea et al., 2017). Des tests exacts de Fisher ont ainsi été réalisés pour évaluer plus précisément l'impact de l'utilisation (ou non) de PRISMA sur les scores PRISMA, AMSTAR et AMSTAR 2, et il a ainsi été produit un test exact de Fisher par item (pour rappel, les items pouvaient être cotés comme « validés » ou « non-validés » ; dans le cas d'AMSTAR 2 ou il était également possible de coter certains items comme étant « partiellement validés », ces réponses ont été regroupées avec

les items validés). Etant donné le nombre de tests, il était évidemment nécessaire de contrôler l'inflation du risque d'erreur de première espèce. La correction de Bonferroni a ainsi été appliquée aux trois listes, et consiste à diviser l'alpha par le nombre de comparaisons à évaluer afin de réduire le seuil de significativité statistique. Pour PRISMA, l'alpha a ainsi été corrigé à 0.002 (0.05/27) ; il a été corrigé à 0.005 pour AMSTAR (0.05/11) et à 0.003 pour AMSTAR 2 (0.05/16). Le coefficient phi (ϕ) a été rapporté pour chacun des tests, afin de quantifier la magnitude de la relation entre la validation de l'item et l'utilisation de PRISMA. Il s'interprète comme un coefficient de corrélation de Pearson, et sa formule usuelle est :

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

et revient aussi à une corrélation de Pearson effectuée sur deux variables binaires.

3) Résultats

AMSTAR

Résultats généraux

La médiane de l'ensemble des scores AMSTAR est égale à 5 [intervalle interquartile = 3 ; 6]. La figure 10 résume les résultats pour chacun des items de la liste.

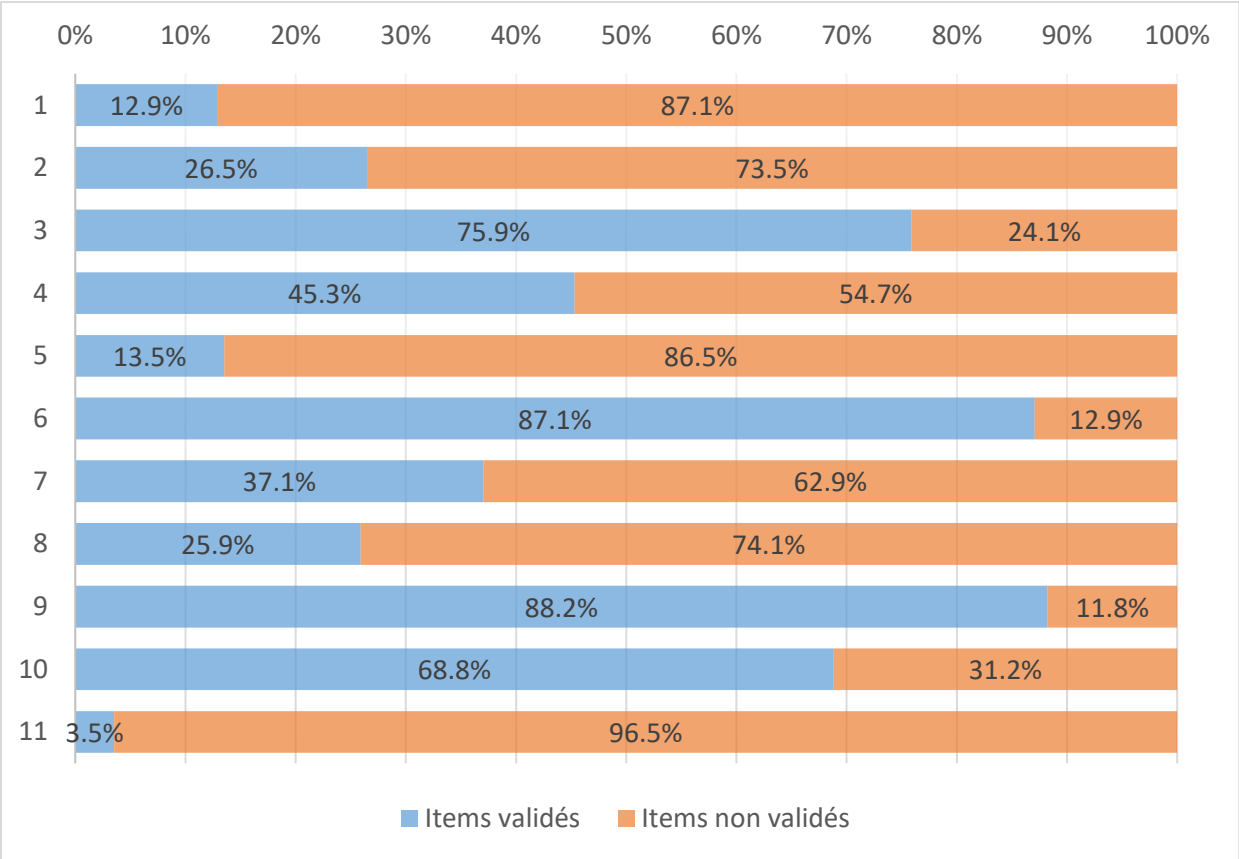


Figure 10 : Pourcentages d'adhérence pour chacun des items d'AMSTAR. Pour chaque ligne, les pourcentages renvoient respectivement aux méta-analyses qui ont rempli (en bleu) ou n'ont pas rempli (en orange) le critère correspondant.
1) Publication d'un protocole
2) Sélection des études et extraction des données par deux auteurs
3) Recherche exhaustive de la littérature
4) Recherche de la littérature non publiée
5) Liste des études inclues et exclues
6) Caractéristiques des études
7) Evaluation de la qualité des études
8) Conclusion incluant une discussion de la qualité des études
9) Méthodes de combinaison des données appropriées
10) Evaluation du biais de publication
11) Déclaration des conflits d'intérêts

Variation de l'adhérence à AMSTAR selon le facteur d'impact, l'index H et l'expérience du premier auteur

Le tableau 3 résume les effets du facteur d'impact, de l'index H et de l'expérience du premier auteur sur les scores AMSTAR.

	Tous les articles (Médiane, rang interquartile)	*Amstar < 5 (N = 75) (Médiane, rang interquartile)*	*Amstar ≥ 5 (N = 89) (Médiane, rang interquartile)*	*Odds ratio et intervalle de confiance (95%)*	*P-valeur*
Facteur d'impact	3.297 [2.290 ; 5.203]	2.976 [2.044 ; 4.519]	3.432 [2.500 ; 6.078]	1.102 [0.980 ; 1.240]	.104
Index H	4 [2 ; 11]	5.5 [2 ; 14]	4 [2 ; 8]	0.959 [0.922 ; 0.997]	.035*
Expérience du premier auteur	2 [1 ; 6]	3 [1 ; 6]	2 [1 ; 5.5]	1.039 [0.994 ; 1.087]	.091

Tableau 3 : Résultats de la régression logistique pour le test du facteur d'impact, de l'index H et de l'expérience du premier auteur sur les scores AMSTAR. Les astérisques désignent les résultats statistiquement significatifs.

AMSTAR 2

Résultats généraux

La médiane de l'ensemble des scores AMSTAR 2 est égale à 6,5 [intervalle interquartile = 5 ; 8,5]. La figure 11 résume les résultats pour chacun des items de la liste.

