Attitude face à la recherche scientifique et lecture critique : une enquête auprès de chercheurs

Auteur : Sensi, Calista
Promoteur(s) : Willems, Sylvie
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
Année académique : 2024-2025
URI/URL : http://hdl.handle.net/2268.2/24529

LIÈGE université
Psychologie, Logopédie
& Sciences de l'Éducation

Attitude face à la recherche scientifique et lecture critique

Une enquête auprès de chercheurs

Mémoire présenté par Sensi Calista

Sous la promotion de Willems Sylvie
Sous la supervision de Blause Sacha

Lecteurs : Rousselle Laurence et Stawarczyk David

En vue de l'obtention du diplôme de Master en Sciences Psychologiques à finalité spécialisée en Psychologie Clinique, filières Neuropsychologie Clinique de l'Enfant et de l'Adulte.

Année académique 2024-2025

Remerciements

L'aboutissement de ce mémoire n'aurait jamais été possible sans le soutien, l'écoute et la bienveillance de nombreuses personnes à qui j'adresse ces quelques mots.

Je tiens tout d'abord à exprimer ma profonde gratitude à Madame Sylvie Willems, ma promotrice. Vous avez été une source d'inspiration précieuse tout au long de mon parcours de formation à la neuropsychologie. Merci pour votre encadrement attentif, vos conseils éclairés et votre bienveillance, qui m'ont permis de progresser et de me dépasser. Je remercie également Madame Sacha Blause pour son accompagnement tout au long de ce travail. Merci pour votre disponibilité et vos retours toujours justes et constructifs. Je suis honorée que ce mémoire s'inscrive dans le cadre de votre thèse et d'avoir pu contribuer, à mon échelle, à votre travail. Enfin, mes remerciements vont à Madame Laurence Rousselle et Monsieur David Stawarczyk pour l'attention portée à ce travail et le temps que vous lui consacrerez. J'espère que la lecture de celui-ci vous sera à la fois agréable et enrichissante.

Je souhaite également remercier les associations professionnelles qui ont permis la diffusion de notre enquête : *Whats'Up Neuropsychologie Clinique*, *PSYNCog*, l'*Association Suisse de Neuropsychologie* (ASNP), la *Société de Neuropsychologie de Langue Française* (SNLF) ainsi que l'*Association Québécoise des Neuropsychologues* (AQNP). Merci pour votre soutien et votre collaboration.

À mes amis les plus proches, merci pour votre présence, vos encouragements et ces moments de répit partagés qui m'ont permis de garder le cap. À vous, mes parents, merci pour votre soutien inconditionnel et d'avoir toujours cru en moi, même dans les moments de doute. À toi, Martin, merci du fond du cœur pour ton écoute, ton soutien et ta patience. Tu as été là à chaque étape de mon parcours universitaire et t'avoir à mes côtés m'a procuré une force inestimable.

Et enfin, merci à toi, Anita, mon binôme de mémoire et bien plus encore. Sans toi, ces cinq années d'études n'auraient pas eu la même saveur, la même intensité, la même lumière. Tu as été présente à chaque étape, tant dans les moments de doute que de joie. Grâce à toi, ce chemin a été bien plus qu'un simple parcours académique. Il a été une aventure humaine profonde, belle et inoubliable. Je n'ai aucun doute : rien de tout cela n'aurait été possible sans toi.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

L'adoption d'une pratique fondée sur les preuves, connue sous le nom *d'Evidence-Based Practice* (EBP), représente une avancée majeure dans plusieurs secteurs de la santé, y compris en neuropsychologie clinique (Blause et al., 2023a). Cette approche repose sur quatre piliers : l'expertise clinique, les meilleures données issues de la recherche, les préférences et les valeurs du patient ainsi que le contexte organisationnel et environnemental. Chacun de ces piliers offre une perspective unique et indispensable, sans qu'aucun ne prédomine sur les autres, et fournit une base solide aux praticiens désireux de s'appuyer sur les avancées récentes en matière de théorie, de méthodologie et de technique (Babione, 2010 ; Melchert et al., 2023 ; Spring, 2007).

Toutefois, l'intégration de l'EBP dans la pratique clinique présente plusieurs défis. Plusieurs études ont mis en évidence les freins rencontrés par les professionnels de la santé, en particulier concernant l'adoption du pilier *recherche*. Ce dernier retient particulièrement l'attention en raison des défis liés à l'utilisation des données issues des articles scientifiques et à la capacité à identifier les biais méthodologiques. Cette problématique est d'autant plus préoccupante que des études récentes mettent en lumière des lacunes dans la rigueur méthodologique des travaux publiés en neuropsychologie (Blause et al., 2023b ; Blause et al., 2025).

Actuellement, aucune étude n'a véritablement exploré la manière dont les neuropsychologues cliniciens et les chercheurs en neuropsychologie lisaient et analysaient concrètement un article scientifique. En effet, les pratiques de lecture critique ont principalement été examinées chez les cliniciens par le biais de questionnaires auto-rapportés, ne permettant pas d'observer directement leurs comportements. Notre étude vise donc à aller un pas plus loin en combinant un questionnaire avec une tâche concrète de lecture d'articles scientifiques.

Par ailleurs, cette étude s'inscrit dans le cadre d'un mémoire réalisé en binôme partiel. Le présent travail se concentre sur la manière dont les chercheurs lisent et évaluent la qualité méthodologique d'un essai contrôlé randomisé – considéré comme le *« gold standard »* en matière de recherche (Hariton & Locascio, 2018). Le travail réalisé en parallèle par Ceman Anita porte sur ces mêmes dimensions auprès des neuropsychologues cliniciens. Cette complémentarité nous permettra de croiser les regards de ces deux acteurs centraux et d'apporter une vision plus globale de la situation. Nous espérons ainsi être en mesure de proposer des pistes pour diminuer le fossé entre la recherche scientifique et la pratique clinique.

Pour ce faire, nous proposerons tout d'abord une revue de la littérature introduisant *l'Evidence-Based Practice* et son application en neuropsychologie clinique. Nous explorerons ensuite les piliers de l'EBP en mettant l'accent sur le pilier *recherche* et sur les étapes essentielles de son utilisation en pratique clinique. Nous discuterons ensuite des défis liés à la mise en œuvre de l'EBP avant de nous concentrer plus spécifiquement sur le pilier *recherche*. Nous nous intéresserons notamment aux habitudes des neuropsychologues cliniciens, ainsi qu'aux difficultés qu'ils rencontrent, en termes de lecture critique d'articles scientifiques. Ensuite, nous aborderons les questions de qualité et de conception des études et, plus particulièrement, des essais contrôlés randomisés. Cette revue permettra ainsi de mieux appréhender les enjeux actuels de l'intégration des preuves scientifiques dans la pratique clinique. À la suite de cette introduction théorique, nous présenterons les objectifs de cette étude ainsi que nos hypothèses. Nous détaillerons par après la méthodologie employée et présenterons les résultats obtenus. Enfin, nous discuterons nos résultats au regard de nos hypothèses et des données de la littérature, en soulignant les principaux apports et limites de notre étude, ainsi qu'en formulant des perspectives concrètes pour l'avenir.

Introduction théorique

1. L'*Evidence-Based Practice*

De nos jours, la formation du neuropsychologue clinicien vise à ce qu'il adopte une approche intégrée, s'appuyant sur les principes de la pratique fondée sur les preuves ou *Evidence-Based Practice* (EBP). En effet, l'utilisation d'une pratique fondée sur les preuves est de plus en plus recommandée dans le domaine de la neuropsychologie clinique (Blause et al., 2023a). L'EBP est une pratique utilisée dans de nombreux domaines de la santé. Initialement, elle est née en médecine dans les années 1980, sous le nom d'*Evidence-Based Medecine* (EBM), dans le but de promouvoir l'utilisation plus systématique des preuves scientifiques dans la formation des médecins et la pratique clinique (Rousseau & Gunia, 2016).

Cette approche a pour objectifs de réduire le recours à des interventions de soins de santé inefficaces et de promouvoir une prise de décision clinique basée sur une utilisation judicieuse des meilleures données probantes actuelles (Pagoto et al., 2007). Dans le domaine de la psychologie, c'est en 2005 que l'*American Psychological Association* (APA) adopte cette pratique fondée sur les preuves en psychologie, ou *Evidence-Based Practice in Psychology* (EBPP). La définition de L'EBPP selon l'APA (APA Presidential Task Force on Evidence-Based Practice, 2006) est la suivante :

> *« L'EBPP est l'intégration des meilleures données disponibles issues de la recherche scientifique à l'expertise clinique, dans le contexte des caractéristiques, de la culture et des préférences du patient »*.

Cette définition peut être représentée à l'aide de la métaphore d'un tabouret à trois pieds où les pieds représentent l'expertise clinique, les caractéristiques, la culture et les préférences du patient ainsi que la recherche scientifique. Cette métaphore met en évidence l'idée que si l'un des pieds n'est pas présent, le tabouret ne peut tenir debout. Ainsi, ces trois pieds, appelés également *« piliers »* de l'EBP, sont essentiels à la pratique des soins de santé (Melchert et al., 2023 ; Spring, 2007) et doivent être considérés de manière complémentaire, imbriquée et non hiérarchique (Babione, 2010 ; Blause et al., 2023a).

Certains auteurs mettent en avant une quatrième composante à l'EBP. C'est le cas de Satterfield et ses collègues (2009) qui ont élaboré un modèle révisé de l'EBP (voir Figure 1). Dans ce modèle, le contexte environnemental et organisationnel est considéré comme englobant les trois composantes précédemment évoquées. Par ailleurs, ce modèle adopte une perspective transdisciplinaire et met l'accent sur l'importance de la prise de décision partagée, se trouvant au centre du modèle.

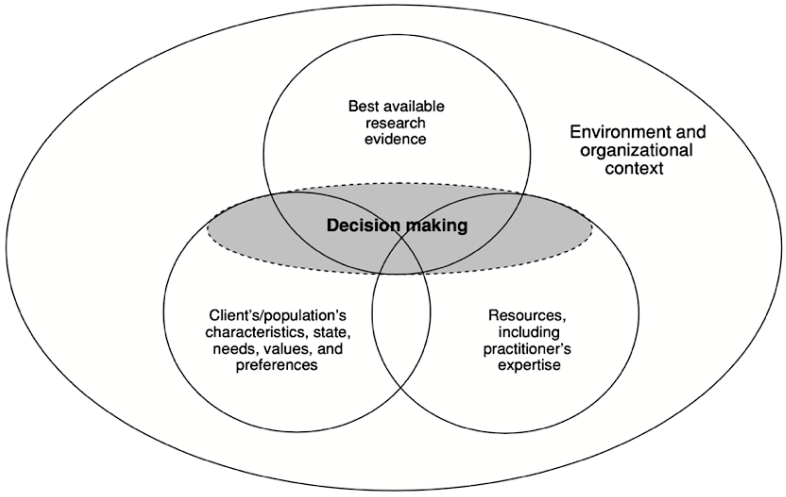

Figure 1. Le modèle révisé de l'EBP (Satterfield et al., 2009)

Le psychologue clinicien doit ainsi tenir compte de quatre piliers essentiels qui représentent chacun une base d'informations pertinentes à exploiter et à combiner pour guider une décision clinique. Parmi ces quatre piliers, on retrouve (1) le pilier *recherche*, qui implique l'intégration des meilleures données issues de la recherche scientifique dans les décisions cliniques, (2) le pilier *expertise*, qui fait référence à l'expérience et aux compétences du praticien, (3) le pilier *patient*, qui implique de prendre en considération les valeurs, les préférences et les attentes du patient dans le processus de soins et (4) le pilier *contexte*, qui met en avant l'importance de prendre en compte le contexte clinique, incluant les ressources disponibles et les contraintes, dans les décisions de soins de santé (Babione, 2010 ; Melchert et al., 2023).

Concernant le pilier *recherche* en particulier, le clinicien, adoptant une approche fondée sur des preuves, devrait notamment être capable de remettre en question ses pratiques et de reconnaître le besoin d'informations lorsqu'il se trouve dans une situation d'incertitude concernant un diagnostic, une action préventive ou le choix d'un traitement pour un patient. En effet, cette approche répond aux besoins des cliniciens soucieux de prendre en considération les avancées récentes en matière de théorie, de méthodologie et de techniques. Elle guide les interventions

dans le but d'améliorer la qualité des soins aux patients et de diminuer l'incertitude du clinicien quant à ses choix cliniques. Ainsi, lorsque ce dernier se posera une question clinique structurée et précise, il recherchera les meilleures données issues de la recherche, évaluera celles-ci de manière critique, appliquera les résultats dans sa pratique et évaluera sa performance (Durieux et al., 2017).

Par ailleurs, certains auteurs parlent d'*Evidence-Based Clinical Neuropsychological Practice* (EBCNP) dont la définition n'est pas encore établie. Toutefois, cette définition devrait probablement inclure les mêmes caractéristiques fondamentales de l'EBM ou de l'EBPP, mais également un cinquième pilier : les attentes et besoins spécifiques des différents référents, c'est-à-dire des personnes susceptibles de solliciter un avis neuropsychologique sur une question précise. Ainsi, dans le domaine de la neuropsychologie clinique, la démarche et les piliers de l'EBP restent fondamentalement identiques, mais sont adaptés pour répondre aux exigences spécifiques des patients, des praticiens de santé et des contextes cliniques propres à la neuropsychologie (Atzeni & Follenfant, 2013 ; Chelune, 2010).

Selon Atzeni et Follenfant (2013), l'EBCNP suit une démarche en cinq étapes. Tout d'abord, il s'agit de formuler une question clinique claire à partir des problèmes et plaintes exprimés par le patient, son entourage et/ou l'équipe médicale. Ensuite, la deuxième étape consiste principalement à collecter et à intégrer deux types de données : celles relatives à l'histoire du patient et celles tirées des données scientifiques existantes en lien avec la question clinique posée. La troisième étape est de choisir un protocole pertinent pour traiter la question clinique en s'appuyant sur l'intégration de ces données. La quatrième étape implique d'analyser statistiquement les données et de les interpréter dans le contexte particulier de l'individu concerné. Enfin, la cinquième étape consiste à évaluer l'efficacité du traitement en vérifiant s'il permet de répondre à la question initiale, s'il améliore la qualité de vie du patient et s'il présente un intérêt pour de futures situations analogues.

Cette démarche présente de nombreuses similarités avec la méthode scientifique utilisée en recherche. Toutefois, l'EBP ne doit pas être réduite à cette approche structurée, mais doit plutôt être perçue comme un rappel de l'importance de mettre en pratique certaines composantes fondamentales du pilier *recherche*. À chaque étape de la démarche clinique, le clinicien peut ainsi s'appuyer sur ce pilier qui englobe l'identification, la sélection et l'analyse d'articles basés sur des preuves de haute qualité (Blause et al., 2025). Dans cette perspective, il est pertinent de s'intéresser plus en détail au pilier *recherche*, qui, parmi les quatre piliers de l'EBP, suscite une

attention particulière en raison des défis associés à son intégration dans la prise de décision clinique (Blause et al., 2023a).

2. L'intégration du pilier *recherche* dans la pratique clinique

Des chercheurs ayant étudié l'EBP s'accordent à dire que, pour être pertinente et rentable, cette intégration requiert que les cliniciens s'engagent dans une démarche réflexive de recherche d'informations qui peut se diviser en cinq étapes (Blause et al., 2023a ; Chelune, 2010 ; Straus et al., 2011).

La *première étape* consiste à transformer un besoin d'informations en une question à laquelle il est possible de répondre (Blause et al., 2023a ; Thyer & Pignotti, 2011). Pour cela, le praticien doit structurer sa réflexion, car la manière dont une question est formulée va impacter le processus de recherche. C'est pourquoi la formulation des questions EBP implique souvent l'utilisation de modèles qui structurent la question afin de faciliter la recherche (Rousseau & Gunia, 2016). Dans le domaine des soins de santé, de nombreuses questions suivent la structure la plus courante, reprise sous l'acronyme PICO – Population, Intervention, Comparaison, Résultats ou *Outcomes* (Durieux et al., 2017 ; Falzon et al., 2010).

La *seconde étape* consiste à recueillir des données pertinentes, en déterminant notamment le *design* d'étude le plus approprié pour répondre à la question posée (Blause et al., 2023a ; Thyer & Pignotti, 2011). Pour ce faire, le praticien peut se tourner vers la pyramide des niveaux de preuves (Agoritsas et al., 2015 ; OCEBM, 2011). Dans cette pyramide des niveaux de preuves, on dispose tout d'abord de données issues d'études primaires ; les Essais Contrôlés Randomisés (ECR), les études de cohorte et les études de cas sont des exemples de *designs* d'études dans le domaine de la santé. Ces données sont les plus abondantes. Ensuite, il existe des synthèses méthodiques de la littérature scientifique, telles que les revues systématiques ou encore les méta-analyses, fondées sur une recherche systématique d'études primaires portant sur une même problématique. Enfin, on dispose des recommandations de bonne pratique clinique, ou *clinical guidelines*, décrivant des recommandations destinées à optimiser les soins apportés à un patient. Elles s'appuient sur une synthèse méthodique des données issues de la recherche scientifique et mentionnent les avantages et inconvénients des options alternatives de soins. Toutefois, les études de synthèse étant plus rares, il est courant de devoir se tourner vers les études primaires. Ainsi, selon plusieurs auteurs, les ECR devraient constituer une priorité lorsqu'il s'agit d'évaluer l'efficacité et l'efficience des traitements, dans la mesure où ils

représentent la forme d'étude primaire la moins sujette aux biais et aux erreurs (Durieux et al., 2017 ; Spring, 2007). Les ECR permettent d'examiner les effets d'une variable indépendante sur une variable dépendante tout en minimisant les influences des variables environnementales. Ils fournissent alors au praticien des données précieuses sur l'efficacité d'une intervention pour un type de patient spécifique, en la comparant à une autre intervention ou à l'absence d'intervention. Par ailleurs, ils facilitent la généralisation des résultats et permettent de déterminer la taille d'effet de l'intervention. Cette dernière permet de rendre compte de l'ampleur de l'impact d'une variable indépendante sur une variable dépendante (Téllez et al., 2015). Elle constitue ainsi une information cruciale pour que le praticien puisse choisir une intervention adéquate (Blause et al., 2025).