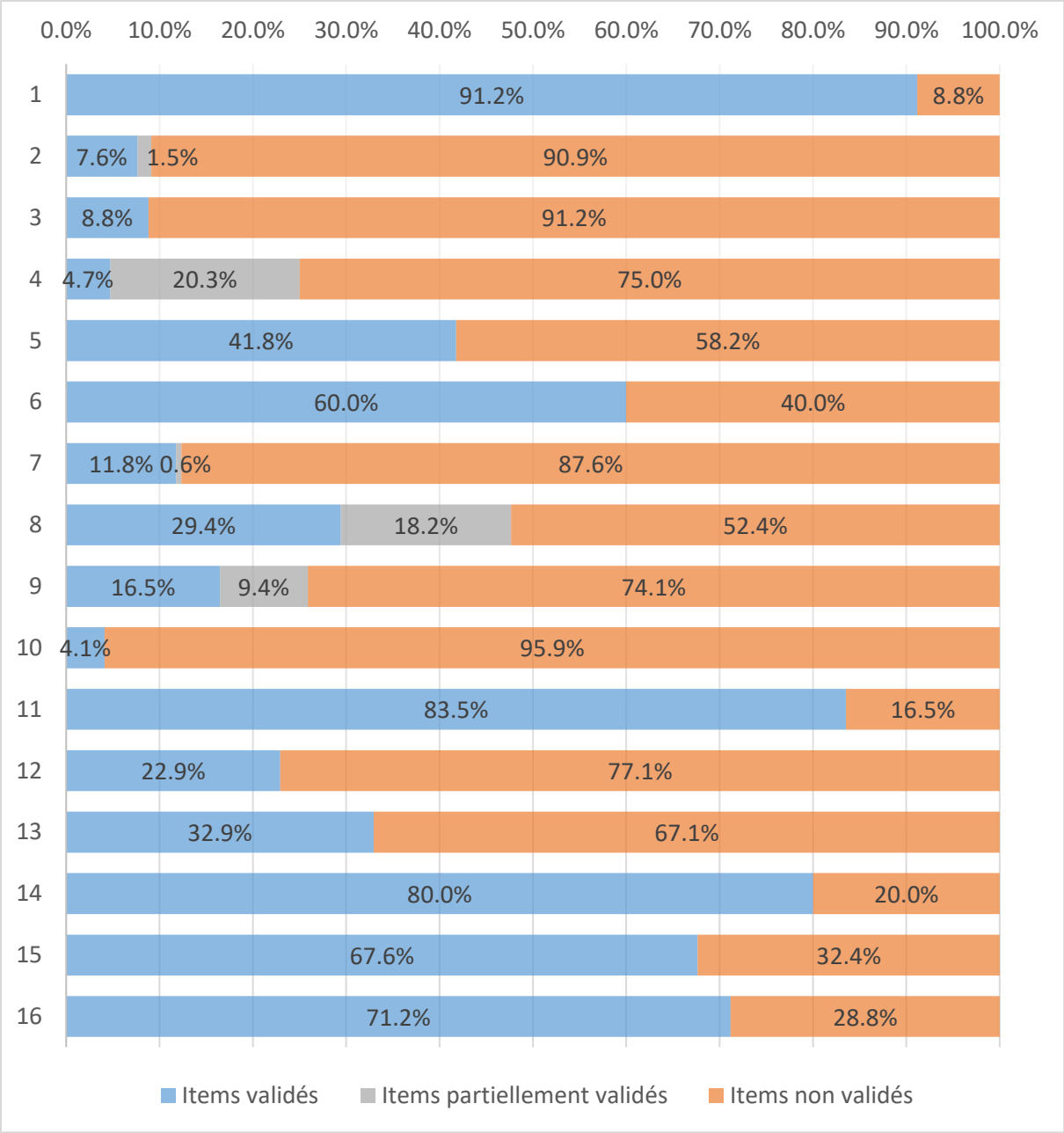


Figure 11 : Pourcentages d'adhérence pour chacun des items d'AMSTAR 2. Pour chaque ligne, les pourcentages renvoient respectivement aux méta-analyses qui ont rempli (en bleu), partiellement rempli (en gris) ou n'ont pas rempli (en orange) le critère correspondant.

1) Question et critères d'inclusion incluent les composants du PICO
2) Publication d'un protocole et justification des déviations par rapport à celui-ci
3) Justification des designs des études à inclure dans la synthèse
4) Recherche exhaustive de la littérature
5) Sélection des études par deux auteurs
6) Extraction des études par deux auteurs
7) Liste des études exclues et leur justifications
8) Description des études
9) Evaluation adéquate des risques de biais dans les études
10) Exploration des sources de financement des études
11) Méthodes de combinaison des données appropriées
12) Evaluation de l'impact des risques de biais sur les résultats de la synthèse
13) Prise en compte du risque de biais des études lors de la discussion des résultats
14) Justification et explication de l'hétérogénéité des résultats
15) Evaluation du biais de publication et discussion de son impact sur les résultats
16) Rapport des conflits d'intérêts et des sources de financement

Variation de l'adhérence à AMSTAR 2 selon le facteur d'impact, l'index H et l'expérience du premier auteur

Le tableau 2 résume les effets du facteur d'impact, de l'index H et de l'expérience du premier auteur sur les scores AMSTAR 2.

	Tous les articles (Médiane, rang interquartile)	*Amstar 2 < 6,5 (N = 66) (Médiane, rang interquartile)*	*Amstar 2 ≥ 6,5 (N = 98) (Médiane, rang interquartile)*	*Odds ratio et intervalle de confiance (95%)*	*P-valeur*
Facteur d'impact	3.297 [2.290 ; 5.203]	2.614 [2.005 ; 3.866]	3.508 [2.570 ; 6.257]	1.187 [1.035 ; 1.360]	.014*
Index H	4 [2 ; 11]	5 [2 ; 14]	4 [2 ; 9]	0.925 [0.882 ; 0.969]	.001*
Expérience du premier auteur	2 [1 ; 6]	3 [1 ; 6]	2 [1 ; 5.5]	1.062 [1.006 ; 1.120]	.029*

Tableau 4 : Résultats de la régression logistique pour le test du facteur d'impact, de l'index H et de l'expérience du premier auteur sur les scores AMSTAR 2. Les astérisques désignent les résultats statistiquement significatifs

Classement de la qualité des articles selon les critères AMSTAR 2

La figure 12 résume les résultats de la classification de Shea et al. (2017) appliquée à l'échantillon de cette étude.

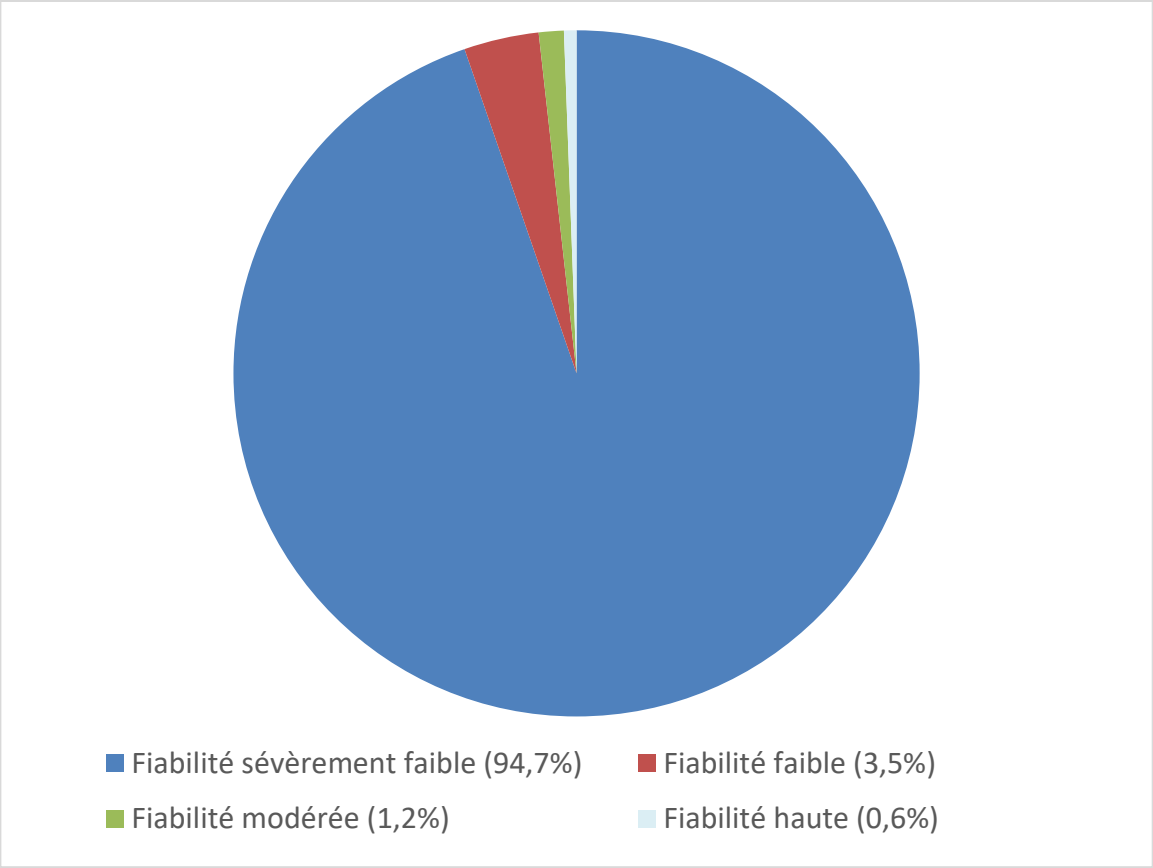


Figure 12 : Classement des articles selon les critères de Shea et al. (2017).

PRISMA

Résultats généraux

La médiane de l'ensemble des scores PRISMA est égale à 21 [intervalle interquartile = 18 ; 24]. La figure 13 résume les résultats pour chacun des items de la liste, et la figure 14 détaille les résultats pour les groupes PRISMA et non PRISMA.