La *troisième étape* implique une analyse critique des données collectées pour en déterminer la validité, l'impact (taille d'effet) et l'utilité pour la clinique (Blause et al., 2023a ; Thyer & Pignotti, 2011). Cette étape requiert du praticien des compétences analytiques et une compréhension des biais méthodologiques potentiels, comme le manque d'informations, l'absence de mise en aveugle des participants et des observateurs, les résultats mal définis, la taille de l'échantillon mal calculée ainsi que l'interprétation inexacte des résultats (Blause et al., 2023a ; Blause et al., 2025). Des grilles de lecture ont dès lors été élaborées pour guider le praticien dans sa lecture. Par exemple, la grille RoB2 a pour objectif d'estimer le risque de biais dans les résultats de tout type d'ECR (Sterne et al., 2019). Nous aborderons cette grille plus en détails ultérieurement.

La *quatrième étape* implique que le praticien prenne une décision éclairée par les preuves issues de la recherche documentaire et intègre son analyse critique des données collectées à son expertise clinique ainsi qu'aux valeurs et à la situation unique de son patient (Blause et al., 2023a ; Thyer & Pignotti, 2011). Cette décision doit tenir compte des préférences du patient dans une perspective de compréhension et de prise de décision partagée concernant le traitement à adopter, qui doit être en adéquation avec le contexte dans lequel se déroule l'intervention. En outre, l'expertise même du praticien constitue une autre source d'information cruciale, permettant un juste équilibre entre les différents éléments de preuves (Blause et al., 2023a).

La *cinquième étape* consiste pour le clinicien à évaluer ses propres compétences dans l'application des quatre premières étapes et à explorer des pistes d'amélioration pour la prochaine fois (Blause et al., 2023a ; Thyer & Pignotti, 2011). Il doit alors s'interroger sur la pertinence et la clarté de la question formulée ainsi que sur la qualité des sources consultées

(Durieux et al., 2017). Cette étape consiste également à évaluer la pratique, c'est-à-dire si le traitement choisi a été efficace et si les effets observés sont attribuables à ce traitement. Cette évaluation nécessite ainsi l'expertise du praticien (Blause et al., 2023a).

Bien que cette approche en cinq étapes vise à améliorer les stratégies de mise en œuvre d'une pratique basée sur les preuves, des défis persistent quant à l'intégration du pilier *recherche* dans la pratique clinique. Nous allons examiner ces défis en adoptant d'abord une perspective globale sur l'EBP, puis en nous concentrant plus spécifiquement sur le pilier *recherche*.

3. Les défis liés à l'application de l'EBP

Malgré les nombreuses études soulignant son potentiel bénéfique, l'application de l'EBP en milieu clinique semble être relativement négligée (Babione, 2010). Ce constat est particulièrement marqué dans le domaine de la neuropsychologie clinique, où cette approche demeure mal comprise et sous-utilisée (Chelune, 2010). En effet, plusieurs études montrent que l'approche EBP est souvent mal interprétée, probablement en raison d'un manque de connaissances sur le sujet.

Une première étude, menée en Belgique auprès de 110 psychologues francophones (comprenant des neuropsychologues), révèle que seulement un tiers des participants connait le concept d'EBP, qu'environ un tiers des répondants n'a jamais entendu parler de cette pratique, tandis qu'un autre tiers connait le terme sans pouvoir en définir l'approche. Quant à l'application de l'EBP dans la pratique clinique, cette même étude montre que les comportements des psychologues ne sont pas entièrement conformes aux recommandations de l'EBP. Les auteurs soulignent ainsi la nécessité de former les cliniciens afin d'améliorer la compréhension de cette approche (Blause et al., 2024).

En ce qui concerne les neuropsychologues plus spécifiquement, une autre étude de Blause et ses collègues (2023a) révèle un manque de connaissance de l'EBP parmi un échantillon de neuropsychologues francophones. En effet, sur les 392 participants, seulement 35 % donnent une définition correcte de l'EBP, soulignant une fois de plus la confusion autour de cette approche. Ce constat n'est d'ailleurs pas spécifique à la neuropsychologie, des résultats similaires étant observés dans d'autres disciplines cliniques (Aarons, 2004 ; Emwodew et al., 2021 ; Li et al., 2024 ; Nakamura et al., 2011 ; Pagoto et al., 2007).

Selon Pagoto et ses collègues (2007), le fait de comprendre les obstacles et les facilitateurs pertinents de l'EBP pourrait faciliter son approbation et sa mise en œuvre. Les auteurs expliquent que des obstacles peuvent surgir à chaque étape de la démarche EBP, freinant les progrès vers une mise en œuvre généralisée. Cependant, des éléments facilitateurs peuvent également intervenir à tout moment pour améliorer le processus de mise en œuvre. Dans leur étude qualitative, ils observent que les praticiens évoquent davantage d'obstacles (64 %) que de facilitateurs (36 %) pour la mise en œuvre de l'EBP, les obstacles principaux étant associés aux attitudes négatives à l'égard des traitements validés empiriquement, ou *« Empirically Supported Treatment »* (EST). Ces attitudes semblent refléter une conception erronée de l'EBP, la réduisant à l'utilisation exclusive d'EST. Luebbe et ses collègues (2007) soulignent quant à eux le fait que de nombreux professionnels de la santé, y compris des étudiants, utilisent parfois les termes EBP et EST de manière interchangeable. Pourtant, il est essentiel de ne pas assimiler l'EBP aux seuls EST (Spring, 2007 ; Thyer & Pignotti, 2011). Alors que les EST reposent principalement sur des ECR, l'EBP constitue un processus plus large, visant à collecter et évaluer les meilleures preuves scientifiques disponibles, qu'il s'agisse d'ECR ou d'autres sources fiables telles que les revues systématiques (Luebbe et al., 2007).

En somme, bien que l'importance de l'EBP en milieu clinique soit largement reconnue, son application pratique semble être freinée par divers obstacles. Parmi ceux-ci figurent notamment un manque de connaissances et, dans certains cas, des attitudes négatives à son égard, lesquelles résulteraient en grande partie d'une formation insuffisante (Blause et al., 2023a ; Pagoto et al., 2007). D'autres obstacles, qui relèvent plus spécifiquement du pilier *recherche* de l'EBP, sont développés dans la section suivante.

4. Les défis spécifiques au pilier *recherche*

La mise en œuvre efficace de l'EBP repose à la fois sur les compétences de lecture critique des cliniciens, que nous regrouperons sous le terme de *« méta-pratique »* et sur la production de connaissances fiables et rigoureuses par les chercheurs, que nous désignerons sous le terme de *« méta-recherche »*. Ainsi, pour appréhender pleinement les défis propres au pilier *recherche*, il est nécessaire d'examiner conjointement ces deux dimensions : la méta-pratique, en s'intéressant aux habitudes de lecture critique des cliniciens, et la méta-recherche, en analysant la qualité des publications scientifiques et les pratiques méthodologiques des chercheurs.

4.1. La méta-pratique

4.1.1. Les défis liés à la lecture d'articles scientifiques

Comme précédemment évoqué, il est difficile pour les cliniciens d'incorporer les données de la recherche dans leur pratique, et ce, pour plusieurs raisons. Dans cette section, nous allons nous intéresser à une liste non exhaustive de défis en lien avec la lecture d'articles scientifiques auxquels sont confrontés les neuropsychologues cliniciens, notamment le manque de temps, le manque d'accès à l'information scientifique ainsi que le manque de compétences.

4.1.1.1. Le manque de temps

Selon Durieux et ses collègues (2017), les cliniciens sont souvent confrontés à une charge de travail importante et manqueraient donc de temps pour se consacrer à la lecture d'articles scientifiques et à l'actualisation de leurs connaissances. Cette difficulté est particulièrement marquée en milieu hospitalier où le rythme est soutenu. Plusieurs études montrent en effet que le manque de temps est l'un des principaux obstacles à l'intégration de la recherche dans la pratique clinique. Nelson et ses collègues (2006) soulignent notamment que de nombreux cliniciens, travaillant déjà au-delà de 40 heures par semaine, estiment ne pas avoir le temps nécessaire pour se tenir à jour sur la littérature de la recherche clinique. Ce constat est d'ailleurs confirmé par une enquête menée auprès de 352 neuropsychologues cliniciens, qui montre que le manque de temps est l'obstacle le plus fréquemment mentionné, par 59 % des participants (Blause et al., 2023a). Cette tendance est également observée chez d'autres professionnels de la santé mentale. Par exemple, dans une autre étude menée auprès de psychologues cliniciens francophones (n = 110), 59 % des répondants citent le manque de temps comme obstacle principal à l'intégration de la recherche dans leur pratique (Blause et al., 2024). Une autre enquête, réalisée auprès de logopèdes (n = 410), confirme également cette tendance, avec 54 % des répondants évoquant le manque de temps (Durieux et al., 2012). Dans une étude de Stewart et ses collègues (2012), les cliniciens interrogés soulignent d'ailleurs que les articles scientifiques sont longs à lire et que le gain d'information ne justifie pas toujours l'effort.

4.1.1.2. Le manque d'accès à l'information scientifique

Au-delà du manque de temps, l'accès limité à des ressources documentaires constitue un obstacle supplémentaire pour les praticiens souhaitant rester à jour en termes de méthodologies empiriquement fondées. Les praticiens ne peuvent en effet pas se satisfaire des enseignements

reçus lors de leur formation initiale et sont dès lors amenés à actualiser régulièrement leurs connaissances et compétences pour intégrer les avancées scientifiques dans leur pratique professionnelle. Or, l'accès à l'information scientifique est souvent entravé par des obstacles logistiques et financiers (Durieux et al., 2017). Par exemple, les cliniciens peuvent ne pas avoir accès à des bases de données scientifiques ou à des revues spécialisées en raison de coûts d'abonnement élevés (Blause et al., 2023a). Selon des enquêtes récentes, 51 % des neuropsychologues cliniciens (*n* = 352) et 43 % des psychologues cliniciens (*n* = 110) identifient le manque d'accès aux ressources comme un obstacle majeur à la lecture d'articles scientifiques (Blause et al., 2023a ; Blause et al., 2024).

4.1.1.3. Le manque de compétences

Dans l'étude menée par Blause et ses collaborateurs (2023a), les neuropsychologues cliniciens interrogés expriment un sentiment de manque de compétences en recherche d'informations. Ils rapportent notamment un manque de maîtrise, voire de connaissances, des outils de recherche ainsi que des difficultés à sélectionner et évaluer des articles pertinents. Baker et ses collègues (2008) expliquent que les cliniciens peuvent être dissuadés de s'engager dans la littérature scientifique pour plusieurs raisons comme le style d'écriture, l'utilisation d'un langage spécifique ou encore la complexité croissante des méthodes scientifiques et statistiques. En effet, cette évolution constante signifie que les cliniciens sont de plus en plus susceptibles de se retrouver face à des méthodes auxquelles ils n'ont pas été formés (Lilienfeld et al., 2014).

Notons que les barrières linguistiques peuvent également contribuer à ce manque de compétences, comme le signalent 26 % de neuropsychologues cliniciens (*n* = 352) (Blause et al., 2023a). Ce constat se retrouve également chez d'autres professionnels de la santé tels que les logopèdes (Durieux et al., 2012). Cette lacune est particulièrement préoccupante, car l'analyse critique des études scientifiques repose en grande partie sur une bonne compréhension des méthodologies de recherche et des analyses statistiques (Blause et al., 2023a ; Blause et al., 2024).

4.1.2. Les comportements des cliniciens en matière de recherche d'informations

Après les défis auxquels peuvent être confrontés les cliniciens lors de la lecture d'articles scientifiques, il convient maintenant de s'intéresser à leurs comportements en matière de recherche d'informations scientifiques. Plusieurs études suggèrent que, bien qu'ils reconnaissent l'importance du pilier *recherche* de l'EBP, les praticiens ont recours à la

littérature scientifique de manière limitée. Blause et ses collaborateurs (2023a) montrent qu'un nombre significatif de neuropsychologues cliniciens ne consultent que rarement des articles scientifiques, même lorsqu'ils sont confrontés à des situations d'incertitude clinique. Dans ce contexte, la majorité déclare préférer solliciter l'avis d'un collègue ou effectuer une recherche rapide via Google, tandis que seulement la moitié envisage la consultation d'articles scientifiques comme une option pertinente.

Cette tendance à privilégier le savoir de ses pairs ou l'expérience personnelle est largement documentée. En 2004, Aarons souligne déjà que les professionnels accordent davantage de crédit aux informations provenant de leurs pairs qu'aux publications de recherche ou aux ouvrages. De même, Gyani et ses collègues (2014) montrent que les thérapeutes fondent souvent leurs décisions cliniques sur leur propre expérience plutôt que sur les données empiriques disponibles. Une étude plus récente menée par Blause et ses collaborateurs (2024) auprès de psychologues francophones belges confirme cette prédominance de l'expertise clinique dans la prise de décision. En revanche, la recherche scientifique, tout comme les préférences et opinions du patient concernant le traitement ainsi que les avis d'experts, influenceraient peu leur pratique.

Plusieurs facteurs peuvent expliquer ce désintérêt relatif pour la littérature scientifique. D'une part, les cliniciens sont confrontés aux défis évoqués précédemment, et, d'autre part, certains expriment un scepticisme quant au transfert des résultats en argumentant que les patients inclus dans les études scientifiques sont différents de leurs propres patients en termes d'histoire, de difficultés, de ressources ou d'aspirations (Durieux et al., 2017). Ce scepticisme semble renforcé par un manque de formation à l'EBP qui limite leur capacité à interpréter et à intégrer les données probantes dans leur pratique quotidienne (Blause et al., 2023a).

Ainsi, il apparaît que les cliniciens consultent rarement la littérature scientifique pour guider le choix d'une intervention thérapeutique, préférant s'appuyer sur les techniques qui leur ont été enseignées ou qu'ils estiment acceptables et efficaces (Aarons, 2004). Or, il n'existe qu'une faible corrélation entre la confiance du thérapeute envers une intervention et son efficacité réelle (Miller et al., 2015). Ces pratiques peuvent conduire à l'utilisation de traitements moins efficaces, voire, dans certains cas, à une aggravation de l'état des patients à la suite d'une prise en charge psychologique (Lilienfeld et al., 2007).

Nous allons à présent nous pencher sur les comportements de lecture adoptés par les cliniciens qui s'engagent dans la consultation d'articles scientifiques.

4.1.3. Les comportements de lecture des cliniciens

En ce qui concerne les comportements de lecture des cliniciens, une étude montre que plus de la moitié d'entre eux ont des pratiques inefficaces, ne lisant pas les articles dans leur intégralité. Les parties les plus lues sont le résumé, la conclusion et la discussion, tandis que l'attention accordée à l'introduction, à la méthode et à la description des instruments de mesure est moindre. Par ailleurs, lors de la lecture des résultats d'un article, une grande importance est accordée à la *p*-valeur et moins à la taille d'effet, relevant d'un manque de compétences des méthodes statistiques (Blause et al., 2023a). En effet, cette *p*-valeur ne renseigne que sur la probabilité d'obtenir un tel résultat si l'hypothèse nulle est vraie, ce qui ne constitue pas une information suffisante d'un point de vue clinique (Faulkner et al., 2008).

Concernant les articles portant sur des outils d'évaluation, les praticiens montrent un intérêt élevé pour la lecture de la composition et de l'âge des normes ainsi que pour la validité théorique de l'étude. D'autres s'intéressent également à la sensibilité, à la spécificité ainsi qu'à la fidélité test-retest. En revanche, l'erreur standard de mesure, la fidélité inter-juges et la consistance interne sont moins prioritaires pour eux (Blause et al., 2023a).

4.1.4. Le regard critique des cliniciens

4.1.4.1. La conscience des biais

Dans une enquête auprès de neuropsychologues cliniciens (n = 350) concernant les biais dans la littérature, 94 % reconnaissent l'existence de biais dans la recherche et 70 % disent les prendre en compte lorsqu'ils lisent un article scientifique. Cependant, la moitié des participants indiquent qu'ils ne possèdent pas les compétences nécessaires pour effectuer correctement cette analyse. Cette enquête révèle ainsi que, bien qu'ils soient conscients de l'existence de biais dans les études scientifiques, les cliniciens ne se sentent pas toujours en mesure de les identifier ni de les interpréter correctement (Blause et al., 2023a). Ces difficultés ne sont pas propres aux neuropsychologues et rejoignent celles rencontrées par d'autres professionnels de santé. C'est notamment le cas des infirmiers qui rapportent eux aussi un manque de formation à l'évaluation critique des études scientifiques. Nombre d'entre eux se sentent insuffisamment préparés pour apprécier la qualité méthodologique des recherches (Mahmoud & Abdelrasol, 2019).

4.1.4.2. Les outils pour surmonter les biais

Pour surmonter les biais et faciliter l'évaluation critique des articles scientifiques par les praticiens, des grilles d'analyse ont été élaborées. La *checklist* JBI pour les ECR est l'une d'entre elles. Son objectif est d'évaluer la qualité méthodologique d'une étude et de déterminer dans quelle mesure elle aborde les biais possibles dans sa conception, sa mise en œuvre et son analyse (Barker et al., 2023 ; Tufanaru et al., 2020). Nous disposons également de l'outil RoB2, spécifiquement conçu pour évaluer le risque de biais dans les résultats de tout type d'ECR. Cet outil se structure en cinq domaines correspondant aux étapes méthodologiques d'un ECR, dans lesquels des biais sont susceptibles d'altérer les résultats de l'étude, à savoir : biais dans le processus de randomisation, biais dû à un changement dans l'intervention, biais dû à des données manquantes, biais dans la mesure des résultats et biais dans le report des résultats (Sterne et al., 2019).

Ces grilles peuvent également être employées par les chercheurs eux-mêmes pour la rédaction de leurs articles scientifiques, d'autant que certaines sont spécifiquement conçues à cet effet, comme nous le verrons par la suite. Notons que l'utilisation effective de ces grilles peut représenter un défi pour les cliniciens, car nécessitant un large éventail de compétences, entre autres statistiques et méthodologiques.

4.1.5. Conclusion

La méta-pratique met en évidence la difficulté des cliniciens à intégrer des preuves scientifiques dans leur pratique clinique. Si l'importance d'une pratique fondée sur les preuves est largement reconnue, sa mise en œuvre se heurte à de nombreux obstacles parmi lesquels figurent le manque de temps, l'accès restreint à l'information scientifique ainsi que des compétences limitées en recherche documentaire. Ces défis, particulièrement le manque de temps et de compétences, peuvent compromettre la capacité des praticiens à évaluer de manière adéquate la qualité des données probantes disponibles. Si les cliniciens ne s'engagent pas pleinement dans une analyse critique, il devient alors impératif que les données probantes sur lesquelles reposent leurs décisions soient de qualité irréprochable. En effet, une mauvaise évaluation de la qualité des études peut entraîner l'adoption de pratiques inappropriées, au détriment de l'efficacité des interventions et de la qualité des soins. C'est pourquoi il convient dès lors de porter une attention particulière à la rigueur méthodologique des études cliniques et à la transparence de leurs résultats.

4.2. La méta-recherche

4.2.1. L'importance d'une rigueur méthodologique

Afin que ses conclusions soient généralisables et significatives, il est primordial qu'une recherche soit planifiée et menée de façon rigoureuse. En effet, les études quantitatives qui ne respectent pas les normes de rigueur et de qualité requises par les revues scientifiques peuvent être sujettes à différentes critiques, notamment en raison d'échantillons non représentatifs, d'une collecte de données peu fiable, de résultats cliniquement peu pertinents ou encore de la présence de facteurs confondants susceptibles de compromettre la validité interne et externe de l'étude (Sink & Mvududu, 2010).

Selon Sink et Mvududu (2010), un des principaux facteurs de mauvaise performance en recherche est souvent attribué à un manque de compétences méthodologiques de la part des chercheurs. Bien que certaines erreurs puissent être évitées, elles restent fréquentes et ont parfois des conséquences graves et imprévues. Dans certains cas, ces erreurs peuvent amener les chercheurs à écarter une hypothèse correcte et à tirer des conclusions erronées. En outre, la pression de publier dans des revues prestigieuses les mène parfois à prioriser des résultats statistiquement significatifs au détriment d'une rigueur méthodologique, affectant ainsi la fiabilité des conclusions. Certains peuvent également être tentés de choisir leurs sujets de recherche et de concevoir leurs études en fonction de ce qui est susceptible d'être publié plutôt que de ce qui est scientifiquement le plus pertinent ou innovant (Morales et al., 2021). Cette tendance s'explique notamment par le biais de publication, qui réduit les chances de publication des études aboutissant à des résultats non statistiquement significatifs (Ioannidis, 2014 ; Scheel et al., 2021).

Parmi les biais courants dans la littérature psychologique, on trouve notamment les petites tailles d'effet et les petits échantillons (Leichsenring et al., 2016), l'importance excessive accordée à la *p*-valeur, le manque d'interprétation de la taille d'effet (Faulkner et al., 2008), le manque de transparence dans la rédaction des articles et l'utilisation de pratiques douteuses telles que le *« p-hacking »* qui consiste à recourir à différentes techniques pour augmenter la probabilité d'obtenir des résultats statistiquement significatifs (Scheel et al., 2021). Ce type de pratiques peut conduire à des résultats de recherche erronés, tels que des faux positifs ou une surestimation des effets (Blause et al., 2025 ; Hardwicke et al., 2021 ; Ioannidis, 2005).

Selon Simon (2001), toute recherche comporte des biais – certains sont mineurs et n'altèrent pas la validité des résultats tandis que d'autres, plus conséquents, peuvent compromettre leur utilité. L'enjeu central réside ainsi dans la capacité à évaluer l'impact de ces biais sur la solidité des conclusions d'une recherche. C'est pourquoi le recours aux ECR apparaît particulièrement pertinent, ceux-ci étant considérés comme la référence absolue pour l'évaluation des interventions en raison de leur capacité à minimiser, voire à éviter les biais (Moher et al., 2010).

4.2.2. Les essais contrôlés randomisés comme référence en matière de recherche

Comme précédemment évoqué, lorsque les cliniciens se posent une question, ils sont encouragés à rechercher en premier lieu des synthèses cliniques ou des guides de pratique clinique. Toutefois, les études de synthèse étant plus rares, il est courant qu'ils doivent se tourner vers les études primaires (Durieux et al., 2017). Les ECR sont largement reconnus comme étant le *« gold standard »* en matière de recherche, fournissant des informations précieuses sur l'efficacité potentielle d'une méthode d'intervention (Hariton & Locascio, 2018). En effet, lorsqu'ils sont correctement planifiés et rigoureusement conduits, les ECR demeurent la méthode de recherche la plus robuste pour identifier l'effet réel d'une intervention (Bhide et al., 2018). En revanche, pour être considérés comme fiables, les ECR doivent répondre à des critères de qualité. Peu importe si les résultats atteignent ou non une signification statistique, la conception, la réalisation et la publication d'un ECR doivent être de haute qualité (Moher et al., 1995). Si les critères de qualité ne sont pas respectés, les conclusions tirées de ces études deviennent peu fiables et donc inutilisables par les cliniciens, ce qui compromet la qualité d'une pratique clinique basée sur des données probantes (Blause et al., 2023a).

Bien que des guides existent pour orienter les chercheurs dans la planification, la conduite, l'analyse et le *reporting* des ECR (Begg et al., 1996 ; Bhide et al., 2018 ; Wolfenden et al., 2021), certains ECR peuvent présenter des biais méthodologiques. Ceux-ci vont compromettre la qualité de l'étude et, par conséquent, en limiter la pertinence et la transférabilité pour la pratique clinique. Par exemple, l'utilisation de méthodes inappropriées et une communication incomplète des résultats peuvent introduire des biais et entraîner une surestimation des effets des traitements (Moher et al., 2010). Des recherches montrent d'ailleurs que les ECR utilisant une allocation non randomisée ou décrivant de manière incomplète leur procédure de randomisation ont tendance à surestimer les effets des traitements par rapport aux études respectant ces standards méthodologiques. Ces lacunes méthodologiques peuvent nuire à la validité interne de l'étude, en réduire la transparence, compliquer son évaluation critique et

ainsi augmenter le risque de conclusions erronées et de décisions cliniques inappropriées (Blause et al., 2025 ; Moher et al., 2010 ; Munafò et al., 2017). Nous allons alors aborder le risque de biais ainsi que la qualité du *reporting* et de la conception des ECR.

4.2.2.1. Le risque de biais dans les essais contrôlés randomisés

Des études dans le domaine de la psychologie montrent que plusieurs biais ne sont pas encore suffisamment contrôlés et affectent encore la qualité des recherches publiées et leur reproductibilité (Leichsenring et al., 2016).

Dans le domaine de la neuropsychologie, Blause et ses collaborateurs (2023b) ont évalué la qualité d'un échantillon de 27 ECR examinant l'efficacité des interventions cognitives sur l'attention et les fonctions exécutives chez des enfants et adolescents atteints d'un Trouble du Déficit de l'Attention avec ou sans Hyperactivité (TDA/H). Leurs résultats indiquent qu'après évaluation avec la grille RoB2, 89 % des articles présentent un risque élevé de biais méthodologique, ce qui interroge la solidité des conclusions qui en sont tirées. Les lacunes les plus fréquentes concernent le processus de randomisation et la mesure des résultats. Plus récemment, une évaluation de la qualité d'un échantillon de 30 ECR portant sur la rééducation de la mémoire de travail chez les enfants montre que 26 ECR sont jugés à haut risque de biais selon la grille RoB2, 4 présentent des préoccupations modérées et aucun n'est considéré comme présentant un faible risque de biais (Blause et al., 2025). De manière générale, plusieurs revues dans le domaine de la neuropsychologie ou de la psychologie cognitive évaluent une majorité d'ECR comme étant à haut risque de biais, affaiblissant ainsi la portée de leurs résultats (Elbe et al., 2023 ; Miskowiak et al., 2016 ; Rivero et al., 2015).

Il apparaît donc essentiel d'évaluer systématiquement le risque de biais dans tout ECR à l'aide d'un outil spécifiquement conçu à cet effet, tel que RoB2 (Sterne et al., 2019). Ce dernier est largement utilisé par les auteurs de méta-analyses pour apprécier la qualité méthodologique des études qu'ils intègrent à leurs revues (Blause et al., 2025 ; Gates & March, 2016). Idéalement, il devrait également être employé par les cliniciens désireux de vérifier la solidité d'une intervention avant de l'adopter en pratique. Toutefois, comme nous l'avons précédemment souligné, l'utilisation de tels outils demeure complexe pour bon nombre d'entre eux en raison de contraintes de temps et de compétences méthodologiques. Du côté des chercheurs, Blause et ses collaborateurs (2025) insistent sur l'intérêt pour eux de se familiariser avec RoB2 en

amont de la réalisation d'une étude : cette démarche pourrait leur permettre d'anticiper certaines erreurs méthodologiques et d'améliorer la qualité de leur *reporting*.

4.2.2.2. La qualité du *reporting* et de la conception des essais contrôlés randomisés

Plusieurs auteurs considèrent que la qualité du *reporting* des ECR nécessite d'être améliorée. C'est notamment le cas de Faulkner et ses collègues (2008) qui ont étudié les pratiques statistiques dans 193 ECR portant sur des thérapies psychologiques et psychiatriques, parus entre 1999 et 2003 dans des revues de psychologie. Leurs résultats révèlent que seulement 46 % de ces ECR prennent en compte la puissance statistique, que 31 % interprètent l'ampleur de l'effet et seulement 2 % les intervalles de confiance. Ces résultats suggèrent donc que les ECR se concentrent davantage sur la significativité statistique, au détriment de l'interprétation clinique des effets observés, et mettent ainsi en évidence des lacunes pouvant compromettre la reproductibilité des études. Les auteurs soulignent également un écart entre l'évaluation des informations cruciales à inclure dans les ECR faite par les chercheurs et leur pratique de *reporting*, c'est-à-dire les informations qu'ils incluent effectivement dans leurs articles. Tandis que 86 % des participants considèrent comme importantes les informations statistiques relatives à l'existence d'un effet réel, à sa taille et sa précision, ainsi qu'à sa pertinence clinique, seuls 13 % des ECR fournissent effectivement ces données. Cette contradiction met en évidence des lacunes en matière de communication de la part des chercheurs.

En neuropsychologie, une étude de Blause et ses collaborateurs (2023b) met en évidence que les sous-dimensions les mieux rapportées dans plus de 50 % d'un échantillon de 27 ECR sont : le *background* scientifique, les données démographiques des sujets, les limitations de l'étude, la possible généralisation des résultats et les potentiels conflits d'intérêt. Cependant, les auteurs mettent en avant un manque de transparence général, caractérisé par un manque d'informations sur des éléments tels que le *blinding*, le calcul de la taille de l'échantillon ou encore les méthodes statistiques employées.

Notons que les défis liés au *reporting* et à la conception d'articles scientifiques ne se limitent pas au domaine de la psychologie. En effet, le *reporting* des ECR a toujours été insatisfaisant dans les revues médicales, qu'elles soient générales ou spécialisées (Schulz, 1996). Or, un *reporting* rigoureux permet non seulement aux cliniciens d'adapter les interventions à leur pratique, mais favorise également la réplication des études par d'autres chercheurs pour vérifier ou contester les résultats obtenus (Blause et al., 2025). Dès lors, il est crucial de réduire les biais

et de surmonter le manque de transparence dans le *reporting* des articles scientifiques (Faulkner et al., 2008 ; Munafò et al., 2017 ; Scheel et al., 2021). Pour ce faire, un groupe de scientifiques et d'éditeurs a développé la déclaration CONSORT que nous allons aborder dans le point suivant (CONsolidated Standards Of Reporting Trials ; Begg et al., 1996).

4.2.2.3. La déclaration CONSORT pour l'amélioration du *reporting* des ECR

Initialement publiée en 1996, puis mise à jour en 2001 et 2010, cette déclaration a été étendue au domaine des sciences comportementales et sociales en 2018. En effet, une extension de la déclaration CONSORT 2010, nommée CONSORT-SPI, a été développée pour aider les auteurs d'ECR d'interventions sociales et psychologiques (Montgomery et al., 2018). Peu importe sa version, cette déclaration a pour objectif principal de faciliter l'évaluation critique et l'interprétation des résultats des ECR. C'est pourquoi les revues scientifiques recommandent de plus en plus, voire imposent pour certaines, l'utilisation de cet outil (Shamseer et al., 2016). En exigeant une description détaillée et transparente des méthodes, la déclaration CONSORT permet aux lecteurs de distinguer les ECR rigoureux de ceux dont les résultats pourraient être biaisés ou peu fiables. Pour ce faire, elle se compose d'une *checklist* et d'un organigramme fournissant un cadre pour la rédaction des ECR (Grant et al., 2018 ; Moher et al., 2010).

La *checklist* de la déclaration CONSORT-SPI couvre des éléments comme le titre, le résumé, l'introduction, la méthodologie, la randomisation, les résultats, la discussion ainsi que d'autres informations importantes. Elle contient 45 *items* dont la communication insuffisante est associée à des estimations biaisées de l'effet du traitement, rendant ces éléments cruciaux pour évaluer la fiabilité et la pertinence des résultats. L'organigramme, quant à lui, illustre le parcours des participants tout au long de l'ECR et permet aux auteurs de fournir des informations sur les quatre phases de l'étude : inscription, allocation des interventions, suivi et analyse (Grant et al., 2018 ; Moher et al., 2010).

Selon Moher et ses collaborateurs (2010), suivre les recommandations de la déclaration CONSORT pourrait non seulement améliorer la qualité des ECR, mais aussi en faciliter la compréhension pour les lecteurs. Nous examinons cela plus en détail dans la section suivante.

4.2.2.4. La déclaration CONSORT et l'amélioration de la qualité du *reporting*

Après la publication de la déclaration CONSORT en 2001, une étude de Hopewell et ses collaborateurs (2010) révèle une amélioration dans la manière dont certains aspects méthodologiques sont présentés entre 2000 et 2006. Il y a une meilleure inclusion de détails concernant le critère de jugement principal, le calcul de la taille de l'échantillon ainsi que les méthodes de génération de séquence aléatoire et d'assignation secrète. Toutefois, la qualité des articles demeure inférieure aux normes souhaitables.

Une étude plus récente menée par Kilicoglu et ses collègues (2023) montre une amélioration de la qualité du *reporting* au fil du temps, la plus forte augmentation ayant eu lieu à la suite de la première publication de CONSORT en 1996. De nos jours, les auteurs indiquent une tendance à la hausse du nombre d'éléments rapportés, mais précisent que la plupart des critères CONSORT ne le sont pas toujours. Pour remédier à cela, nous avons vu précédemment que certaines revues scientifiques recommandent de plus en plus l'utilisation de la grille CONSORT (Shamseer et al., 2016). À ce propos, des auteurs montrent une amélioration de la qualité des revues qui demandent de soumettre une liste de contrôle complète (Jin et al., 2018).

Dans le domaine de la neuropsychologie, plusieurs méta-recherches mettent en évidence un écart entre les recommandations de la déclaration CONSORT et leur mise en application effective. Blause et ses collègues (2023b) révèlent que cette grille est souvent sous-utilisée ou mal comprise, avec, en moyenne, seulement 8 *items* correctement rapportés sur 45 dans un échantillon de 27 ECR. Une autre méta-recherche confirme cette tendance en montrant que seuls 10,8 des 45 *items* de CONSORT sont entièrement rapportés dans un échantillon de 30 ECR. Certains des éléments sous-rapportés sont pourtant essentiels à la compréhension et à l'application des résultats dans la pratique clinique (Blause et al., 2025).

Enfin, les auteurs de cette méta-recherche soulignent que si la checklist CONSORT est très complète, elle reste complexe à utiliser. La rédaction d'un ECR exige un travail conséquent et l'intégration d'outils d'aide au *reporting* ne devrait pas être reléguée aux dernières étapes de rédaction. Selon eux, une adoption plus précoce et systématique de ces outils pourrait ainsi améliorer la transparence et la qualité des publications (Blause et al., 2025).