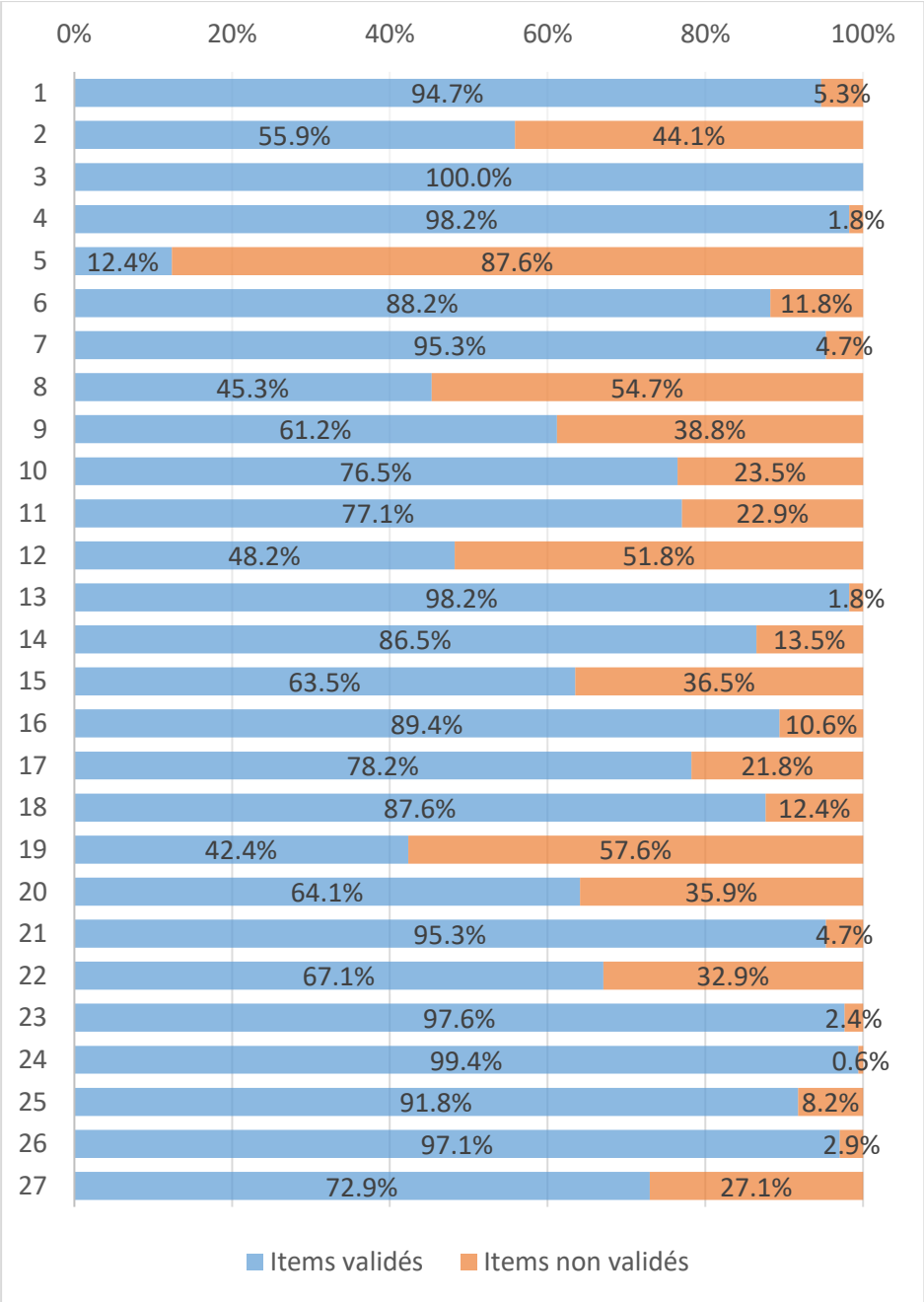


Figure 13 : Pourcentages d'adhérence pour chacun des items de PRISMA. Pour chaque ligne, les pourcentages renvoient respectivement aux méta-analyses qui ont rempli (en bleu), ou n'ont pas rempli (en orange) le critère correspondant.

1) Identification de l'article comme une « revue systématique » ou une « méta-analyse » ;
2) Abstract structuré ;
3) Raisons de l'étude ;
4) Questions clairement définies selon le PICO ;
5) Publication d'un protocole ;
6) Critères d'éligibilité des études ;
7) Source des études et date de la dernière recherche ;
8) Publication d'une stratégie de recherche complète ;
9) Façon de la sélection des études ;
10) Façon d'extraire les données ;
11) Liste de toutes les variables étudiées ;
12) Façon d'évaluer les risques de biais des études ;
13) Nature de la taille d'effet méta-analytique ;
14) Méthode de combinaison des données et évaluation de l'hétérogénéité ; 15) Evaluation des risques de biais qui peuvent altérer l'effet résumé, comme le biais de publication ;
16) Rapport des analyses additionnelles ;
17) Nombre d'études évaluées et raisons des exclusions ;
18) Caractéristiques des études ;
19) Présentation de l'évaluation des risques de biais ;
20) Pour chaque résultat méta-analytique, présentation de l'effet de chaque étude ainsi que les intervalles de confiance ;
21) Présentation des résultats de chaque méta-analyse, avec les intervalles de confiance et les mesures de l'hétérogénéité ;
22) Présentation des risques de biais pour l'ensemble des études (relatifs à l'item 15) ;
23) Résultats des analyses additionnelles (relatives à l'item 16) ;
24) Résumé des résultats et leur importance pratique ;
25) limite des études et de la méta-analyse ;
26) Discussion des résultats dans le contexte d'autres preuves et implication pour la recherche future ;
27) Déclaration des sources de financement et de support ainsi que le rôle du financeur.

Figure 14 : Pourcentages d'adhérence aux items de PRISMA selon le groupe des méta-analyses : a) Groupe PRISMA b) Groupe non PRISMA

Adoption de PRISMA selon le facteur d'impact, l'index H et l'expérience du premier auteur

Le tableau 5 résume les effets du facteur d'impact, de l'index H et de l'expérience du premier auteur sur la propension des auteurs à utiliser PRISMA ou non.

	Tous les articles (Médiane, rang interquartile)	*N'utilisent pas PRISMA (N = 78) (Médiane, rang interquartile)*	*Utilisent PRISMA (N = 86) (Médiane, rang interquartile)*	*Odds ratio et intervalle de confiance (95%)*	*P-valeur*
Facteur d'impact	3.297 [2.290 ; 5.203]	2.614 [2.044 ; 4.362]	3.657 [2.809 ; 6.190]	1.147 [1.019 ; 1.291]	.024*
Index H	4 [2 ; 11]	5 [2 ; 13]	4 [2 ; 9,5]	0.984 [0.949 ; 1.021]	.397
Expérience du premier auteur	2 [1 ; 6]	3 [1 ; 5]	2 [1 ; 6.5]	1.011 [0.981 ; 1.042]	.474

Tableau 5 : Résultats de la régression logistique pour le test du facteur d'impact, de l'index H et de l'expérience du premier auteur sur la propension à utiliser PRISMA. Les astérisques désignent les résultats statistiquement significatifs.

Comparaison entre les méta-analyses qui adhèrent à PRISMA et celles qui n'adhèrent pas à PRISMA au niveau des scores PRISMA, AMSTAR et AMSTAR 2

Utilisation de PRISMA et adhésion à PRISMA

Le test de Wilcoxon a révélé une différence entre les groupes PRISMA et non PRISMA quant à l'adhésion des items de la liste PRISMA ($Z = 6{,}694700$; $p < .0001$, $r = 0{,}513$). La figure 15 résume les résultats.

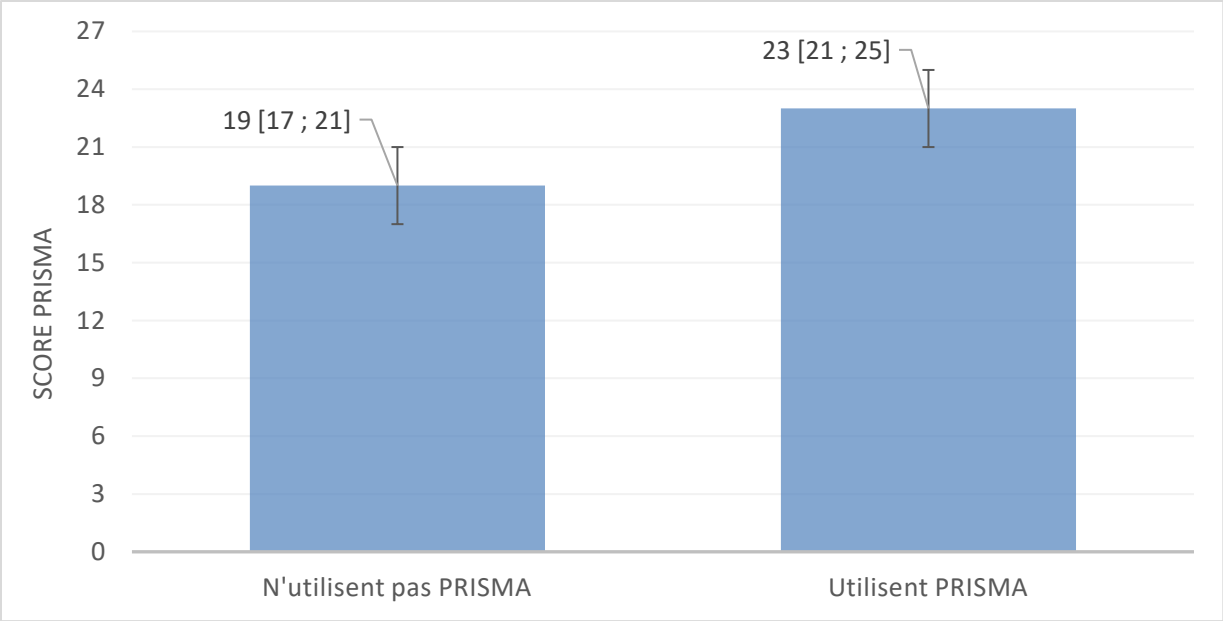


Figure 15 : Comparaison des groupes PRISMA et non PRISMA quant à l'adhérence à PRISMA. Les chiffres au-dessus des barres indiquent les médianes et les intervalles interquartiles.

Le tableau 6 résume les comparaisons entre les groupes PRISMA et non PRISMA pour chacun des items de PRISMA.

	Groupe PRISMA ou non (coefficient phi)	*P-valeur*
P1	0.0650	0.3100
P2	0.3867	<.0001*
P3	/	/
P4	0.1264	0.1460
P5	0.2907	0.0001*
P6	0.1248	0.0813
P7	0.2095	0.0053
P8	0.3732	<.0001*
P9	0.2916	0.0001*
P10	0.0784	0.2002
P11	-0.0462	0.7838

P12	0.3635	<.0001*
P13	0.0369	0.5443
P14	0.0628	0.2772
P15	0.2002	0.0069
P16	0.0180	0.5074
P17	0.4687	<.0001*
P18	0.2107	0.0049
P19	0.4083	<.0001*
P20	0.3368	<.0001*
P21	0.0982	0.1805
P22	0.1844	0.0122
P23	-0.0091	0.7327
P24	0.0725	0.5294
P25	0.0681	0.2733
P26	0.0246	0.5557
P27	0.1763	0.0162

Tableau 6 : Tests exacts de Fisher (unilatéraux à droite) comparant les niveaux d'adhérence à PRISMA entre les groupes PRISMA et non PRISMA. Les astérisques désignent les résultats significatifs à un seuil alpha adapté à 0.001. Pour l'item 3, le test n'a pas été effectué car toutes les méta-analyses ont adhéré au critère.

Utilisation de PRISMA et adhésion à AMSTAR

Le test de Wilcoxon a révélé une différence entre les groupes PRISMA et non PRISMA quant à l'adhésion des items de la liste AMSTAR [Z = 5.3893 ; p <.0001, r = 0,413]. La figure 16 résume les résultats.

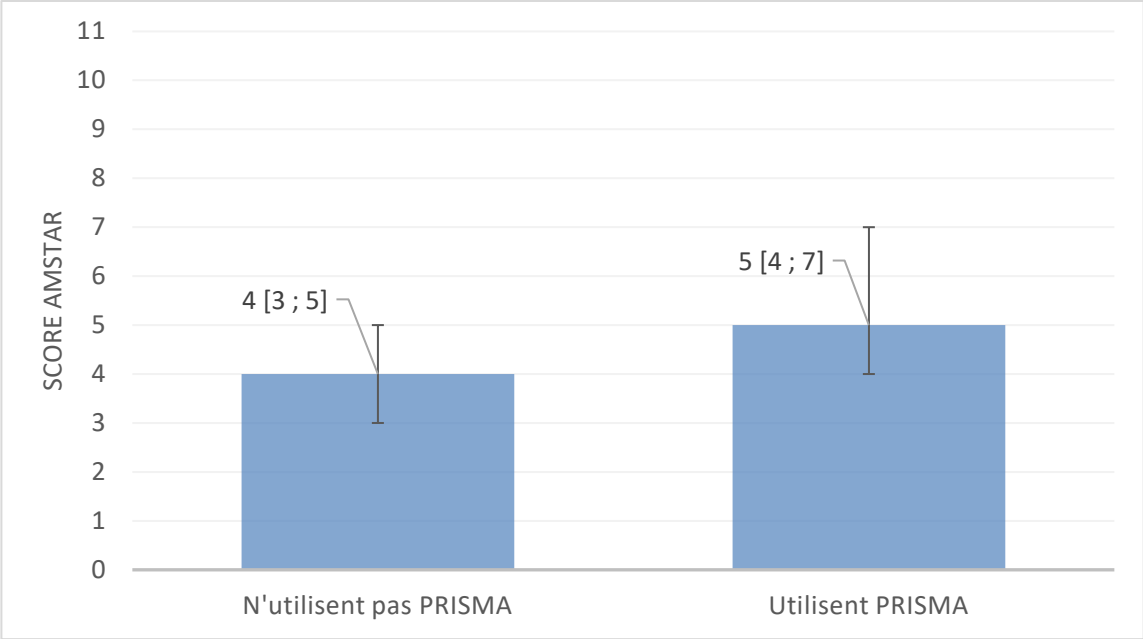


Figure 16 : Comparaison des groupes PRISMA et non PRISMA quant à l'adhérence à AMSTAR. Les chiffres au-dessus des barres indiquent les médianes et les intervalles interquartiles.

Le tableau 7 résume les comparaisons entre les groupes PRISMA et non PRISMA pour chacun des items d'AMSTAR.

	Utilisation de PRIMA ou non (coefficient phi)	*P-valeur*
A1	0.3036	<.0001*
A2	0.3158	<.0001*
A3	0.2836	0.0002*
A4	-0.1239	0.9619
A5	0.1784	0.0175
A6	0.2582	0.0005*
A7	0.3746	<.0001*
A8	0.3308	<.0001*
A9	0.1980	0.0082
A10	0.1003	0.1267
A11	0.0751	0.2869

Tableau 7 : Tests exacts de Fisher (unilatéraux à droite) comparant les niveaux d'adhérence à AMSTAR entre les groupes PRISMA et non PRISMA. Les astérisques désignent les résultats significatifs à un seuil alpha adapté à 0.005.

Utilisation de PRISMA et adhésion à AMSTAR 2

Le test de Wilcoxon a révélé une différence entre les groupes PRISMA et non PRISMA quant à l'adhésion des items de la liste AMSTAR 2 (Z = 5.9061 ; p <.0001, r = 0,130). La figure 17 résume les résultats.

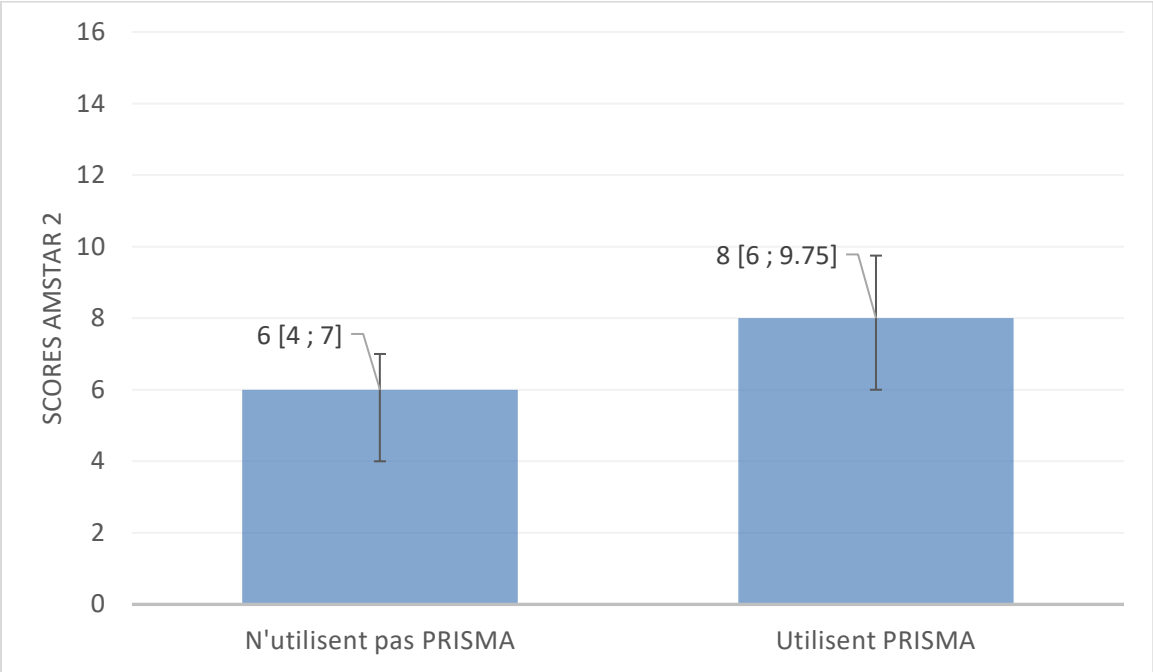


Figure 17 : Comparaison des groupes PRISMA et non PRISMA quant à l'adhérence à AMSTAR 2. Les chiffres au-dessus des barres indiquent les médianes et les intervalles interquartiles.