4.2.3. Conclusion

En conclusion, bien que la qualité irréprochable des données probantes soit jugée essentielle pour guider les décisions des cliniciens, particulièrement lorsqu'ils manquent de temps et de compétences en lecture critique d'articles scientifiques, les ECR en neuropsychologie ne sont pas toujours exempts de biais et peuvent manquer de transparence. En effet, nous avons constaté que plusieurs ECR – pourtant considérés comme le *« gold standard »* pour évaluer l'efficacité des interventions – pouvaient présenter un risque élevé de biais selon la grille RoB2 et omettre plusieurs éléments jugés essentiels d'après la grille CONSORT (Blause et al., 2023b ; Blause et al., 2025). Par ailleurs, des auteurs ont mis en évidence un écart significatif entre les éléments que les chercheurs eux-mêmes considéraient comme essentiels à inclure dans un ECR et ceux qui étaient effectivement rapportés dans leurs publications (Faulkner et al., 2008).

Cette problématique de qualité méthodologique et de transparence dans les ECR n'est pas nouvelle et a d'ailleurs conduit au développement d'initiatives visant à y remédier, comme la déclaration CONSORT (CONsolidated Standards Of Reporting Trials ; Begg et al., 1996). Toutefois, des méta-recherches récentes indiquent que son application demeure encore insuffisante et que ses recommandations restent largement méconnues dans le domaine de la neuropsychologie (Blause et al., 2023b ; Blause et al., 2025). Par conséquent, de nombreux ECR dans ce domaine ne répondent pas aux critères méthodologiques requis pour fournir aux cliniciens des données réellement fondées sur des preuves fiables.

Cette situation conduit à s'interroger : si les ECR présentent encore aujourd'hui des faiblesses méthodologiques notables, ne pourrait-on pas y voir, en amont, une difficulté des chercheurs eux-mêmes à en évaluer la qualité ? Autrement dit, bien qu'il soit attendu des cliniciens qu'ils aient un regard critique face aux données issues des ECR, il apparaît tout aussi pertinent de se questionner sur les compétences des chercheurs en matière d'évaluation méthodologique – d'autant plus qu'à notre connaissance, aucune étude ne s'est encore spécifiquement penchée sur cette question.

Objectifs et hypothèses

1. Objectifs

Comme évoqué précédemment, la littérature scientifique en neuropsychologie peut présenter plusieurs limites sur le plan méthodologique (Blause et al., 2023b ; Blause et al., 2025). De plus, les travaux de Faulkner et ses collègues (2008) montrent un décalage important entre les éléments que les chercheurs considèrent comme essentiels à rapporter dans un ECR et ceux qu'ils mentionnent effectivement dans leurs publications. Ces constats soulèvent alors une question centrale : dans quelle mesure les chercheurs sont-ils aptes à évaluer la qualité méthodologique d'un article scientifique ?

Dans la continuité de ces travaux, l'objectif principal de ce mémoire est d'examiner l'attitude des chercheurs en neuropsychologie face à la recherche scientifique ainsi que leurs pratiques en matière de lecture critique. Plus précisément, nous explorons la fréquence à laquelle les chercheurs lisent des articles portant sur des interventions cliniques au quotidien. En pratique, nous regardons leurs comportements de lecture, notamment s'ils lisent les articles dans leur intégralité ou si certaines parties sont préférées, mais aussi s'ils s'aident d'outils d'évaluation critique lors de leur lecture. Par ailleurs, nous cherchons à déterminer si leur niveau de confiance dans les résultats d'un article varie en fonction de sa qualité méthodologique, quels critères ils mobilisent pour évaluer cette qualité et dans quelle mesure ils sont capables d'identifier d'éventuels biais méthodologiques. Enfin, ce mémoire étant réalisé en binôme partiel, notre objectif est de comparer les performances de lecture critique des chercheurs avec celles des neuropsychologues cliniciens.

Pour ce faire, notre échantillon de chercheurs a été invité à participer à une étude à distance consistant en la complétion d'un questionnaire ainsi qu'en la lecture de deux ECR se distinguant par leur qualité méthodologique et rédactionnelle, l'un étant considéré comme ayant un *reporting* de haute qualité (ci-après *« article reporting + »*) et l'autre un *reporting* de faible qualité (ci-après *« article reporting - »*). La qualité méthodologique des articles a été évaluée à l'aide de la grille CONSORT-SPI (Montgomery et al., 2018), afin d'apprécier l'exhaustivité et la transparence du *reporting*, ainsi que de la grille RoB2 (Sterne et al., 2019), destinée à estimer le risque de biais (voir infra).

2. Hypothèses

Compte tenu du caractère exploratoire de cette étude et du nombre limité de recherches antérieures sur le sujet, il est difficile de s'appuyer sur un cadre théorique solide pour formuler des hypothèses précises. Pour autant, nous présentons ci-après quelques hypothèses préliminaires susceptibles d'orienter notre investigation.

Tout d'abord, nous émettons l'hypothèse que les chercheurs liront les articles scientifiques dans leur intégralité. En l'absence de données spécifiques concernant les pratiques de lecture des chercheurs en neuropsychologie, cette hypothèse repose sur les résultats d'études menées auprès de cliniciens. Ces travaux indiquent que les cliniciens ont tendance à adopter une lecture partielle des articles, accordant moins d'attention à certaines sections telles que l'introduction, la méthodologie ou encore les éléments statistiques, comme la taille d'effet (Blause et al., 2023a). Ainsi, en raison de leur formation scientifique, nous supposons que les chercheurs, à l'inverse, mèneront une lecture intégrale et porteront une attention accrue à ces sections, particulièrement à la méthodologie et aux résultats.

Ensuite, nous formulons l'hypothèse que les chercheurs percevront l'article *reporting* + comme étant supérieur à l'article *reporting* – sur le plan de la qualité méthodologique ainsi que sur celui de la suffisance d'informations pour reproduire l'intervention, en apprécier l'efficacité et en juger l'utilité, mais aussi qu'ils accorderont une confiance plus élevée aux résultats de l'ECR et exprimeront une meilleure compréhension des différentes sections de l'article. Cette hypothèse s'aligne sur la littérature montrant que les recommandations de la déclaration CONSORT peuvent non seulement améliorer la qualité des ECR, mais aussi faciliter la compréhension pour les lecteurs, leur permettant notamment de distinguer les études méthodologiquement solides de celles dont les conclusions peuvent être biaisées ou peu fiables (Grant et al., 2018 ; Moher et al., 2010). Quant au jugement de confiance, nous formulons l'hypothèse que celui-ci pourrait changer après la complétion des différentes sections de questions qui font suite à la lecture de chaque article. Pour l'article *reporting* +, nous prévoyons une augmentation du niveau de confiance, nos questions permettant de mieux se rendre compte des éléments méthodologiques solides ou non. À l'inverse, pour l'article *reporting* -, nous nous attendons à une diminution du niveau de confiance. Nous émettons également l'hypothèse que les chercheurs identifieront un plus grand nombre de biais dans l'article *reporting* -, en raison de son risque de biais méthodologique plus élevé.

Par ailleurs, en lien avec l'étude de Faulkner et ses collaborateurs (2008), nous avançons l'hypothèse que les éléments statistiques tels que la *p*-valeur, la taille d'effet et l'intervalle de confiance seront des éléments pris en compte par les chercheurs pour évaluer la qualité méthodologique d'un ECR. Pour rappel, leur étude montre que ces éléments sont jugés importants par les chercheurs, bien qu'ils ne soient pas toujours rapportés dans leurs écrits. Ainsi, selon nous, les chercheurs restent conscients que ces éléments sont des critères de qualité d'un article scientifique.

Enfin, dans le contexte de ce mémoire en binôme, notre dernière hypothèse est que les chercheurs, en raison de leur formation et de leur expertise méthodologique plus approfondies, évalueront plus favorablement que les cliniciens l'article *reporting* +, tant sur le plan de la qualité méthodologique que sur celui de la suffisance d'informations pour reproduire l'intervention, en apprécier l'efficacité et en juger l'utilité. Ils devraient également rapporter une meilleure compréhension des différentes sections de l'article. Par ailleurs, il est attendu que les chercheurs identifient un plus grand nombre de biais méthodologiques que les cliniciens au vu du fait que ces derniers éprouvent, de par leur manque de compétences scientifiques et statistiques, des difficultés à évaluer des articles (Baker et al., 2008 ; Blause et al., 2023a).

Méthodologie

La section suivante détaille la méthodologie de notre étude, depuis le recrutement des participants jusqu'à la conception du questionnaire et la collecte des données. Nous y abordons également les analyses statistiques prévues en dernier lieu.

1. Participants

Initialement, les critères d'inclusion exigeaient que les participants soient titulaires du titre de psychologue spécialisé en neuropsychologie et exercent une activité de recherche en neuropsychologie, au moins à temps partiel, dans un pays francophone, au moment de l'étude ou au cours de la dernière année. Cependant, face au nombre insuffisant de participants, nous avons élargi nos critères en incluant également des chercheurs en psychologie cognitive et en neurosciences cognitives. Pour ce qui est de la taille d'échantillon, celle-ci n'a pas été déterminée à l'avance étant donné le caractère exploratoire de cette étude. Notre objectif a été d'obtenir le plus grand nombre possible de répondants, avec un minimum souhaité de vingt participants.

En ce qui concerne le recrutement, nous avons contacté par courriel des chercheurs exerçant au sein d'universités francophones, que ce soit en Belgique, en France, en Suisse, au Grand-Duché de Luxembourg ou au Québec. Nous avons également diffusé le questionnaire au sein de groupes de neuropsychologues francophones sur les réseaux sociaux, notamment *Facebook* et *LinkedIn*. Lors du lancement du questionnaire en mars 2025, des messages ont également été diffusés par le biais de différentes associations professionnelles, à savoir *Whats'Up Neuropsychologie Clinique*, *PSYNCog*, l'*Association Suisse de Neuropsychologie* (ASNP) et la *Société de Neuropsychologie de Langue Française* (SNLF). L'*Association Québécoise des Neuropsychologues* (AQNP) a également diffusé notre questionnaire début mai 2025.

2. Matériel

Afin d'évaluer les capacités de lecture critique des participants, ceux-ci ont été amenés à lire deux ECR ayant pour objectif d'évaluer l'efficacité de deux types de prises en charge en neuropsychologie, avec des différences au niveau de leur qualité méthodologique. Nous détaillons ci-après les différentes étapes de notre démarche, de la sélection des articles à l'élaboration du questionnaire.

2.1. Choix des articles

Les deux articles ont été sélectionnés parmi un échantillon d'ECR publiés entre 1985 et 2022, évaluant l'efficacité d'interventions cognitives sur l'attention et les fonctions exécutives d'enfants et d'adolescents atteints de TDA/H. Les articles ont été sélectionnés à partir de quatre bases de données : *CENTRAL*, *Embase*, *Medline* et *PsycInfo* (Blause et al., 2023b). Parmi ces articles, deux ont été retenus : l'un présentant un *reporting* de haute qualité (*reporting* +), et l'autre, un *reporting* de moindre qualité *(reporting -)*.

2.2. Qualité méthodologique des articles

La qualité des articles retenus a été évaluée selon deux critères : le risque de biais et l'exhaustivité du *reporting*. Tandis que le premier a été évalué à l'aide de la grille RoB2 (Sterne et al., 2019), le second a été déterminé à l'aide de la grille CONSORT-SPI (Montgomery et al., 2018). Initialement conçue comme un outil destiné à accompagner les chercheurs dans la rédaction d'articles scientifiques, cette grille a néanmoins été utilisée dans le cadre de notre étude pour évaluer l'exhaustivité et la transparence du *reporting* des ECR – bien que cette utilisation ne corresponde pas à sa fonction première. La sélection des articles et leur analyse ont été réalisées par deux chercheurs indépendants avec, ensuite, une mise en commun des résultats. Un troisième chercheur est également intervenu pour régler les désaccords (Blause et al., 2023b).

L'article *reporting* +, avec un risque de biais estimé comme faible par RoB2 et avec 17 *items* correctement rapportés sur 45 dans CONSORT-SPI, est le suivant :

Dovis, S., Van Der Oord, S., Wiers, R. W., & Prins, P. J. M. (2015). Improving Executive Functioning in Children with ADHD : Training Multiple Executive Functions within the Context of a Computer Game. A Randomized Double-Blind Placebo Controlled Trial. *PLoS ONE, 10*(4), Article e0121651. https://doi.org/10.1371/journal.pone.0121651.

L'article *reporting* -, avec un risque de biais estimé comme élevé par RoB2 et avec 7 *items* correctement rapportés sur 45 dans CONSORT-SPI, est le suivant :

Kray, J., Karbach, J., Haenig, S., & Freitag, C. (2012). Can Task-Switching Training Enhance Executive Control Functioning in Children with Attention Deficit/-Hyperactivity Disorder ? *Frontiers In Human Neuroscience, 5*. https://doi.org/10.3389/fnhum.2011.00180.

2.2.1. Risque de biais

L'article *reporting* + est considéré par la grille RoB2 comme étant à faible risque de biais. Cette estimation a été obtenue car chacun des domaines méthodologiques évalués par cet outil présente un risque faible de présence de biais (voir Annexe 1). En ce qui concerne le premier domaine, relatif aux biais dans le processus de randomisation, le risque de biais est faible. En effet, la séquence d'allocation a été générée aléatoirement et dissimulée aux participants jusqu'à leur assignation à l'une des conditions expérimentales. De plus, les différences initiales entre les groupes ne suggèrent pas de problème dans la randomisation. Pour le deuxième domaine, portant sur les biais liés aux écarts par rapport à l'intervention prévue, le risque est également faible. Dans cette étude, ni les enfants, ni les parents, ni les personnes qui délivrent l'intervention ne sont conscients du groupe dans lequel l'enfant a été assigné durant l'intervention. De plus, des analyses considérées comme appropriées ont été utilisées afin de déterminer l'effet de l'assignement à l'intervention. Concernant le troisième domaine, relatif aux biais dus aux données manquantes, le risque est faible en raison d'un faible taux d'abandon au cours de l'étude. Pour le quatrième domaine, traitant des biais dans la mesure des résultats, la méthode d'évaluation des *outcomes* est jugée adaptée, les procédures de mesure ont été appliquées de manière similaire entre les groupes et les évaluateurs étaient également aveugles à l'affectation des participants. Ces différents éléments conduisent à une estimation faible du risque de biais à l'étape de mesure des *outcomes*. Enfin, pour le cinquième domaine, le risque de biais lié à la sélection des données reportées dans l'article est faible également, car ce qui est réalisé semble être en accord avec le plan d'analyse qui a été prédéfini par les auteurs. De plus, tous les résultats et les analyses effectués sont reportés dans la partie résultats de l'article.

À l'inverse, l'article *reporting* - est considéré par la grille RoB2 comme étant à haut risque de biais. Cette estimation résulte du fait que plusieurs des domaines méthodologiques évalués par cet outil présentent un risque élevé de présence de biais (voir Annexe 1). Concernant le premier domaine, des préoccupations subsistent quant à de possibles biais liés au processus de randomisation. L'étude fournit très peu d'informations à ce sujet, ce qui empêche le lecteur d'évaluer clairement le risque de biais méthodologique. Cependant, l'absence de différences majeures entre les groupes initiaux ne suggère pas de problème évident dans la randomisation. Pour le deuxième domaine, un risque élevé de biais est associé à des écarts potentiels par rapport à l'intervention prévue. Aucune information n'est donnée sur l'aveuglement des participants ou de la personne administrant l'intervention, ni sur le déroulement concret de cette intervention.

De plus, aucun détail n'est fourni concernant d'éventuelles analyses visant à évaluer l'effet de l'assignation à l'intervention. En ce qui concerne le troisième domaine, le nombre important d'abandons en cours d'étude, associé à l'absence de données permettant de juger de leur impact, soulève certaines préoccupations quant au biais lié aux données manquantes, comme le souligne l'outil RoB2. Pour le quatrième domaine, la méthode de mesure *outcomes* semble appropriée, et les groupes ont été évalués de manière comparable. Toutefois, l'absence d'information sur l'aveuglement des évaluateurs conduit à un risque de biais élevé dans cette phase de l'étude. Enfin, concernant le cinquième domaine, le risque de biais dans la sélection des résultats rapportés est jugé faible. Les analyses présentées sont cohérentes avec le plan d'analyse préétabli, et tous les résultats obtenus sont inclus dans la section correspondante de l'article.

2.2.2. Exhaustivité du *reporting*

Parmi les 45 *items* qui composent la grille CONSORT-SPI, 17 sont rapportés de façon complète dans l'article *reporting* + (voir Annexe 2). Dans les premières pages de l'article, nous pouvons voir que le *design* de l'étude est évoqué dans le titre et que l'introduction de l'article semble complète et compréhensible. En ce qui concerne la méthodologie, les auteurs rapportent qu'il n'y a pas eu de changement par rapport à ce qui était initialement prévu (méthodologie et *outcomes*). Ils détaillent également les critères d'inclusions et d'exclusions des participants, mais aussi comment la taille de l'échantillon a été calculée. Ils fournissent assez de détails concernant l'intervention pour permettre au lecteur de la répliquer. De plus, ils décrivent les similarités entre les groupes d'intervention et comment ils se sont assurés que l'intervention se déroulait comme prévu. Des informations sur la gestion des données manquantes sont également fournies dans l'article. Au niveau des résultats, les pertes et les exclusions de participants après le début de l'intervention sont rapportées de façon précise et toutes les données récoltées sont disponibles. Enfin, la plupart des informations importantes demandées par CONSORT-SPI telles que la déclaration de conflits d'intérêt, les sources de financement, l'implication du développeur de l'intervention, les récompenses données aux participants et l'implication des différentes parties prenantes sont rapportées dans l'article original.

En ce qui concerne l'article reporting -, 7 *items* sur les 45 qui composent la grille CONSORT-SPI sont rapportés de façon complète (voir Annexe 2). Ainsi, en comparaison avec l'article *reporting* +, les éléments manquants réduisent la transparence de cet article. Tout d'abord, l'absence de mention explicite du type de *design* dans le titre complique l'identification de l'article comme étant pertinent ou non pour le lecteur, en fonction de ses besoins. De plus, les

auteurs ne précisent pas si des modifications méthodologiques sont intervenues après le début de l'essai. Ce manque d'informations empêche le lecteur de déterminer si l'étude a été menée conformément au protocole initial ou si des ajustements, potentiellement révélateurs de difficultés d'implémentation, ont été nécessaires – ce qui est pourtant essentiel pour une éventuelle reproduction de l'intervention. L'absence de données sur le calcul de la taille d'échantillon limite également la compréhension de la puissance statistique de l'étude ainsi que de l'ampleur de l'effet a priori. Cette omission nuit à la transparence méthodologique et complique la reproductibilité de la recherche. Par ailleurs, les abandons de participants sont fréquents dans ce type d'étude. L'absence d'informations concernant la gestion de ces abandons, leur nombre et les raisons de ceux-ci entrave la compréhension des résultats et des éventuels effets indésirables liés à l'intervention. Ne pas fournir l'accès aux données collectées empêche les lecteurs de vérifier l'intégrité des résultats et de s'assurer qu'aucune manipulation ou fraude n'a eu lieu. De même, le manque de transparence concernant les parties prenantes, les conflits d'intérêts potentiels et l'implication des développeurs limite la capacité du lecteur à juger de l'impartialité des résultats. Enfin, bien que l'élaboration et la diffusion d'un protocole d'étude soient fortement recommandées (Hardwicke & Wagenmakers, 2023 ; Tan et al., 2019), l'absence de lien vers un tel protocole dans l'article *reporting* - constitue une faiblesse supplémentaire affectant la qualité globale de l'étude. Il convient toutefois de souligner certaines forces de l'article *reporting* -, absentes dans l'article *reporting* +. En effet, bien que présentant certaines limitations, les auteurs en discutent de manière explicite dans la section dédiée. De plus, ils abordent la question de la généralisabilité de leurs résultats dans la discussion, ce qui renforce la portée interprétative de l'étude.

2.3. Travail d'uniformisation des articles

Lors de la création du questionnaire, nous avons travaillé à l'uniformisation des articles. Des modifications spécifiques ont été apportées pour rendre ceux-ci méconnaissables : le nom des auteurs et des revues ainsi que la bibliographie ont été supprimés, la police a été modifiée (*Times New Roman*, taille 12) et les citations bibliographiques ont été remplacées par des références chiffrées. L'objectif de ces ajustements était double : d'une part, éviter que des éléments de forme n'influencent la perception de la qualité méthodologique entre les deux groupes et, d'autre part, garantir que les articles ne soient pas reconnus par les participants. L'article *reporting* + est présenté en Annexe 3 et l'article *reporting* - en Annexe 4.

2.4. Élaboration du questionnaire

Lors de l'enquête, les participants ont non seulement été invités à lire les deux articles scientifiques mentionnés précédemment, mais ont également dû répondre à différentes pages/sections de questions. Ils devaient répondre à la totalité des questions d'une page pour passer à la suivante et n'avaient pas la possibilité de revenir en arrière une fois la page validée.

Pour élaborer le questionnaire, notamment les questions portant sur la qualité méthodologique des articles, nous nous sommes appuyés sur les grilles CONSORT-SPI et RoB2 (Montgomery et al., 2018 ; Sterne et al., 2019). Nous avons également pris en compte les biais principaux et les problèmes méthodologiques susceptibles d'être rencontrés lors de la lecture et de l'analyse critique d'articles scientifiques, tels que présentés dans l'introduction théorique de ce travail (Blause et al., 2023a ; Faulkner et al., 2008).

2.5. Structure et contenu du questionnaire

2.5.1. Introduction

L'introduction du questionnaire avait pour but de fournir aux participants des informations détaillées sur les objectifs de cette enquête, son déroulement ainsi que les conditions de participation. Les participants ont été informés que l'étude visait à explorer la manière dont les psychologues spécialisés en neuropsychologie utilisaient la recherche scientifique dans leur pratique professionnelle. Après cela, un message soulignait l'importance de compléter le questionnaire en une seule fois ou en tout cas de ne pas faire de pause après la lecture de chacun des articles en raison de l'impossibilité de revenir en arrière ou de laisser des réponses en suspens. Cet avertissement visait à minimiser les oublis et/ou les réponses au hasard. Pour encore plus de sécurité, à la fin de l'enquête, il était demandé aux participants de préciser s'ils avaient interrompu le questionnaire et, le cas échéant, à quel moment.

2.5.2. Questions

Le questionnaire comportait une cinquantaine de questions, réparties en une dizaine de pages/sections distinctes. Le format des questions était varié : questions de type oui-non, questions à choix multiples, questions à réponses uniques parmi plusieurs options ou encore questions à réponses ouvertes. Le Tableau 1 présente une synthèse des thématiques abordées dans le questionnaire, ce dernier étant consultable en Annexe 5.

Tableau 1. Synthèse des thématiques abordées au sein du questionnaire

QUESTIONNAIRE
Section 1 : Informations générales
Confirmation d'être psychologue spécialisé en neuropsychologie et d'exercer dans un pays francophone
Genre
Nombre d'années d'exercice en tant que chercheur
Pays d'exercice actuel
Diplôme obtenu
Section 2 : Contexte professionnel
Situation professionnelle actuelle
Statut occupé actuellement
Réalisation de tâches d'enseignement ou non
Activité en neuropsychologie clinique ou non Si oui : - Temps consacré à la clinique par semaine - Type de population rencontré - Type de service
Publication d'articles scientifiques dans une revue internationale avec *peer-reviewing* ou non - Si oui : nombre d'articles publiés
Participation à la réalisation d'un essai contrôlé randomisé ou non
Publication d'articles scientifiques portant sur des prises en charge/ revalidations en neuropsychologie ou des ECR dans une revue internationale avec *peer-reviewing* ou non Si oui : - Nombre d'articles prises en charge/ revalidations publiés en tant que 1er auteur, co-auteur et dernier auteur - Nombre d'ECR publiés en tant que 1er auteur, co-auteur et dernier auteur - Nombre d'ECR prises en charge/ revalidations publiés en tant que 1er auteur, co-auteur et dernier auteur
Utilisation d'un outil de *reporting* pour aider dans la rédaction d'un article portant sur un ECR ou non - Si oui : indication de l'outil - Si non : explication
Section 3 : Habitudes de lecture
Fréquence de lecture d'études scientifiques portant sur des prises en charge au quotidien
Critères vérifiés pour évaluer la qualité des articles scientifiques
Consentement à la lecture des articles ou non
LECTURE DU PREMIER ARTICLE
Section 4 : Pratiques de lecture et questions de contrôle
Confirmation de la lecture de l'article dans son intégralité ou identification des parties lues
Article déjà lu auparavant ou non
Familiarité avec le sujet traité dans l'article ou non
Indication du groupe sur lequel l'intervention a été testée
Indication de l'intervention administrée dans l'étude
Indication du nombre de groupes d'intervention
Section 5 : Confiance donnée aux résultats et identification des biais
Niveau de confiance dans les résultats de l'article (avant)
Identification de biais méthodologiques et précision
Degré d'accord pour la compréhension de chaque section de l'article, pour la qualité méthodologique ainsi que pour la suffisance des informations pour reproduire l'intervention, en apprécier l'efficacité et en juger l'utilité
Recherche de l'article dans un moteur de recherche ou non
Utilisation d'un outil d'aide à la lecture (traducteur, grille de lecture, etc.) ou non - Si oui : indication de l'outil
Section 6 : Évaluation critique de l'article
Indication du degré d'attention pour chaque critère de qualité méthodologique
Niveau de confiance dans les résultats de l'article (après)
LECTURE DU SECOND ARTICLE
Sections 7, 8 et 9
Idem que sections 4, 5 et 6
Section 10 : Fin du questionnaire
Spécification de l'article possédant la meilleure qualité méthodologique

Tout d'abord, une première section permettait de recueillir des données démographiques sur les participants, comme leur genre, leurs années de pratique et leur(s) diplôme(s). Une seconde section explorait leur contexte professionnel en tant que chercheur, mais aussi éventuellement en tant que neuropsychologue clinicien ou enseignant. Une troisième section s'intéressait à leurs habitudes de lecture d'articles scientifiques, notamment les informations auxquelles ils prêtent attention pour déterminer si un article est de bonne qualité.

Ensuite, les participants étaient invités à lire un premier article scientifique, soit l'article *reporting* +, soit l'article *reporting* -. En effet, l'ordre de présentation changeait en fonction de la randomisation (voir infra). Avant de commencer cette étape, leur consentement était requis. En cas de refus, ils étaient directement redirigés vers la fin du questionnaire afin d'éviter des réponses aléatoires. Les participants qui refusaient de participer à l'étude, ceux qui ne répondaient qu'aux questions démographiques ou ceux qui fournissaient des réponses incorrectes aux questions de contrôle sur le contenu des articles n'étaient pas pris en compte.

Avant de débuter la lecture, les participants étaient informés qu'ils devaient lire l'article comme à leur habitude, avec la possibilité de prendre des notes, mais que la lecture se ferait à l'écran. À l'issue de la lecture, une première série de questions visait à évaluer leur compréhension de l'article, incluant également des *items* sur leur familiarité avec celui-ci afin de vérifier s'ils l'avaient déjà lu auparavant. La section suivante portait sur leur niveau de confiance envers les résultats de l'ECR ainsi que sur l'éventuelle présence de biais méthodologiques susceptibles d'altérer ce niveau de confiance. Une troisième section invitait les participants à se positionner sur différents aspects de l'étude à l'aide d'une échelle en 4 points (de 1 = *Pas du tout d'accord* à 4 = *Parfaitement d'accord*), puis à indiquer les éléments pris en compte pour juger de la qualité de l'article via une échelle en 5 points (1 = *Je n'y ai pas prêté attention*, 2 = *J'y ai prêté attention et j'ai eu l'impression que c'était réalisé de façon satisfaisante*, 3 = *J'y ai prêté attention et j'ai eu l'impression que ce n'était pas réalisé ou réalisé de façon insatisfaisante*, 4 = *J'y ai prêté attention mais je ne m'en souviens plus*, et 5 = *J'y ai prêté attention mais je ne peux me positionner quant à la qualité*). Pour terminer, leur niveau de confiance était à nouveau mesuré afin d'identifier d'éventuelles variations après complétion du questionnaire. Celui-ci se clôturait par une question portant sur une éventuelle interruption en cours de passation (et le moment concerné) ainsi que par un espace libre pour formuler leurs commentaires ou suggestions.

2.5.3. Changements amenés

Étant donné le grand nombre d'abandons lors du premier mois de récolte des résultats, nous avons procédé à un remaniement du questionnaire. Nous avons fait passer les sections 2 et 3 avant la section 10 (voir Tableau 1). Le questionnaire commençait donc par les questions démographiques, suivies de celle portant sur les critères utilisés pour évaluer la qualité des articles scientifiques (issue de la section 3), puis enchaînait directement avec la lecture du premier article. Nous avons estimé qu'entamer rapidement la lecture pourrait favoriser l'intérêt et l'implication des participants. En leur proposant d'emblée un premier article, ils seraient ainsi plus à même de comprendre le sens de leur participation et l'objectif de l'étude, ce qui pourrait les inciter à compléter l'intégralité du questionnaire.

3. Procédure générale

La procédure générale de cette étude s'est déroulée en deux étapes principales : une phase de pré-tests visant à valider le questionnaire, suivie d'une phase expérimentale au cours de laquelle l'enquête a été diffusée en ligne.

Dans un premier temps, un pré-test a été réalisé en soumettant le questionnaire à deux pré-évaluatrices : deux chercheuses de l'Unité de Neuropsychologie de l'Adulte de l'Université de Liège. Cette étape visait à s'assurer que le questionnaire était complet et ne présentait aucun problème opérationnel ou de contenu. À la suite de leurs retours, des ajustements ont été apportés afin d'améliorer la clarté du questionnaire et un second pré-test a alors été mené auprès de quatre doctorantes de l'Unité de Neuropsychologie de l'Adulte de l'Université de Liège. Ainsi, cette phase de pré-tests a permis de valider l'exhaustivité et la cohérence du questionnaire avant d'entamer la phase expérimentale.

Dans un second temps vient alors la phase expérimentale au cours de laquelle l'enquête a été diffusée en ligne. Se présentant sous la forme d'une enquête électronique, elle a été hébergée sur la plateforme interne de l'Université de Liège (*Surveys*), plateforme conçue conformément aux normes européennes de protection des données. L'enquête était librement accessible et accompagnée d'une description claire du public cible. Le sondage a commencé le 18 février 2025 et l'analyse des données a été finalisée à la moitié du mois de juin 2025.

Avant de débuter le questionnaire, les participants ont donné leur consentement pour l'utilisation de leurs réponses dans le cadre de cette étude. La réalisation de cette enquête a été approuvée par le comité d'éthique de la Faculté de Psychologie, Logopédie et Sciences de l'Éducation de l'Université de Liège (référence du dossier : 2324-111). Les participants ayant accepté de participer à cette enquête ont reçu un lien unique qui les dirigeait de manière aléatoire vers l'une des deux versions de l'enquête. Dans l'une, l'article *reporting* + était présenté en premier, suivi de l'article *reporting* -, tandis que dans l'autre version, l'ordre était inversé. La durée maximale de complétion de l'enquête ne devait pas dépasser 180 minutes.

4. Analyses statistiques prévues

Toutes les analyses statistiques seront réalisées avec la version 2.3.28 du logiciel Jamovi (The Jamovi project, 2022). Dans un premier temps, une analyse descriptive sera menée afin d'examiner les données démographiques et professionnelles des participants ainsi que leurs habitudes de lecture scientifique. Nous mènerons également des analyses descriptives sur l'identification des biais dans les articles et les critères méthodologiques pris en compte ou non. Ensuite, une analyse qualitative portera sur les réponses des participants quant aux critères mobilisés pour évaluer la qualité d'un article au quotidien ainsi que les biais méthodologiques identifiés dans chacun des articles lus.

En termes d'analyses statistiques, des tests *t* de Student pour échantillons appariés seront effectués, sous réserve que les données suivent une distribution normale (vérifiée à l'aide du test de Shapiro-Wilk), afin de comparer la confiance initiale accordée par les chercheurs aux résultats des deux articles, leur compréhension et leur perception des différentes sections des articles (introduction, méthodologie, résultats, discussion), mais aussi leur évaluation de la qualité méthodologique et leur perception de disposer de suffisamment d'informations pour reproduire l'intervention, en apprécier l'efficacité et en juger l'utilité. En cas de non-normalité, le test des rangs signés de Wilcoxon sera privilégié.

Par ailleurs, afin de comparer ces mêmes éléments entre les chercheurs et les cliniciens en ce qui concerne l'article *reporting* +, des tests *t* de Student pour échantillons indépendants seront effectués, sous réserve du respect des conditions de normalité et d'homogénéité des variances (vérifiées respectivement avec les tests de Shapiro-Wilk et de Levene). En cas de non-respect de la condition d'homogénéité des variances uniquement, le test *t* de Welch sera utilisé. En cas

de non-normalité (ou si les deux conditions sont violées), le test non paramétrique de Mann-Whitney sera privilégié.

Ensuite, pour examiner l'évolution des jugements avant et après les questions portant sur les biais méthodologiques identifiés dans les articles ainsi que sur les critères pris en compte pour évaluer leur qualité, un test *t* de Student pour échantillons appariés sera appliqué, sous réserve du respect de la condition de normalité. En cas de non-respect de cette condition, le test des rangs signés de Wilcoxon sera privilégié.

En ce qui concerne la comparaison des biais méthodologiques identifiés par les chercheurs et les cliniciens dans l'article *reporting* +, un test du khi carré d'indépendance sera effectué. Si certaines modalités présentent des effectifs insuffisants, le test exact de Fisher sera privilégié.

Pour l'ensemble des analyses, l'interprétation des tailles d'effet repose sur les conventions proposées par Cohen (1988). Ainsi, pour le *d* de Cohen, les seuils retenus sont les suivants : 0.20 pour un petit effet, 0.50 pour un effet moyen et 0.80 pour un effet important. Concernant les corrélations bisériées de rang (r_{bis}), les seuils retenus sont les suivants : 0.10 pour un effet faible, 0.30 pour un effet modéré et 0.50 pour un effet élevé. Enfin, pour le coefficient phi (φ), nous nous référons également aux seuils de Cohen (1988), à savoir : 0.10 pour un effet faible, 0.30 pour un effet modéré et 0.50 pour un effet élevé.