Le tableau 8 résume les comparaisons entre les groupes PRISMA et non PRISMA pour chacun des items d'AMSTAR 2.

	Utilisation de PRIMA ou non (coefficient phi)	*P-valeur*
A1	0.1271	0.0813
A2	0.2884	0.0001*
A3	0.0391	0.4043
A4	0.0418	0.3481
A5	0.3486	<.0001*
A6	0.0241	0.4380
A7	0.1632	0.0285
A8	0.3056	<.0001*
A9	0.4134	<.0001*
A10	0.0419	0.4347
A11	0.1962	0.0084
A12	0.2984	<.0001*
A13	0.3673	<.0001*
A14	0.1768	0.0165
A15	0.0978	0.1332
A16	0.1576	0.0291

Tableau 8 : Tests exacts de Fisher (unilatéraux à droite) comparant les niveaux d'adhérence à AMSTAR entre les groupes PRISMA et non PRISMA. Les astérisques désignent les résultats significatifs à un seuil alpha adapté à 0.003.

4) Discussion

Ce travail met en évidence que la qualité méthodologique et la qualité de reporting des méta-analyses qui concernent des questions psychologiques ou qui relèvent de disciplines apparentées sont loin d'être optimales. La magnitude des résultats est globalement assez importante pour que l'on soit réellement interpellé par ce que représentent réellement les conclusions de ces études, qui sont censées représenter le plus haut niveau de la fiabilité scientifique.

Plus précisément, si nous reprenons la classification des items critiques relative à AMSTAR 2 (à savoir l'enregistrement d'un protocole, l'exhaustivité de la recherche de la littérature, les justifications apportées quant à l'exclusion de certaines études, l'évaluation des risques de biais, l'adéquation des méthodes de combinaison des données, la prise en compte des risques de biais dans les études lors de la discussion des résultats et l'évaluation de la présence et de l'impact du biais de publication ; Shea et al., 2017), 94.7% des méta-analyses comportent des manquements à au moins deux de ces items critiques, ce qui selon cette classification signifie qu'elles sont d'une fiabilité *sévèrement faible*. Seules 1.2% des méta-analyses sont d'une fiabilité *modérée*, et seules 0.6% représentent un haut niveau de fiabilité. Les résultats descriptifs d'AMSTAR et de PRISMA vont dans le même sens.

Parmi les problèmes les plus prégnants se trouvent, selon AMSTAR et AMSTAR 2, la publication d'un protocole (respectivement 87,1% et 90,9% des méta-analyses qui n'ont pas adhéré à l'item 1 d'AMSTAR et à l'item 2 d'AMSTAR 2), la sélection et l'extraction des données par deux auteurs (73,5% pour l'item 2 d'AMSTAR ; respectivement 58,2% et 40,0% pour les items 5 et 6 d'AMSTAR 2), l'exhaustivité de la recherche de la littérature (item 3 et 4 d'AMSTAR, 24,1% et 54,7% - l'item 4 porte spécifiquement sur la recherche de la littérature non publiée - ; item 4 d'AMSTAR 2, 75,0%), la publication de la liste des études inclues et exclues (item 5 d'AMSTAR, 86,5% ; item 7 d'AMSTAR 2, 87,6%), l'évaluation et la discussion des risques de biais dans les études (item 7 et 8 d'AMSTAR, 62.9% et 74,1% ; item 9 , 12 et 13 d'AMSTAR 2, 74,1%, 77,1% et 67,1%) et la description des sources de financement des études et de la méta-analyse (item 11 d'AMSTAR, 96.5% ; item 10 et 16 d'AMSTAR 2, 95.9% et 28.8%). Comme déjà développé dans l'introduction de ce mémoire, la sélection et l'extraction des

données par deux auteurs doit permettre de neutraliser les biais et les erreurs que les chercheurs peuvent commettre lors de ces étapes fastidieuses de la méta-analyse. La publication des études inclues et exclues ajoute un degré de transparence toujours appréciable lorsqu'il s'agit d'évaluer la pertinence des choix d'inclusion et d'exclusion, voire de répliquer l'analyse ; l'évaluation et la discussion des risques de biais est évidemment indispensable, car une synthèse basée sur des études de mauvaise qualité risque de ne pas être très informative, et ce même si les autres aspects méthodologiques de celle-ci sont de bonne facture. En ce qui concerne la description des sources de financement, AMSTAR 2 permet de savoir que le problème est surtout relatif à l'évaluation de ces intérêts dans les études inclues dans les méta-analyses (item 10), et moins au niveau des méta-analyses en elles-mêmes (item 16). AMSTAR ne permet pas cette distinction (pour être validé, l'item 11 d'AMSTAR doit remplir deux conditions, qui sont respectivement la déclaration des sources de support au niveau des méta-analyses et celles relatives aux études individuelles, d'où la faible adhérence – 3,5% - des études à cet item). Enfin, l'importance de l'exhaustivité de la recherche des articles est absolument capitale. En réalité, le caractère d'exhaustivité est censé être une particularité fondamentale des revues systématiques et des méta-analyses ; c'est précisément cette qualité qui est censée rendre ces dernières plus fiables que d'autres travaux de synthèse (Boreinstein et al., 2009). Le manque d'exhaustivité n'est ainsi pas seulement un problème méthodologique, mais également un manquement à l'esprit même de la méta-analyse.

La publication d'un protocole, l'un des grands problèmes relevé dans les résultats, mérite un certain développement ; nous y reviendrons. Pour le moment, disons simplement que l'enregistrement des protocoles de recherche est une solution globale proposée par certains chercheurs afin d'améliorer la transparence et la reproductibilité de la recherche (e.g. Chambers, 2017).

Influence du facteur d'impact, de l'index H et de l'expérience du premier auteur sur la qualité méthodologique et de reporting

Le facteur d'impact, l'index H et l'expérience du premier auteur ressortent comme des variables influençant de façon statistiquement significative les scores AMSTAR (Index H, p =

0.0354) et AMSTAR 2 (Facteur d'impact, p = 0.0140 ; index H, p = 0.0011 ; expérience du premier auteur, p = 0.0288). En ce qui concerne la propension des auteurs à utiliser PRISMA ou non, le facteur d'impact rapporte également un p significatif (p = 0.0235).

Cependant, tous les odds ratio associés à ces résultats sont assez faibles (celui qui a la plus grande magnitude, toutes listes confondues et parmi les résultats statistiquement significatifs, est égal à 1.187 [95% CI = 1.035 ; 1.360] et concerne l'influence du facteur d'impact sur les scores AMSTAR 2). Certains d'eux suggèrent même une influence *négative* de l'index H sur les scores AMSTAR et AMSTAR 2 (ils sont respectivement égaux à 0.959 et à 0.925, ce qui donnerait en français pour le premier : « à chaque augmentation d'une unité de l'index H, la « chance » d'obtenir un score reflétant une bonne qualité méthodologique diminue d'un facteur de 0.041 »). Les odds ratio sont connus pour ne pas être nécessairement aisés à interpréter ; en ce qui concerne leur magnitude, Cohen (1988) a cependant suggéré qu'un odds ratio de 1.50 représente un petit effet, un odds ratio de 2.50 représente un effet moyen et un odds ratio de 4.30 représente un grand effet. Si l'on suit cette classification, les effets que nous retrouvons ici sont petits, voire très petits, qu'ils suggèrent une amélioration ou une dégradation des scores AMSTAR et AMSTAR 2, et le même raisonnement peut être effectué quant à l'influence du facteur d'impact, de l'index H et de l'expérience du premier auteur quant à la propension à utiliser PRISMA (Odds ratio = 1.147, p = 0.0235). En somme cela signifie potentiellement que si le facteur d'impact, l'index H et l'expérience du premier auteur ont une véritable influence sur les résultats et la propension à utiliser PRISMA, la magnitude de cet impact est faible. Si ces résultats ont une quelconque forme d'importance au niveau pratique, on peut se risquer à dire que la magnitude de ces effets, tels que retrouvés ici, va à l'encontre des idées largement répandues qui lient crucialement le facteur d'impact et l'index H à la qualité des publications scientifiques, et à fortiori des méta-analyses, et qui déterminent de façon visiblement exagérée jusqu'à la carrière de certains chercheurs. Nous avons déjà largement parlé du facteur d'impact et de l'index H dans l'introduction, et de ce pourquoi il y a des raisons de penser qu'ils ne sont pas crucialement liés à la qualité des publications de façon générale. En ce qui concerne l'expérience du premier auteur, cela signifie peut être que les méta-analystes devraient suivre des formations particulières, ce qui n'est pas nécessairement étonnant compte tenu de la charge de travail et de la complexité associées à la réalisation de certaines de leurs étapes.

Pour résumer, nous pouvons postuler, sur la base de nos résultats, que le facteur d'impact, l'index H et l'expérience du premier auteur n'influencent pas de façon cruciale la qualité méthodologique des méta-analyses, ni la propension des chercheurs à utiliser la liste PRISMA.

Influence de l'utilisation de PRISMA sur les scores PRISMA, AMSTAR et AMSTAR 2

De façon générale, les items de PRISMA pour lesquels l'adhérence est la plus faible touchent les mêmes domaines que ceux d'AMSTAR et d'AMSTAR 2, mais les problèmes de reporting peuvent être assez différents selon que les auteurs des méta-analyses déclarent avoir utilisé PRISMA ou non (cfr figure 14). Selon nos résultats, l'adoption de PRISMA semble avoir un effet positif à la fois sur la qualité de reporting et sur la qualité méthodologique des méta-analyses, et plus précisément au niveau de certains items particuliers des trois check-lists. Pour les résultats significatifs et toutes listes confondues, les coefficients phi révèlent des liens intéressants et dont l'importance pratique est certainement substantielle[1] : le plus petit d'eux est égal à 0.2582 et concerne l'item 6 d'AMSTAR – le rapport des caractéristiques des études individuelles - et le plus grand est égal à 0.4687 et concerne l'item 17 de PRISMA – la publication du nombre des méta-analyses évaluées et inclues, ainsi que les raisons des exclusions de celles qui ont été rejetées. Le tableau 9 résume les résultats significatifs des tests de Fisher. L'utilisation de PRISMA semble avoir une influence sur les résultats des trois check-lists d'autant plus intéressante que beaucoup des items qui bénéficient de cet avantage sont aussi parmi ceux pour lesquels l'adhérence est la plus faible dans les résultats descriptifs (par exemple, on retrouve beaucoup d'items relatifs à l'évaluation et au reporting des risques de biais dans les études, ou encore ceux qui traitent de la publication d'un protocole).

[1] Cohen (1988) propose que les coefficients de corrélation de .10 représentent de petits effets, que les coefficients de .30 représentent des effets modérés et que les coefficients de .50 représentent de grands effets.

PRISMA	*AMSTAR*	*AMSTAR 2*
2) Abstract structuré	*1) Publication d'un protocole*	*2) Publication d'un protocole et justification des déviations par rapport à celui-ci*
5) Publication d'un protocole	*2) Sélection des études et extraction des données par deux auteurs*	*5) Sélection des études par deux auteurs*
8) Stratégie de recherche complète pour au moins une base de donnée	*3) Recherche exhaustive de la littérature*	*8) Description des études*
9) Processus de sélection des études	*6) Caractéristiques des études*	*9) Evaluation adéquate des risques de biais dans les études*
12) Méthodes d'évaluation des risques de biais	*7) Evaluation de la qualité des études*	*12) Evaluation de l'impact des risques de biais sur les résultats de la synthèse*
17) Nombre d'études évaluées pour inclusion, et raison des exclusions	*8) Conclusion incluant une discussion de la qualité des études*	*13) Prise en compte du risque de biais des études lors de la discussion des résultats*
19) Présentation de l'évaluation des risques de biais		
20) Pour chaque résultat méta-analytique, présentation de l'effet de chaque étude ainsi que les intervalles de confiance		

Tableau 9 : résumé des effets statistiquement significatifs relatifs aux tests exacts de Fisher, pour chacun des items de PRISMA, d'AMSTAR et d'AMSTAR 2.

5) Conclusion, et quelques perspectives générales

Les méta-analyses qui concernent des questions psychologiques ou relevant de disciplines apparentées, qui sont censées représenter le plus haut niveau de la preuve scientifique, sont dans la pratique d'une qualité, tant de reporting que méthodologique, qui est suboptimale. Des facteurs traditionnellement reconnus comme témoignant de la qualité des publications – le facteur d'impact et l'index H – ne montrent pas ici un effet crucial sur la qualité des méta-analyses, de même que l'expérience du premier auteur. Relativement à ces mêmes variables, on ne peut pas non plus affirmer qu'elles ont un effet important sur la propension des auteurs de méta-analyses à adopter la liste PRISMA ; l'adoption de cette dernière semble pourtant avoir un effet positif sur la qualité des méta-analyses selon nos résultats. Les problèmes relevés dans ce mémoire font écho à des investigations similaires ayant été menées dans le domaine biomédical.

Si l'adoption de PRISMA à grande échelle peut sans doute améliorer la qualité des méta-analyses de façon générale, il ne s'agit que d'une solution partielle ; il reste indispensable de porter un regard plus vaste sur la recherche telle qu'elle se déroule actuellement. Comme dit dans l'introduction, les méta-analyses s'inscrivent dans un contexte scientifique affecté par des pratiques de recherche discutables et des problèmes méthodologiques qui réduisent la fiabilité de la littérature scientifique ; les méta-analyses se basant sur cette dernière risquent de rapporter des effets déformés de leur véritable signification, étant donné l'ampleur du bais de publication, des problèmes liés à la puissance statistique et aux pratiques de recherches discutables. On peut faire le pari que la sensibilisation des chercheurs aux problèmes généraux que la méta-recherche a soulevé et tant que possible, leur résolution, aura à long terme une influence sur la qualité des méta-analyses, d'une part parce que les études qu'elles incluraient seraient alors davantage susceptibles de rapporter des effets se rapprochant de la vérité, et d'autre part parce que les méta-analystes seraient poussés, d'une manière ou d'une autre, à accorder un soin particulier à la réalisation de leurs investigations. Actuellement, force est de constater qu'on est encore loin de cet idéal. Quelques initiatives ont cependant déjà vu le jour et certaines ont rapporté des effets intéressants.

On peut par exemple nous attarder sur l'enregistrement des protocoles des études, qui se trouve aussi être l'un des problèmes les plus importants rapporté dans ce mémoire. En fait, la publication des protocoles est une solution contre le biais de publication et l'analyse flexible des données. Ils permettent en effet d'une part de contrer la survenue d'une série d'inconduites scientifiques relatives à l'arrangement des données, dans le sens où les analyses sont spécifiées à l'avance, et les auteurs doivent alors s'y tenir et déclarer clairement les nouvelles analyses. D'autre part, les journaux qui ont publié un protocole sont ensuite tenus de publier l'article complet, quels que soient ses résultats. Si, pour une raison ou une autre, le protocole ou l'article sont rejetés, ou que les auteurs décident de se retirer du processus, un enregistrement du retrait est publié et accessible aux autres chercheurs (figure 18). Un désavantage de cette formule est qu'elle peut s'étaler sur un délai relativement important, ce qui n'est pas toujours compatible avec les obligations des chercheurs, spécialement des doctorants. Il existe dans ce cas une seconde forme d'enregistrement des études, qui ne nécessite pas de revue par les pairs ; la base de données Open Science Framework (OSF), dans laquelle le projet dont ce mémoire est issu a été publié, permet de tels enregistrements.

On peut bien sûr se demander à quel point l'enregistrement des protocoles améliore la qualité de la littérature scientifique. Cela est encore rare dans les sciences humaines de façon générale, mais beaucoup plus répandu dans certains champs de la médecine, dans lesquels cette question commence à être adressée. Par exemple, Kaplan et Irvin (2015) ont trouvé que depuis l'enregistrement obligatoire des essais cliniques financés par le National Heart, Lung, and Blood Institute (NHLBI) en 2000, le pourcentage d'essais rapportant un effet significatif est passé de 57% (17 de 30 essais) à 8% (2 de 25 essais). Les avantages de l'enregistrement des protocoles de méta-analyses sont similaire à celle des autres types d'études ; en effet de nombreux choix et analyses sont possibles au cours de leur réalisation, et peuvent aboutir à des résultats différents[2]. Au moment de la rédaction de ce mémoire, 125 journaux proposent la publication de protocoles, en psychologie et dans d'autres disciplines (Center for Open Science, 2018 ; la liste des journaux est mise à jour régulièrement).