Résultats

Dans cette section portant sur les résultats de notre étude, nous commençons par présenter le flux des participants ainsi que leurs caractéristiques démographiques et professionnelles. Nous décrivons ensuite leurs habitudes de lecture, la manière dont ils ont abordé les deux ECR ainsi que leur niveau de confiance envers les résultats. Nous analysons par après leur capacité à identifier les biais, leur compréhension des différentes sections de l'article et les critères utilisés pour en apprécier la qualité. Enfin, nous comparons leurs compétences à celles des neuropsychologues cliniciens avant de conclure sur les principaux résultats obtenus.

1. Résultats côté chercheurs

1.1. Flux des participants

En ce qui concerne la première version du questionnaire, 14 participants l'ont entamée, mais seulement 5 l'ont complétée intégralement (taux d'abandon de 64 %). Quant à la seconde version du questionnaire, 17 participants l'ont débutée, mais seuls 8 sont allés jusqu'au bout (taux d'abandon de 53 %). Au total, 13 participants ont répondu à l'ensemble du questionnaire. Parmi les abandons (N = 18), nous avons choisi de conserver les données des participants ayant complété la section sur les habitudes de lecture (N = 8), c'est-à-dire celle avant la lecture du premier ECR, ces réponses étant pertinentes pour nos objectifs. Par conséquent, les participants ayant abandonné avant cette section (N = 10) ont été exclus. L'échantillon ayant servi à l'analyse des habitudes de lecture comprend ainsi 21 participants (voir Figure 2).

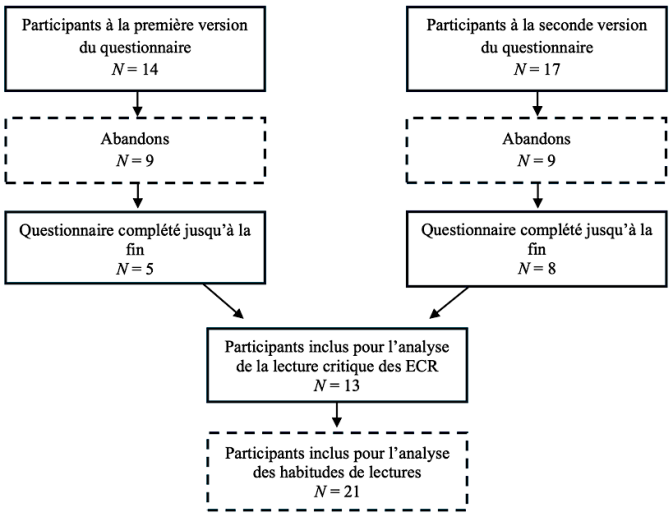

Figure 2. Diagramme du flux des participants

1.2. Caractéristiques démographiques et professionnelles

Parmi les 21 participants inclus dans l'analyse des habitudes de lecture, la majorité se compose de femmes (81 %, *N* = 17) et exerce en Belgique (81 %, *N* = 17). La répartition selon les années d'expérience est relativement équilibrée. La plupart des participants sont titulaires d'un Master en psychologie avec une spécialisation en neuropsychologie (67 %, *N* = 14), plus d'un tiers détient un Doctorat (38 %, *N* = 8) et un peu moins d'un tiers a réalisé un post-doctorat (24 %, *N* = 5). Leurs caractéristiques démographiques et professionnelles sont consultables en Annexe 6 et Annexe 7.

Parmi l'échantillon de participants ayant complété l'intégralité du questionnaire (*N* = 13), incluant la lecture et l'analyse critique des articles, on retrouve une majorité de femmes (77 %, *N* = 10) et une exclusivité de chercheurs exerçant en Belgique. Plus d'un tiers exerce depuis moins d'un an ou depuis 1 à 5 ans (38 %, *N* = 5), et la majorité est titulaire d'un Master en psychologie avec une spécialisation en neuropsychologie (77 %, *N* = 10) (voir Tableau 2).

Tableau 2. Données démographiques des participants

Données démographiques	***N***	**%**
Genre		
Homme	3	23
Femme	10	77
Autre	0	0
Pays d'exercice		
France	0	0
Belgique	13	100
Canada	0	0
Suisse	0	0
Grand-Duché de Luxembourg	0	0
Autre	0	0
Années d'activité		
Moins d'1 an	5	38
De 1 à 5 ans	5	38
De 6 à 10 ans	3	23
Plus de 10 ans	0	0
Diplômes		
Master (ou DESS) de psychologie spécialisée en neuropsychologie ou MAS en neuropsychologie	10	77
Master (ou DESS) de psychologie sans spécialisation en neuropsychologie	2	15
DESS en psychologie obtenu avant 2000 et pratique spécialisée en neuropsychologie depuis plus de 10 ans	0	0
Doctorat	2	15
Post-doctorat	1	8
Diplôme Universitaire (DU)	1	8
Certificat Universitaire (CU)	0	0
Autre	0	0

Notes. N = nombre de participants concernés ; % = pourcentage calculé sur l'échantillon total (N = 13)

Sur le plan professionnel, notre échantillon se compose majoritairement de doctorants (69 %, *N* = 9). La majorité des participants occupe un poste de chercheur à temps plein (69 %, *N* = 9) et près de la moitié participe à l'encadrement de travaux pratiques (46 %, *N* = 6). Seule une minorité exerce parallèlement une activité de neuropsychologue clinicien (23 %, *N* = 3). Enfin, plus d'un tiers a déjà publié dans une revue internationale avec *peer-reviewing* (38 %, *N* = 5) et

près de la moitié a pris part à la réalisation d'un ECR (46 %, *N* = 6). Toutefois, seul un participant a publié un ECR en tant que co-auteur sans avoir utilisé d'outil d'aide à la rédaction, car n'ayant pas de rôle décisionnel dans l'élaboration de l'article (voir Tableau 3).

Tableau 3. Contexte professionnel des participants

Contexte professionnel	***N***	**%**
Situation professionnelle actuelle		
Chercheur à temps plein	9	69
Chercheur à temps partiel – Plus d'un mi-temps mais pas à temps plein	1	8
Chercheur à temps partiel – Mi-temps	2	15
Chercheur à temps partiel – Moins d'un mi-temps	1	8
Sans activité de chercheur	0	0
Statut actuel		
Doctorant	9	69
Post-doctorant	1	8
Chercheur permanent/ enseignant à l'université	0	0
Autre	3	23
Tâches d'enseignement		
Aucune	6	46
Cours théoriques	3	23
Travaux pratiques	6	46
Autre	0	0
Activité de neuropsychologue clinicien		
Oui	3	23
Non	10	77
Temps consacré à la clinique par semaine (*N* = 3)		
Moins de 10 %	0	0
10 à 25 %	0	0
25 à 50 %	3	100
Plus de 50 %	0	0
Population clinique rencontrée (*N* = 3)		
Enfants	1	33
Adolescents	1	33
Adultes	2	67
Adultes âgés (> 60 ans)	2	67
Secteur d'exercice actuel (*N* = 3)		
Neurologie, neuropédiatrie	0	0
Gériatrie (service de gériatrie, EHPAD, etc.)	1	33
Consultation mémoire	0	0
Neurochirurgie	0	0
Oncologie	0	0
Psychiatrie, pédo-psychiatrie, géronto-psychiatrie, addictologie	0	0
Rééducation, réhabilitation	1	33
Médico-social ou médico-éducatif (FAM, UEROS, SESSAD, etc.)	0	0
Recherche, université	1	33
Libéral	1	33
Association	0	0
Organisme de formation	0	0
Autre	1	33
Publication d'articles scientifiques dans une revue internationale avec *peer-reviewing*		
Oui	5	38
Non	8	62
Nombre d'articles scientifiques publiés dans une revue internationale avec *peer-reviewing* (*N* = 5)		
1-5	2	40
5-10	1	20
10-50	2	40
Plus de 50	0	0
Participation à la réalisation d'un essai contrôlé randomisé		
Oui	6	46
Non	7	55
Publication d'articles scientifiques portant sur des prises en charge/ revalidations en neuropsychologie ou des essais contrôlés randomisés dans une revue internationale avec *peer-reviewing* (en 1e auteur ou non)		
Oui	1	8
Non	12	92
Utilisation d'un outil de reporting lors de la rédaction d'un essai contrôlé randomisé (*N* = 1)		
Oui	0	0
Non	1	100

Notes. N = nombre de participants concernés ; % = pourcentage calculé sur l'échantillon total (N = 13) sauf mention spécifique

1.3. Habitudes de lecture

En ce qui concerne leurs habitudes de lecture, nous avons interrogé les participants sur la fréquence à laquelle ils consultaient des articles scientifiques portant sur des prises en charge au quotidien. L'ensemble des 21 répondants rapporte des fréquences de lecture hétérogènes, la plus courante étant une lecture exceptionnelle (43 %, $N = 9$) (voir Figure 3).

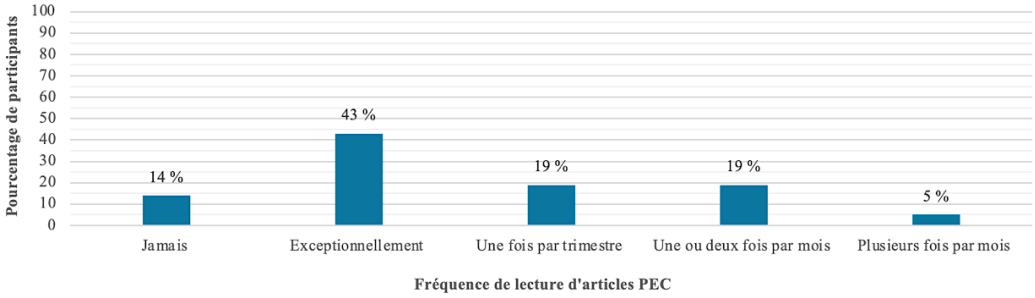

Figure 3. Pourcentage de participants selon la fréquence de lecture d'articles scientifiques portant sur des prises en charge

Par ailleurs, nous avons analysé les critères mobilisés pour évaluer la qualité d'un article scientifique. L'analyse des réponses recueillies met en évidence la récurrence de plusieurs critères méthodologiques, en accord avec les recommandations de la déclaration CONSORT-SPI (voir Annexe 8 pour le détail des réponses).

Le critère le plus fréquemment cité concerne la clarté et la précision de la méthodologie (62 %, $N = 13$), incluant notamment la description des procédures, les outils utilisés et/ou la possibilité de répliquer l'étude. La taille de l'échantillon et/ou la puissance statistique de l'étude sont également mentionnées par plus d'un tiers des participants (38 %, $N = 8$). Par ailleurs, certains soulignent l'importance de la validité des analyses statistiques (24 %, $N = 5$) ainsi que celle d'un plan d'étude rigoureux et approprié (19 %, $N = 4$), intégrant notamment le *blinding*, la randomisation et/ou la comparabilité des groupes. D'autres critères, bien que moins fréquemment évoqués, sont également rapportés : la taille d'effet (14 %, $N = 3$), la clarté de la question de recherche (14 %, $N = 3$), la transparence des auteurs quant aux limites de leur étude (5 %, $N = 1$) et la déclaration des conflits d'intérêts (5 %, $N = 1$).

1.4. Lecture des ECR

Pour rappel, seuls 13 participants ont lu et analysé les deux ECR. Les analyses présentées dans ce point porteront donc exclusivement sur cet échantillon.

1.4.1. Pratiques de lecture

Concernant l'article *reporting* +, la majorité des participants (69 %, *N* = 9) en a effectué une lecture intégrale. Parmi les autres, les omissions concernent principalement les sections relatives aux résultats (textuels ou en tableau) et aux instruments. De manière similaire, l'article *reporting* - a également été lu dans son intégralité par la majorité des participants (62 %, *N* = 8). Quant aux omissions les plus fréquentes, elles portent sur les résultats, les instruments, la discussion ou encore, pour un participant, l'introduction et la conclusion.

Quel que soit l'article, aucun participant ne l'a lu auparavant et près de la moitié se déclare familière avec le sujet abordé (46 %, *N* = 6). Tous ont répondu correctement aux questions de vérification de la compréhension. Hormis un participant ayant recherché l'article *reporting* + en ligne, aucun participant n'a eu recours à un moteur de recherche ou n'a utilisé le moindre outil tel qu'un traducteur ou une grille d'analyse critique.

1.4.2. Niveau de confiance accordé aux résultats

Nous nous intéressons ici au niveau de confiance accordé par les participants aux résultats de chaque article. Pour rappel, cette évaluation a été réalisée, à l'aide d'une échelle graduée de 1 à 10, à deux moments distincts.

1.4.2.1. Comparaison des niveaux de confiance initiaux

Nous avons examiné si le niveau de confiance initial différait d'un article à l'autre. L'hypothèse alternative unilatérale postulait que le niveau de confiance initial pour l'article *reporting* + serait supérieur à celui pour l'article *reporting* - (H_a : μ*Reporting*+ > μ*Reporting*-).

Le test de normalité révèle que, pour les deux articles, les niveaux de confiance initiaux suivent une distribution normale (voir Annexe 9). Les résultats du test *t* de Student montrent quant à eux une différence significative de la confiance initiale, avec une forte taille d'effet.

Tableau 4. Résultats du test t de Student pour échantillons appariés

	M(ET)		***t***	***ddl***	***p***	***d***	***IC 95%***	
	R+	***R-***					***Min***	***Max***
Confiance *reporting* + vs. *reporting* -	7.77 (1.36)	4.54 (2.96)	3.41	12	.003*	0.95	0.27	1.59

*Notes. * p < .05 ; M(ET) = Moyenne et écart-type ; t = Statistique t de Student ; ddl = degrés de liberté ; d = d de Cohen ; IC 95% = Intervalle de confiance à 95%*

1.4.2.2. Évolution du jugement après l'analyse approfondie des biais

Après analyse du niveau de confiance initial, nous avons examiné l'évolution des jugements de confiance des participants quant aux résultats de chaque ECR. Nos hypothèses alternatives unilatérales postulaient que, pour l'article *reporting* +, le niveau de confiance augmenterait (H_a : μAvant < μAprès), tandis que, pour l'article *reporting* -, il diminuerait (H_a : μAvant > μAprès). Ces observations pourraient s'expliquer par la nature des questions posées, incitant les participants à adopter une lecture critique de la qualité méthodologique des articles.

Pour les deux articles, le test de Shapiro-Wilk révèle un écart significatif à la normalité (voir Annexe 10). Les résultats n'indiquent pas de variation significative du niveau de confiance après l'analyse critique, quel que soit l'article. La taille d'effet observée est modérée pour l'article *reporting* +, tandis qu'elle reste faible pour l'article *reporting* -.

Tableau 5. Résultats du test des rangs signés de Wilcoxon

	Me		*M(ET)*		*W*	*p*	r_{bis}
	Confiance avant	**Confiance après**	**Confiance avant**	**Confiance après**			
Confiance avant vs. après – *reporting* +	8	8	7.77 (1.36)	6.31 (3.07)	24	.963	.71
Confiance avant vs. après – *reporting* -	5	6	4.54 (2.96)	5.85 (2.34)	18	.781	.29

Notes. Me = Médiane ; M(ET) = Moyenne et écart-type ; W = Statistique W de Wilcoxon ; r_{bis} = corrélation entre rangs bisériés

1.4.3. Identification des biais méthodologiques

Pour rappel, il a été demandé aux participants de déterminer s'ils avaient identifié spontanément des biais méthodologiques dans chaque ECR, via une réponse binaire (oui/non). Pour l'article *reporting* +, plus d'un tiers des participants ont indiqué avoir identifié des biais (38 %, $N = 5$). En ce qui concerne l'article *reporting* -, plus de deux tiers en ont relevé (77 %, $N = 10$). Ces participants ont alors été interrogés librement sur les biais identifiés dans chacun des deux articles (voir Annexe 11 et Annexe 12 pour le détail des réponses).

Tableau 6. Proportion de participants ayant repéré des biais par article

Identification de biais	**Article R+**		**Article R-**	
	N	**%**	***N***	**%**
Oui	5	38	10	77
Non	8	62	3	23

Notes. N = nombre de participants concernés ; % = pourcentage calculé sur l'échantillon total (N = 13)

1.4.3.1. Analyse des réponses recueillies

L'analyse des réponses recueillies met en évidence la récurrence de plusieurs biais méthodologiques pouvant être reliés aux recommandations de la déclaration CONSORT-SPI. Certains biais méthodologiques sont relevés dans les deux articles, notamment l'absence de test de puissance, le manque de précision des critères d'inclusion, le manque de mesures, la pertinence des outils et la pertinence du *design* d'étude. D'autres biais, tels que la petite taille de l'échantillon, l'absence de randomisation et de *blinding*, l'absence d'analyse en intention de traiter, la récolte des résultats et la qualité de la discussion, sont essentiellement relevés dans l'article *reporting* -. La répartition détaillée des biais est présentée ci-dessous.

Tableau 7. Biais cités par les participants

Biais cités	**Article R+** (*N* = 5)		**Article R-** (*N* = 10)	
	N	**%**	***N***	**%**
Pas de test de puissance	1	20	3	30
Petit échantillon	0	0	6	60
Absence de randomisation	0	0	1	10
Absence de *blinding*	0	0	2	20
Manque de précision quant aux critères d'inclusion	1	20	5	50
Pas d'analyse en intention de traiter	0	0	1	10
Manque de mesures	1	20	2	20
Pertinence des outils	1	20	3	30
Récolte des résultats	0	0	4	40
Pertinence du *design* d'étude	1	20	4	40
Qualité de la discussion	0	0	1	10

Notes. N = nombre de participants concernés ; % = pourcentage calculé sur l'échantillon spécifié

1.4.3.2. Analyses complémentaires

Premièrement, nous avons analysé de façon descriptive la relation entre la fréquence de lecture d'articles de prise en charge et l'identification de biais dans l'article *reporting* -. Parmi les participants, les trois ayant déclaré ne jamais lire d'articles ont identifié des biais méthodologiques dans l'article. Sur les six participants lisant exceptionnellement ce type d'article, cinq en ont identifié et sur les trois lisant des articles une à deux fois par mois, deux. Quant au participant ayant déclaré lire des articles une fois par semaine ou plus, il n'en a identifié aucun (voir Annexe 13).

Deuxièmement, nous avons examiné de façon descriptive la relation entre l'identification de biais dans l'article *reporting* - et le choix du meilleur article. Sur notre échantillon, 11 participants ont identifié le meilleur article (85 %) et, parmi eux, trois n'ont pas identifié de biais dans l'article *reporting* - (voir Annexe 14).

1.4.4. Compréhension et perception des articles

Pour rappel, les participants ont été interrogés sur leur compréhension et leur perception des différentes sections des articles. Pour chaque item, ils ont dû indiquer leur degré d'accord, sur une échelle de Likert allant de 1 (pas du tout d'accord) à 4 (parfaitement d'accord). Le questionnaire comportait 8 *items* évaluant la compréhension des différentes sections de l'article (introduction, méthodologie, résultats et discussion), l'appréciation globale de la qualité méthodologique et la perception de disposer de suffisamment d'informations pour reproduire l'intervention, en apprécier l'efficacité et en juger l'utilité. Notre hypothèse alternative unilatérale stipulait que les participants attribueraient aux différentes variables un degré d'accord globalement plus élevé pour l'article *reporting* + que pour l'article *reporting* - (H_a : μ*Reporting*+ > μ*Reporting*-).

La normalité de chaque variable a été testée, ce qui a permis de déterminer le test statistique approprié pour chaque comparaison (voir Annexe 15). Les jugements sont significativement plus favorables à l'article *reporting* + concernant la compréhension de l'introduction, l'évaluation de la qualité méthodologique ainsi que la perception de disposer de suffisamment d'informations pour reproduire l'intervention, en apprécier l'efficacité et en juger l'utilité, avec de très fortes tailles d'effet. Aucune différence significative n'a été observée pour la compréhension de la méthodologie, des résultats et de la discussion, bien que les tailles d'effet soient faibles à fortes.

Tableau 8. Résultats des tests des rangs signés de Wilcoxon

	Me		***M(ET)***		***W***	***p***	***r_{bis}***
	R+	**R-**	**R+**	**R-**			
Compréhension de l'introduction	4	3	3.62 (0.51)	3.08 (0.51)	21	.013*	1
Compréhension de la méthodologie	3	3	3.31 (0.63)	3.15 (0.56)	14	.242	.33
Compréhension des résultats	3	3	2.85 (0.56)	2.62 (0.65)	15	.187	.43
Compréhension de la discussion	3	3	3.15 (0.38)	2.69 (0.75)	18.5	.052	.76
Évaluation de la qualité méthodologique	3	2	3.23 (0.44)	2.38 (0.65)	36	.005*	1
Suffisamment d'informations pour juger l'efficacité	3	2	2.92 (0.64)	2.08 (0.64)	36	.006*	1
Suffisamment d'informations pour juger l'utilité	3	2	2.69 (0.48)	2.00 (0.58)	36	.004*	1

*Notes. * p < .05 ; Me = Médiane ; M(ET) = Moyenne et écart-type ; W = Statistique W de Wilcoxon ; r_{bis} = corrélation entre rangs bisériés*

Par ailleurs, les participants jugent l'article *reporting* + significativement plus reproductible, avec une grande taille d'effet.

Tableau 9. Résultats des tests t de Student pour échantillons appariés

	M(ET)		***t***	***ddl***	***p***	***d***	***IC 95%***	
	R+	**R-**					***Min***	***Max***
Suffisamment d'informations pour reproduire l'intervention	3.38 (0.77)	2.46 (0.88)	3.21	12	.004*	0.89	0.23	1.53

*Notes. * p < .05 ; M(ET) = Moyenne et écart-type ; t = Statistique t de Student ; ddl = degrés de liberté ; d = d de Cohen ; IC 95% = Intervalle de confiance à 95%*

1.4.5. Critères examinés pour évaluer la qualité des articles

Une section de l'enquête s'intéressait à l'attention portée par les participants aux critères utilisés pour évaluer la qualité des articles scientifiques. Les réponses possibles étaient : (1) *« Je n'y ai pas prêté attention »*, (2) *« J'y ai prêté attention et j'ai eu l'impression que c'était réalisé de façon satisfaisante »*, (3) *« J'y ai prêté attention et j'ai eu l'impression que ce n'était pas réalisé ou réalisé de façon insatisfaisante »*, (4) *« J'y ai prêté attention, mais je ne m'en souviens plus »*, et (5) *« J'y ai prêté attention, mais je ne peux me positionner quant à la qualité ».*

Dans un premier temps, nous avons analysé le degré d'attention porté par les participants aux critères de qualité d'un ECR. Pour ce faire, les réponses ont été dichotomisées en deux catégories. La catégorie *« Critère non pris en compte ou évaluation non claire »* regroupe les réponses (1), (4) et (5), tandis que la catégorie *« Critère pris en compte »* reprend les réponses (2) et (3). L'Annexe 16 présente en détail la répartition des critères pris en compte ou ignorés par les participants pour chacun des deux articles. L'analyse révèle qu'une majorité de critères ont été pris en compte par plus de la moitié des participants dans les deux articles. Seuls deux critères n'atteignent pas le seuil de 50 % dans les deux articles : la présence d'un intervalle de confiance et l'absence de conflit d'intérêts.

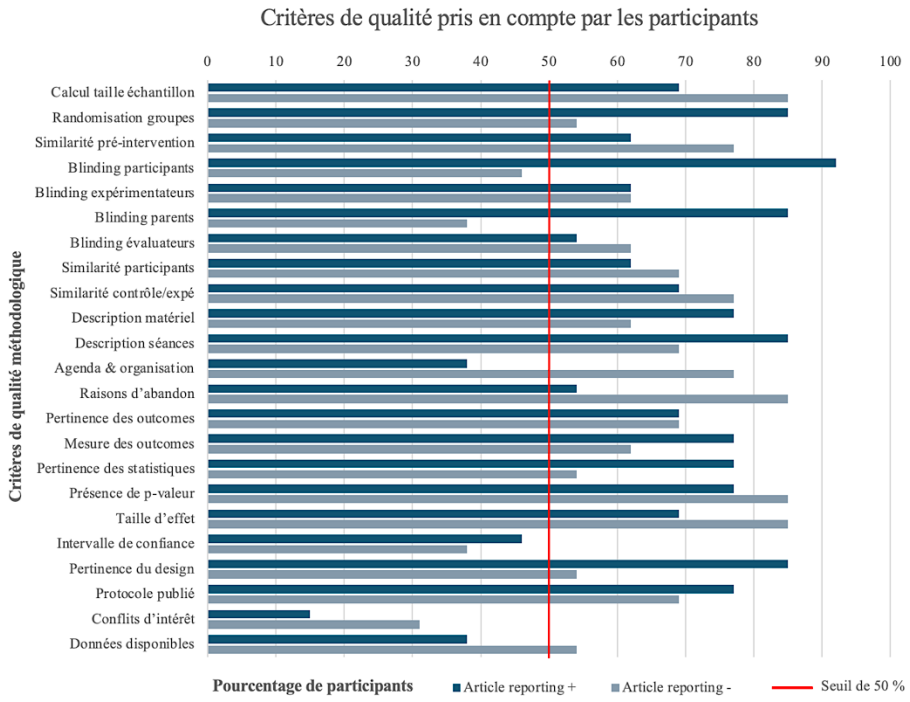

Figure 4. Pourcentage de participants selon l'attention portée aux critères méthodologiques pour les deux articles lus

Nous avons ensuite examiné le pourcentage de bonnes réponses pour chaque critère pris en compte par les participants. Pour ce faire, nous avons uniquement pris en compte les réponses (2) et (3). En fonction de l'article, si la bonne réponse était celle indiquant que le critère est satisfaisant, le participant devait sélectionner la réponse (2). En revanche, si le critère n'était pas jugé satisfaisant, le participant devait choisir la réponse (3). L'Annexe 17 présente en détail la répartition des réponses correctes et incorrectes concernant les critères considérés par les participants lors de l'évaluation de la qualité des deux ECR.

Parmi les critères les plus pris en compte et bien évalués par les participants (≥ 50 %) lors de la lecture des deux articles, on retrouve : le calcul de la taille de l'échantillon, le processus de randomisation, la similarité pré-intervention, le processus de *blinding* des expérimentateurs et des évaluateurs, la *p*-valeur, la taille d'effet ainsi que la publication d'un protocole. Les autres critères, en revanche, ne franchissent ce seuil que dans un seul des deux articles, voire dans aucun d'entre eux.

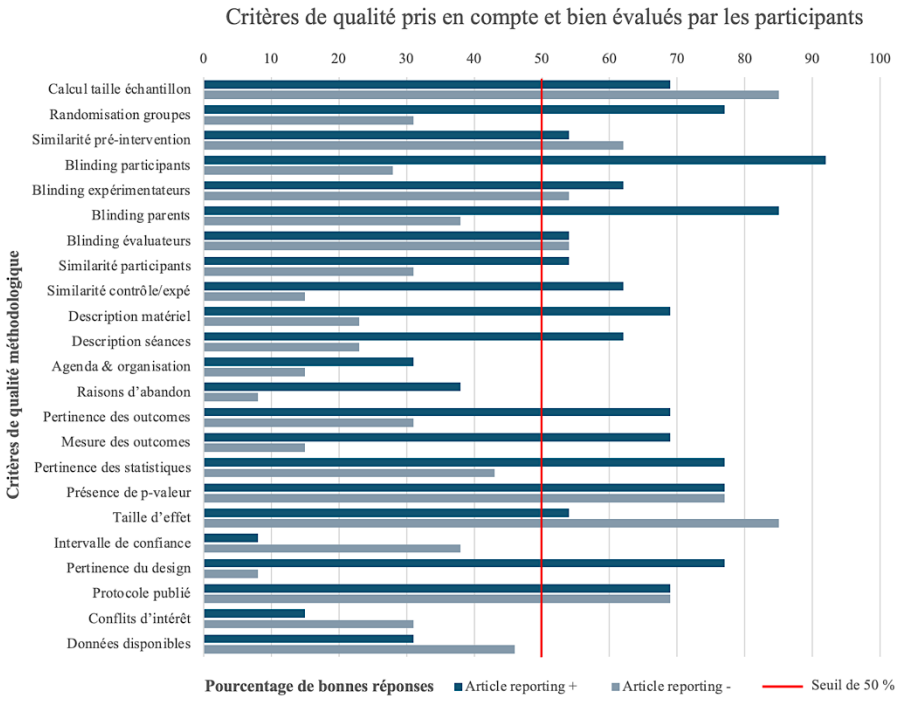

Figure 5. Pourcentage de bonnes réponses pour chaque critère méthodologique pris en compte par les participants dans les deux articles

2. Comparaison entre les chercheurs et les cliniciens

Dans le cadre de ce mémoire réalisé en binôme, nous avons comparé les évaluations de l'article *reporting* + entre chercheurs (65 %, N = 13) et cliniciens (35 %, N = 7). L'échantillon se compose ainsi de 20 participants dont les caractéristiques démographiques et professionnelles sont présentées dans le tableau ci-dessous.

Tableau 10. Données démographiques des participants chercheurs et cliniciens

Données démographiques	***N***	**%**
Statut		
Clinicien	13	65
Chercheur	7	35
Genre		
Homme	5	25
Femme	15	75
Autre	0	0
Pays d'exercice		
France	2	10
Belgique	17	85
Canada	1	5
Suisse	0	0
Grand-Duché de Luxembourg	0	0
Années d'activités		
Moins d'1 an	7	35
De 1 à 5 ans	7	35
De 6 à 10 ans	4	20
Plus de 10 ans	2	10
Diplômes		
Master (ou DESS) de psychologie spécialisée en neuropsychologie ou MAS en neuropsychologie	16	80
Master (ou DESS) de psychologie sans spécialisation en neuropsychologie	2	10
DESS en psychologie obtenu avant 2000 et pratique spécialisée en neuropsychologie depuis plus de 10 ans	0	0
Doctorat	3	15
Post-doctorat	1	5
Diplôme Universitaire (DU)	1	5
Certificat Universitaire (CU)	0	0
Autre	0	0

Notes. N = nombre de participants concernés ; % = pourcentage calculé sur l'échantillon total (N = 20)

2.1. Identification des biais méthodologiques

Nous avons examiné s'il existait une différence entre chercheurs et cliniciens quant à leur capacité à identifier les biais méthodologiques dans l'article. Notre hypothèse alternative unilatérale stipulait que les chercheurs identifieraient davantage de biais que les cliniciens (H_a : P(chercheurs) > P(cliniciens)) et a été évaluée à l'aide du test exact de Fisher.

Les résultats descriptifs montrent une proportion plus élevée de détections de biais chez les chercheurs (N = 6) par rapport aux cliniciens (N = 2). Cependant, le test exact de Fisher n'a pas montré de différence statistiquement significative entre les groupes (voir Annexe 18).

2.2. Compréhension et perception de l'article

Après l'identification des biais, nous avons comparé l'ensemble des scores relatifs à la compréhension et à la perception des différentes sections de l'article, à l'évaluation de sa qualité méthodologique, mais aussi à la perception de disposer de suffisamment d'informations pour reproduire l'intervention, en apprécier l'efficacité et en juger l'utilité, entre chercheurs et cliniciens. Notre hypothèse alternative unilatérale supposait que le degré d'accord des chercheurs serait supérieur à celui des cliniciens pour chacune des variables (H_a : μChercheurs > μCliniciens).

En l'absence de respect des conditions de normalité et d'homogénéité, le test non paramétrique de Mann-Whitney a été retenu pour chaque variable (voir Annexe 19). Seule la perception de la reproductibilité de l'intervention diffère significativement entre les groupes, avec une grande taille d'effet, les chercheurs estimant davantage avoir suffisamment d'informations pour reproduire l'intervention. Aucune différence significative n'est observée pour les autres dimensions.

Tableau 11. Résultats des tests de Mann-Whitney

	Me		*M(ET)*		*U*	*p*	r_{bis}
	Chercheurs	**Cliniciens**	**Chercheurs**	**Cliniciens**			
Compréhension de l'introduction	4	4	3.62 (0.51)	3.57 (0.54)	43.5	.444	.04
Compréhension de la méthodologie	3	3	3.31 (0.63)	3.00 (0.58)	33.5	.148	.26
Compréhension des résultats	3	3	2.85 (0.56)	2.86 (0.69)	45.5	.519	.00
Compréhension de la discussion	3	3	3.15 (0.38)	3.29 (0.49)	39.5	.771	.13
Évaluation de la qualité méthodologique	3	3	3.23 (0.44)	3.00 (0.00)	35	.100	.23
Suffisamment d'informations pour reproduire l'intervention	4	3	3.38 (0.77)	2.43 (0.79)	18	.012*	.60
Suffisamment d'informations pour juger l'efficacité	3	3	2.92 (0.64)	3.14 (0.38)	37	.811	.19
Suffisamment d'informations pour juger l'utilité	3	3	2.69 (0.48)	3.14 (0.69)	29	.945	.36

*Notes. * $p < .05$; Me = Médiane ; M(ET) = Moyenne et écart-type ; U = Statistique U de Mann-Whitney ; r_{bis} = corrélation entre rangs bisériés*

3. Conclusion

Les résultats de cette étude mettent en évidence plusieurs éléments clés concernant les habitudes de lecture, la compréhension et l'évaluation critique d'articles scientifiques par notre échantillon de chercheurs. Tout d'abord, en ce qui concerne les habitudes de lecture, les chercheurs déclarent consulter des articles portant sur des prises en charge à des fréquences variées, la lecture exceptionnelle étant la plus fréquemment rapportée (43 %, $N = 9$). Les critères les plus souvent mobilisés pour évaluer la qualité d'un article sont la clarté et la précision de la méthodologie (62 %, $N = 13$), la taille de l'échantillon (38 %, $N = 8$), l'importance de la validité des analyses statistiques (24 %, $N = 5$) ainsi que celle d'un plan d'étude rigoureux et approprié (19 %, $N = 4$).

Concernant la lecture des ECR, les résultats montrent que le niveau de confiance initial était significativement plus élevé pour l'article *reporting* + (M = 7.77) que pour l'article *reporting* - (M = 4.54), avec une forte taille d'effet. Cette confiance initiale est restée globalement stable après l'analyse des biais méthodologiques, aucune évolution significative n'ayant été observée pour les deux articles. Par ailleurs, plus d'un tiers des participants ont identifié des biais dans l'article *reporting* + (38 %, N = 5), contre plus de deux tiers dans l'article *reporting* - (77 %, N = 10).

Les résultats montrent également que les participants ont attribué à l'article *reporting* + des évaluations significativement plus élevées concernant la compréhension de l'introduction (M = 3.62 vs. 3.08), l'évaluation de la qualité méthodologique (M = 3.23 vs. 2.38) ainsi que la perception de disposer de suffisamment d'informations pour reproduire l'intervention (M = 3.38 vs. 2.46), mais aussi en apprécier l'efficacité (M = 2.92 vs. 2.08) et en juger l'utilité (M = 2.69 vs. 2.00).