[2] La base de données « PROSPERO » est dédiée à l'enregistrement des revues systématiques et des méta-analyses.

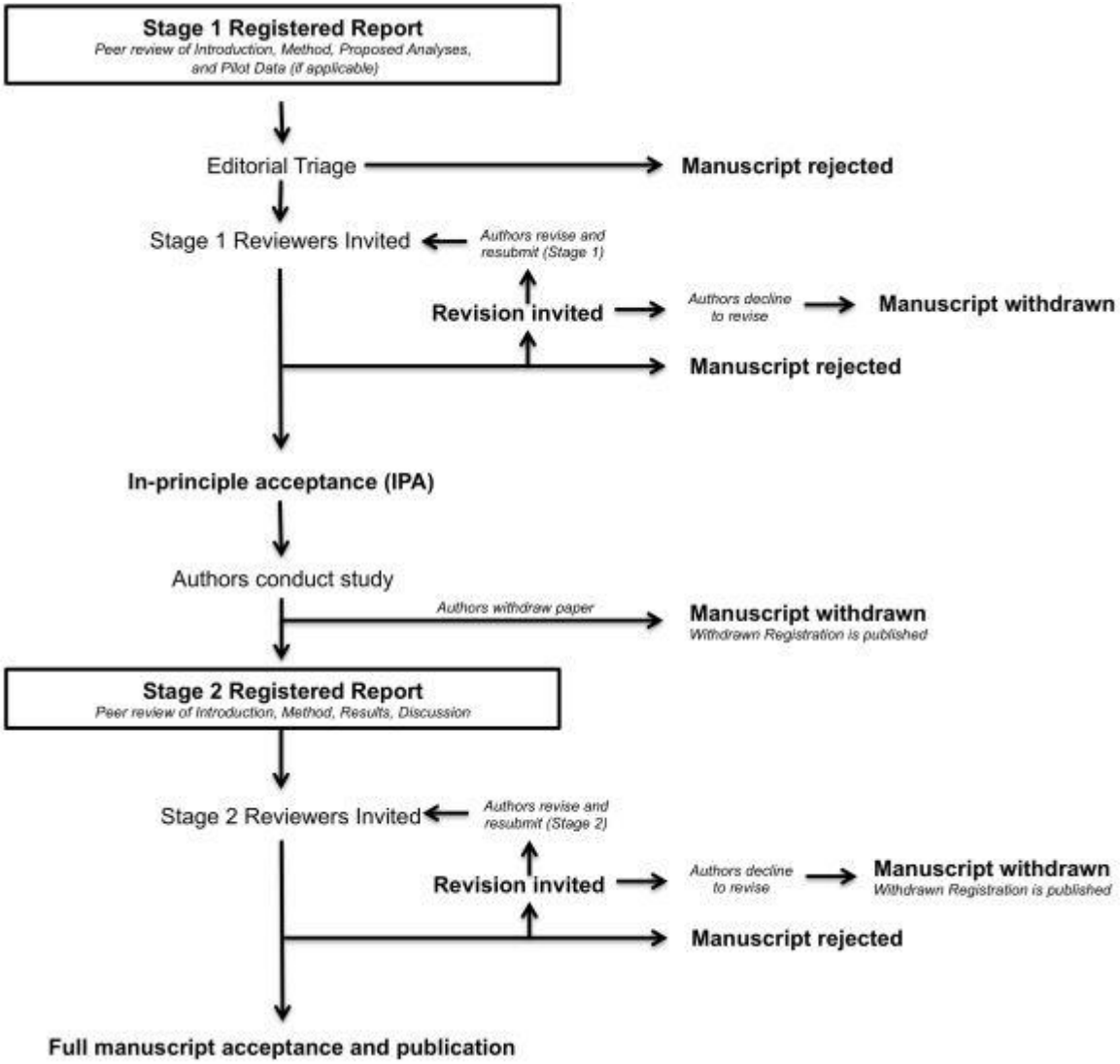


Figure 18 : Le processus de soumission des protocoles pour le journal *Cortex* et pour d'autres journaux. Le processus comprend deux phases : dans une première étape, les auteurs soumettent une introduction, une méthode et un plan d'analyses avant la collection des données. Si le protocole est accepté, ce dernier est « accepté par principe » par le journal (*In-principle acceptance,* « IPA »). Les chercheurs collectent ensuite leurs données, et soumettent dans une deuxième étape un nouveau manuscrit contenant les résultats et la discussion, et s'en suit finalement la publication de l'article après une nouvelle étape de revue par les pairs. Tiré de Chambers (2017).

Il existe bien sûr d'autres formes de solutions qui ont été proposées pour améliorer la qualité de la littérature scientifique. Munafò et al. (2017) listent par exemple une série de mesures qui pourraient améliorer la reproductibilité de la science ; ils insistent par exemple sur l'importance de la formation statistique et méthodologique des chercheurs, spécialement en ce qui concerne l'interprétation des tailles d'effet, la puissance ou encore la signification de la p-valeur, l'amélioration du reporting des études (avec des listes telles que PRISMA pour les méta-analyses), le renforcement de l'*Open Science*, ou encore l'introduction d'incitations à de

bonnes pratiques de recherche, comme l'adoption de certains « badges » que des journaux attribuent à des articles selon différentes formes de mérites (par exemple, le journal *Psychological Science* a introduit en janvier 2014 des badges récompensant les articles dont les auteurs partageaient leurs donnéesmmmm. Avant l'introduction de ces badges, c'était le cas de moins de 5% des articles ; environ un an après l'introduction de cette mesure, ce chiffre est passé à près de 40%). Les initiatives pour lesquelles des études ont déjà questionné l'efficacité semblent réellement avoir une influence positive sur la transparence et la qualité de la recherche, même s'il faudra encore de nombreuses années avant que ces mesures ne soient davantage généralisées.

Bibliographie

Adler, R., Ewing, J., & Taylor, P. (2008). Joint committee on quantitative assessment of research: Citation statistics. *Australian Mathematical Society Gazette*, *35*(3), 166-188.

Barbour, V., Clark, J., Peiperl, L., Veitch, E., Wong, M., & Yamey, G. (2008). Making sense of non-financial competing interests. *PLoS Medicine*, *5*(9), 1299–1301. doi:10.1371/journal.pmed.0050199

Begley, C. G., Buchan, A. M., & Dirnagl, U. (2015). Robust research: Institutions must do their part for reproducibility. *Nature*, *525*(7567), 25–27. doi:10.1038/525025a

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531-533. doi:https://doi.org/10.1038/483531a

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open, 7*(2), 1–7. doi:10.1136/bmjopen-2016- 012545

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ : Wiley-Blackwell.

Bornmann, L., & Daniel, H. D. (2009). The state of h index research: Is the h index the ideal way to measure research performance? *EMBO Reports*, *10*(1), 2–6. doi:10.1038/embor.2008.233

Broad, W., & Wade, N. (1994). *La souris truquée. Enquête sur la fraude scientifique.* Paris, France: Points sciences.

Bronner, G. (2013). *La démocratie des crédules*. Paris, France: Presses universitaires de France.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi:10.1038/nrn3475

Casadevall, A., & Fang, F. C. (2015). Impacted science: Impact is not importance. *MBio*, *6*(5), 1–4. doi:10.1128/mBio.01593-15

Center for Open Science (2018). *Registered reports: Peer review before results are known to align scientific values and practices*. Retrieved from https://cos.io/rr/

Chalmers, I. (2007). The lethal consequences of failing to make use of all relevant evidence about the effects of medical treatments: the need for systematic reviews. In P.M. Rothwell (Ed.), *Treating individuals: From randomised trials to personalised medicine* (pp. 37-58). New York, NY: Elsevier.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice.* Princeton, NJ: Princeton University Press.

Cochrane. (2018). *Glossary*. Retrieved from https://community.cochrane.org/glossary#letter-B

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145–153.

Cullis, P. S., Gudlaugsdottir, K., & Andrews, J. (2017). A systematic review of the quality of conduct and reporting of systematic reviews and meta-analyses in paediatric surgery. *PLoS ONE*, *12*(4), 1–24. doi:https://doi.org/10.1371/journal.pone.0175213

Ebrahim, S., Bance, S., Athale, A., Malachowski, C., & Ioannidis, J. P. A. (2016). Meta-analyses with industry involvement are massively published and report no caveats for antidepressants. *Journal of Clinical Epidemiology*, *70*, 155–163. doi:10.1016/J.JCLINEPI.2015.08.021

Ellis, P. D. (2010). *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results***.** Cambridge, Royaume-Uni: Cambridge University Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

Fanelli, D. (2010). ‘“ Positive ”’ results increase down the hierarchy of the sciences. *PLoS ONE, 5*(4), e10068. doi:10.1371/journal.pone.0010068

Fang, F. C., Steen, R.G., & Casadevall, A. (2013). Misconducts accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *110*(3), 17028–17033. doi:https://doi.org/10.1073/pnas.1212247109

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE*, *9*(10), e109019. doi:10.1371/journal.pone.0109019

Gagnier, J. J., & Kellam, P. J. (2013). Reporting and methodological quality of systematic reviews in the orthopaedic literature. *The Journal of Bone and Joint Surgery. American volume, 77*, 1–7. doi:http://dx.doi.org/10.2106/JBJS.L.00597

Garfield, E. (2006). The history and meaning of the journal impact factor. *Journal of he American Medical Association, 295*(1), 90-93. doi:10.1001/jama.295.1.90

Gómez-García, F., Ruano, J., Aguilar-Luque, M., Gay-Mimbrera, J., Maestre-Lopez, B., Sanz-Cabanillas, J. L., … Isla-Tejera, B. (2017). Systematic reviews and meta-analyses on psoriasis: role of funding sources, conflict of interest and bibliometric indices as predictors of methodological quality. *British Journal of Dermatology*, *176*(6), 1633–1644. doi:10.1111/bjd.15380

Grapov, D. (2013, January 17). *Power Calculations – relationship between test power, effect size and sample size*. Retrieved from https://imdevsoftware.wordpress.com/2013/01/17/255/

Hartoupian, G. (2016). *La petite histoire des grandes impostures scientifiques*. Paris, France: Editions du Chêne.

Hasan, H., Muhammed, T., Yu, J., Taguchi, K., Samargandi, O. A., Howard, A. F., … Goddard, K. (2017). Assessing the methodological quality of systematic reviews in radiation oncology : A systematic review. *Cancer Epidemiology*, *50*, 141–149. doi:https://doi.org/10.1016/j.canep.2017.08.013

Higgins, J. P. T., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions*. Hoboken, NJ: John Wiley & Sons.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, *102*(46), 16569-16572. doi:https://*doi*.org/*10.1073*/*pnas*.*0507655102*

Howick, J., Chalmers. I., Glasziou, P., Greenhalgh, T., Heneghan, C., Liberati, A., … Thornton, H. (2018) *Levels of evidence: Introductory document.* Retrieved from https://www.cebm.net/2011/06/2011-oxford-cebm-levels-evidence-introductory-document/

Ioannidis, J. P. A. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Quarterly*, *94*(3), 485–514. doi:10.1111/1468-0009.12210

Ioannidis, J. P. A. (2018). Meta-research: Why research on research matters. *PLOS Biology*, *16*(3), e2005468. doi:https://doi.org/10.1371/journal.pbio.2005468

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. doi:10.1177/0956797611430953

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PloS one*, *10*(8), e0132382. doi:https://doi.org/10.1371/journal.pone.0132382

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, *9*(9), 1–8. doi:10.1371/journal.pone.0105825

Leclercq, V., Beaudart, C., Ajamieh, S., Rabenda, V., Tirelli, E., & Bruyère, O. (2018). *Assessment of the reporting and methodological qualities and associated factors of a sample of meta-analyses recently indexed in PsycINFO (2016)*. Retrieved from https://osf.io/hjybx/

Lipsey, M. W. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. Thousand Oaks, CA: SAGE Publications.

Lozano, G. A., & Larivière, V. (2012). The weakening relationship between the impact factor and papers ' citations in the digital age. *Journal of the American Society for Information Science and Technology, 63*(11), 2140–2145. doi:10.1002/asi.22731

Maiväli, Ü. (2015). *Interpreting biomedical science: Experiment, evidence, and belief*. Cambridge, MA: Academic Press.

Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science, 122*(3159), 108–111. doi:10.1126/science.122.3159.108

Medina, J., & Cason, S. (2017). No evidential value in samples of transcranial direct current stimulation (tDCS) studies of cognition and working memory in healthy populations. *Cortex*, *94*, 131–141. doi:http://dx.doi.org/10.1016/j.cortex.2017.06.021

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502–1505. doi:10.1126/science.1255484

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., … Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. doi:10.1371/ journal.pmed.1000097

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. doi:https://doi.org/10.1038/s41562-016-0021

Brembs, B., Button, K., & Munafò, M. (2013). Deep impact : Unintended consequences of journal rank. *Frontiers in Human Neuroscience, 7*, 1–12. doi:10.3389/fnhum.2013.00291

Oxford University Press. (2018 a). *Instructions to authors*. Retrieved from https://academic.oup.com/brain/pages/General_Instructions

Oxford University Press. (2018 b) *Instructions for authors*. Retrieved from https://academic.oup.com/cercor/pages/Instructions_For_Authors

Retraction Watch. (2015). *Diederik Stapel now has 58 retractions*. Retrieved from https://retractionwatch.com/2015/12/08/diederik-stapel-now-has-58-retractions/#more-34952

Plos. (2018). *Competing Interests*. Retrieved from http://journals.plos.org/plosbiology/s/competing-interests

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331–363. doi:10.1037/1089-2680.7.4.331

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., … Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, *7*(10), 1-7. doi:10.1186/1471-2288-7-10

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., … Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *British Medical Journal*, *358,* j4008*.* doi: 10.1136/bmj.j4008

Simon, S. D. (2001). Understanding the odds ratio and the relative risk. *Journal of andrology*, *22*(4), 533-536. doi:https://doi.org/10.1002/j.1939-4640.2001.tb02212.x

Siontis, K. C., Hernandez-Boussard, T., & Ioannidis, J. P. A. (2013). Overlapping meta-analyses on the same topic: survey of published studies. *British Medical Journal, 347*, f4501. doi:https://doi.org/10.1136/bmj.f4501

Society for Science & the Public. (2005). *Rating Researchers*. Retrieved from https://www.sciencenews.org/article/rating-researchers

Suttle, C. M., Lawrenson, J. G., & Conway, M. L. (2018). Efficacy of coloured overlays and lenses for treating reading difficulty: An overview of systematic reviews. *Clinical and Experimental Optometry*, 514–520. doi:10.1111/cxo.12676

Taylor and Francis Online. (2018). *Instructions for authors*. Retrieved from https://www.tandfonline.com/action/authorSubmission?show=instructions&journalCode=hpli20&

The Royal Society. (2018). *Publishing metrics*. Retrieved from http://rsob.royalsocietypublishing.org/citation-metrics

Tijdink, J. K., Verbeke, R., & Smulders, Y. M. (2014). Publication pressure and scientific misconduct in medical scientists. *Journal of Empirical Research on Human Research Ethics*, *9*(5), 64–71. doi:10.1177/1556264614552421

UNESCO. (2015). *Rapport de l'UNESCO sur la science, vers 2030*. Retrieved from https://fr.unesco.org/node/252295

University College Dublin. (2017). *Bibliometrics: Journal impact factor*. Retrieved from http://libguides.ucd.ie/bibliometrics/JIF

Useem, J., Brennan, A., Lavalley, M., Vickery, M., Ameli, O., Reinen, N., & Gill, C. J. (2015). Systematic differences between Cochrane and non-Cochrane meta-Analyses on the same topic : A matched pair analysis. *PLOS One, 111*, 1–17. doi:10.1371/journal.pone.0144980

Wasiak, J., Tyack, Z., Ware, R., Goodwin, N., & Jr, C. M. F. (2016). Poor methodological quality and reporting standards of systematic reviews in burn care management. *International Wound Journal, 14*(5), 754–763. doi: 10.1111/iwj.12692

Wu, X. Y., Lam, V. C. K., Yu, Y. F., Ho, R. S. T., Feng, Y., & Chung, V. C. H. (2016). Epidemiological characteristics and methodological quality of meta-analyses on diabetes mellitus treatment : A systematic review. *European Journal of Endocrinology, 175*, 353–360. doi:10.1530/EJE-16-0172

Yong, E. (2012). Bad copy. *Nature*, *485*(7398), 298-300. doi:10.1038/485298a

Young, N. S., Ioannidis, J. P. A., & Al-ubaydli, O. (2008). Why current publication practices may distort science. *PLOS Medecine 5*(10), e201. doi:10.1371/journal. pmed.0050201

Zhang, H., Han, J., Zhu, Y., Lau, W., Schwartz, M. E., Xie, G., … Yang, T. (2016). Reporting and methodological qualities of published surgical meta-analyses. *Journal of Clinical Epidemiology*, *70*, 4–16. doi:http://dx.doi.org/10.1016/j.jclinepi.2015.06.009

Zhu, Y., Fan, L., Zhang, H., Wang, M., Mei, X., Hou, J., … Shen, Y. (2016). Is the best evidence good enough: Quality assessment and factor analysis of meta-analyses on depression. *PLoS ONE*, *11*(6), e0157808. doi:10.1371/journal.pone.0157808