En ce qui concerne les critères de qualité méthodologique, les résultats indiquent qu'une majorité d'entre eux ont été pris en compte par au moins 50 % des participants, quel que soit l'article. En revanche, le taux de bonnes évaluations (c'est-à-dire les critères à la fois identifiés et correctement jugés) varie selon les critères.

Enfin, la comparaison entre chercheurs et cliniciens ne révèle aucune différence significative en ce qui concerne la capacité à identifier les biais méthodologiques, la compréhension des différentes sections de l'article ou la perception de disposer de suffisamment d'informations pour apprécier l'efficacité et l'utilité de l'intervention. La seule différence significative concerne la perception de disposer de suffisamment d'informations pour reproduire l'intervention, pour laquelle les chercheurs ont plus souvent estimé disposer de suffisamment d'éléments par rapport aux cliniciens (M = 3.38 vs. 2.43).

Discussion

Pour rappel, cette étude visait à explorer les comportements de chercheurs en neuropsychologie clinique, en psychologie cognitive ou en neurosciences cognitives en matière de lecture critique. Nos objectifs principaux étaient d'analyser leur manière d'évaluer deux ECR de qualité méthodologique contrastée, notamment en termes de confiance accordée aux résultats, d'identification des biais et de critères mobilisés pour juger la qualité. En outre, nous avions également pour objectif de comparer notre échantillon de chercheurs avec un échantillon de neuropsychologues cliniciens.

Ainsi, dans la discussion de ce travail, nous reprenons nos hypothèses initiales et exploratoires, confrontons les résultats attendus à ceux observés dans notre échantillon, puis proposons une interprétation à la lumière des connaissances scientifiques actuelles. Nous mettons également en avant les limites de notre étude, avant de suggérer des perspectives concrètes pour l'avenir.

1. Pratiques de lecture

Étant donné le caractère exploratoire de cette étude, qui s'intéresse notamment aux pratiques de lecture des chercheurs, nous avons souhaité analyser plus en détail les comportements observés au sein de notre échantillon.

Dans un premier temps, nous nous sommes intéressées à la fréquence à laquelle les participants lisaient des ECR portant sur des interventions cliniques. Nos résultats indiquent qu'une majorité des répondants déclarent consulter ce type d'articles de manière peu fréquente dans le cadre de leur pratique professionnelle. Plusieurs pistes explicatives peuvent être avancées pour éclairer cette tendance. Tout d'abord, certains chercheurs, en fonction de leur domaine d'investigation, peuvent ne pas être directement concernés par ce type de recherches. Ensuite, certains peuvent manquer de temps pour s'y consacrer pleinement, préférant alors s'orienter vers des revues systématiques ou des méta-analyses. Tout comme les cliniciens, il est possible que les chercheurs perçoivent la lecture d'articles scientifiques comme particulièrement chronophage et l'investissement nécessaire à leur lecture disproportionnel par rapport au bénéfice informatif (Stewart et al., 2012). Il est également important de noter qu'une proportion non négligeable de notre échantillon de chercheurs exerce parallèlement une activité clinique (33 %, $N = 7$), ce qui pourrait constituer un frein à la lecture régulière d'articles scientifiques. En effet, la charge de

travail associée à la pratique clinique est souvent rapportée comme un obstacle majeur à l'intégration de la littérature scientifique dans la routine professionnelle (Blause et al., 2023a ; Durieux et al., 2017 ; Nelson et al., 2006).

Ensuite, nous avons vérifié l'hypothèse selon laquelle les chercheurs allaient lire les articles scientifiques dans leur intégralité en raison de leur formation scientifique. Les résultats obtenus vont dans le sens de cette hypothèse puisque la majorité des participants indique avoir lu les deux articles en entier et non de manière partielle. Cette préférence pour une lecture intégrale peut s'expliquer par leur familiarité avec la littérature scientifique ainsi que par leurs compétences méthodologiques et statistiques. Des travaux antérieurs montrent que les cliniciens ne lisent que partiellement les articles, car ils se sentent souvent mal à l'aise face au vocabulaire spécialisé ou à la complexité des méthodes utilisées, en particulier sur le plan statistique (Baker et al., 2008 ; Blause et al., 2023a). Ainsi, à la différence des cliniciens, les chercheurs semblent accorder une attention plus soutenue aux sections portant sur la méthodologie et les résultats. La majorité des chercheurs de notre échantillon démontre ainsi une aptitude accrue à appréhender ces sections techniques, ce qui leur permet de s'engager dans une lecture approfondie et analytique de leurs contenus. Il convient toutefois de souligner que leurs déclarations pourraient avoir été influencées par un biais de désirabilité sociale, certains participants ayant pu surestimer leur implication réelle afin de se conformer à ce qu'ils percevaient comme étant attendu.

En ce qui concerne les outils d'analyse critique, notre étude s'est intéressée à leur utilisation éventuelle par les chercheurs et il s'avère qu'aucun des participants ne rapporte y avoir eu recours lors de la lecture des deux ECR. Cela peut s'expliquer, une fois de plus, par un manque de temps de leur part. Il est possible que, pour certains, l'investissement nécessaire à l'emploi de tels outils soit perçu comme coûteux en temps et en effort, pour un bénéfice jugé limité. Ce ressenti pourrait d'ailleurs s'accentuer lorsque l'article lu s'éloigne de leur domaine de spécialité – ce qui est le cas pour un peu moins de la moitié de l'échantillon. En outre, étant donné que peu de participants déclarent lire régulièrement des ECR dans leur pratique quotidienne, on peut supposer que leur niveau de familiarité avec ces outils d'évaluation critique demeure restreint. Il convient toutefois de noter que certains participants ont pu décider de ne pas utiliser ce type d'outils pendant l'étude, bien qu'ils les emploient dans leur pratique courante. Dès lors, bien que nous puissions formuler l'hypothèse que des contraintes de temps ou un manque de familiarité puissent freiner l'usage de ces outils, les obstacles à leur intégration dans la pratique professionnelle méritent d'être explorés plus en profondeur dans le cadre de

recherches ultérieures. En l'absence de recours explicite à ces outils dans notre étude, l'évaluation de la qualité méthodologique semble ainsi reposer avant tout sur une lecture critique intuitive et spontanée des articles, point que nous développons dans la section suivante.

2. Influence de la qualité méthodologique

Pour rappel, nous avions formulé l'hypothèse selon laquelle l'article *reporting* + serait perçu par les chercheurs comme supérieur à l'article *reporting* - sur plusieurs dimensions, notamment la qualité méthodologique et la suffisance d'informations pour reproduire l'intervention, en apprécier l'efficacité et en juger l'utilité. Nous anticipions également que cet article susciterait un niveau de confiance initial plus élevé ainsi qu'une meilleure compréhension des différentes sections de l'article. Par ailleurs, nous supposions que ce jugement de confiance initial pourrait évoluer à la suite de la complétion du questionnaire.

Les résultats obtenus vont globalement dans le sens de notre hypothèse. En effet, les chercheurs perçoivent l'article *reporting* + comme significativement supérieur sur le plan de la qualité méthodologique ainsi que sur celui de la suffisance d'informations pour reproduire l'intervention, en apprécier l'efficacité et en juger l'utilité. Concernant la compréhension des différentes sections de l'article, seule la section *« Introduction »* a obtenu des scores significativement plus élevés pour l'article *reporting* +. Nos résultats tendent ainsi à confirmer l'idée que les chercheurs sont sensibles à la structuration et à la clarté méthodologique des articles qu'ils lisent. Cette observation rejoint les données de la littérature qui indiquent que la qualité du *reporting* devrait faciliter l'évaluation critique des ECR (Moher et al., 2010). Toutefois, nos résultats ne permettent pas de corroborer l'hypothèse selon laquelle un *reporting* de meilleure qualité favoriserait également la compréhension de l'ensemble des sections de l'article, comme Moher et ses collègues (2010) le suggèrent.

Concernant le niveau de confiance initial, les résultats montrent que l'article *reporting* + suscite un degré de confiance significativement plus élevé que l'article *reporting* -, ce qui vient à nouveau soutenir notre hypothèse. En revanche, aucun changement significatif du niveau de confiance n'est observé entre les deux temps de mesure, que ce soit pour l'article *reporting* + ou *reporting* -. Ces résultats suggèrent que les chercheurs mobilisent spontanément des critères implicites leur permettant de porter un jugement relativement rapide sur la rigueur méthodologique d'un article, même en l'absence d'un outil d'analyse critique. Toutefois, cette

compétence intuitive ne semble pas garantir une identification précise et approfondie des biais méthodologiques, comme nous allons le voir dans les sections suivantes.

3. Identification des biais méthodologiques

Pour rappel, nous avions pour hypothèse que les participants identifieraient davantage de biais méthodologiques dans l'article *reporting* -, celui-ci présentant un risque de biais élevé tel qu'objectivé par la grille RoB2. Sur le plan descriptif, nos résultats vont dans le sens de cette hypothèse : un nombre plus élevé de participants rapporte avoir effectivement identifié des biais dans l'article *reporting* -, comparativement à l'article *reporting* +. L'analyse des réponses montre qu'une majorité des biais évoqués pour l'article *reporting* - peuvent être reliés aux recommandations de la déclaration CONSORT-SPI, suggérant une certaine capacité des chercheurs à détecter les lacunes méthodologiques sans outil d'analyse critique. Ce constat pourrait également témoigner d'une certaine familiarité des participants avec ces recommandations, bien que la présente étude n'ait pas exploré leur connaissance explicite de la grille CONSORT. Il est donc envisageable que certains participants aient mobilisé, de manière implicite, des critères issus de cette grille au moment de formuler leurs réponses. Dans la mesure où aucun article scientifique n'est exempt de biais (Simon, 2001), il convient de noter que les participants auraient aussi pu identifier des biais dans l'article *reporting* +. Quant aux participants n'ayant identifié aucun biais dans les deux articles, on peut y voir une forme d'incertitude quant à l'identification des biais, voire un manque de temps ou d'investissement lors de la réponse à la question ouverte. En effet, certains chercheurs, en particulier ceux lisant rarement des ECR cliniques, peuvent éprouver des difficultés à repérer ou interpréter les biais méthodologiques, bien qu'ils soient conscients de leur existence – un phénomène déjà mis en évidence chez les cliniciens (Blause et al., 2023a).

Par conséquent, nous avons voulu examiner de façon exploratoire l'éventualité d'un lien entre la fréquence de lecture d'ECR cliniques et la capacité à identifier des biais méthodologiques. Notre hypothèse exploratoire était que des lectures plus régulières pourraient favoriser la détection des biais. Toutefois, les résultats obtenus ne permettent pas de confirmer cette hypothèse. En effet, certains participants déclarant ne jamais lire ce type d'articles ont été en mesure d'identifier plusieurs biais dans l'un ou les deux articles, tandis qu'un participant affirmant en lire une fois par semaine ou plus n'en a relevé aucun. Ces résultats appellent à la prudence et suggèrent qu'il serait nécessaire de disposer de données supplémentaires pour dégager des tendances plus solides.

En outre, une autre analyse exploratoire visait à déterminer si l'identification de biais dans l'article *reporting* - pouvait constituer un prédicteur de la capacité à reconnaître le meilleur article en termes de qualité méthodologique. Notre hypothèse exploratoire postulait que plus les participants identifiaient de biais dans l'article *reporting* -, plus ils étaient susceptibles de juger correctement de la supériorité méthodologique de l'article *reporting* +. Cette hypothèse est partiellement confirmée sur base d'analyses descriptives : parmi les 11 participants ayant correctement identifié le meilleur article, seule une minorité (27 %, $N = 3$) n'avait relevé aucun biais dans l'article *reporting* -.

4. Critères de qualité pris en compte lors de la lecture

Dans la continuité de l'étude de Faulkner et ses collaborateurs (2008), nous avions formulé l'hypothèse que les éléments statistiques tels que la *p*-valeur, la taille d'effet et l'intervalle de confiance seraient des critères pris en compte par les chercheurs pour évaluer la qualité méthodologique d'un ECR. Nos résultats confirment en partie cette hypothèse étant donné que la *p*-valeur et la taille d'effet ont largement été prises en compte comme critères de qualité méthodologique. En effet, plus de la moitié des participants en ont tenu compte dans leur lecture critique des deux articles. Par contre, moins de la moitié des participants a pris en considération la présence d'un intervalle de confiance autour de la taille d'effet dans les deux articles. Or, dans l'étude de Faulkner et ses collègues (2008), une grande majorité des répondants considère comme essentielles les informations statistiques relatives à l'existence d'un effet réel, à sa taille et sa précision, ainsi qu'à sa pertinence clinique. À cet égard, nous aurions pu nous attendre à ce que, dans notre échantillon, l'intervalle de confiance soit pris en considération au même titre que la *p*-valeur ou la taille d'effet. Une hypothèse explicative plausible réside dans l'absence explicite de ces intervalles de confiance dans les articles proposés, ce qui a pu conduire à un oubli ou à une omission involontaire de la part des participants.

Par ailleurs, selon une méta-recherche récente, Blause et ses collègues (2025) montrent que, dans un échantillon d'ECR en neuropsychologie, plusieurs éléments de la grille CONSORT-SPI étaient rarement rapportés. Parmi ces éléments figurent notamment les objectifs spécifiques et les hypothèses précises de l'étude, la description détaillée de l'intervention afin d'en permettre la réplication, la définition des *outcomes*, le processus de *blinding*, les résultats précis pour chaque *outcome* ainsi que l'interprétation clinique des résultats. Pourtant, nos résultats montrent que plus de la moitié des participants ont tenu compte, lors de la lecture des deux articles, de la plupart de ces critères. Cela soulève une question importante : comment expliquer

que ces éléments, bien qu'identifiés comme essentiels par les lecteurs, soient si peu présents dans les publications ? Ces observations pourraient indiquer que, bien que les auteurs reconnaissent l'importance de ces critères pour l'évaluation de la qualité d'un article scientifique, ils éprouvent des difficultés à les appliquer dans leurs propres publications, potentiellement en raison d'un manque de directives méthodologiques explicites. C'est précisément pour répondre à ces difficultés que la grille CONSORT a été élaborée (Begg et al., 1996). Celle-ci vise à fournir un cadre structuré et standardisé permettant aux auteurs de rapporter les éléments méthodologiques essentiels de manière claire, transparente et exhaustive. En facilitant la rédaction et la lecture critique des ECR, la déclaration CONSORT a pour objectif de renforcer la qualité des publications scientifiques et d'en améliorer la reproductibilité (Moher et al., 2010). Ainsi, l'écart observé entre l'importance accordée à certains critères par les lecteurs et leur sous-représentation dans les publications souligne la nécessité de promouvoir l'usage effectif de ces recommandations au sein de la communauté scientifique.

Enfin, en ce qui concerne l'analyse du pourcentage de bonnes réponses pour chaque critère méthodologique pris en compte par les participants, nos résultats révèlent que plusieurs critères ont été jugés satisfaisants ou insatisfaisants à tort. Ainsi, comme précédemment mentionné, la compétence intuitive des chercheurs ne semble pas garantir une identification précise et approfondie des biais méthodologiques. Ces résultats soulignent l'intérêt d'un recours plus systématique aux outils d'évaluation critique. Par ailleurs, de tels outils pourraient non seulement renforcer la rigueur de la lecture critique des chercheurs, mais aussi leur permettre d'anticiper certaines erreurs méthodologiques ou d'améliorer la qualité de leur *reporting*, comme l'ont récemment souligné Blause et ses collègues (2025).

5. Comparaison entre les chercheurs et les cliniciens

Dans le cadre de ce mémoire en binôme, une comparaison a été effectuée entre les chercheurs et les cliniciens. L'hypothèse posée était que, du fait de leur formation et de leur expertise méthodologique plus poussées, les chercheurs évalueraient plus favorablement que les cliniciens l'article *reporting* + sur le plan de la qualité méthodologique ainsi que sur celui de la suffisance d'informations pour reproduire l'intervention, en apprécier l'efficacité et en juger l'utilité. Nous nous attendions également à ce qu'ils rapportent une meilleure compréhension des différentes sections de l'article (introduction, méthodologie, résultats et discussion) et identifient un plus grand nombre de biais méthodologiques que les cliniciens.

Les résultats obtenus ne permettent pas de confirmer clairement notre hypothèse. Aucune différence statistiquement significative n'a été observée entre les deux groupes, que ce soit concernant l'identification des biais méthodologiques, la compréhension des différentes sections de l'article ou l'évaluation de sa qualité méthodologique. La seule différence significative portait sur la perception d'avoir suffisamment d'informations pour reproduire l'intervention, en faveur des chercheurs. Cette observation est en accord avec les résultats de Blause et ses collègues (2023a), qui montrent que seuls 6 % des 350 neuropsychologues cliniciens interrogés considèrent que les informations contenues dans les articles scientifiques sont directement applicables à leur pratique clinique.

Ces résultats peuvent donner lieu à deux interprétations complémentaires. Tout d'abord, même si l'article *reporting* + utilisé dans notre étude a un score de 17/45 à la grille CONSORT-SPI – score supposé refléter un niveau de transparence suffisant pour juger l'étude réplicable –, il est possible que les cliniciens aient néanmoins éprouvé des difficultés à en extraire des éléments directement exploitables dans leur pratique. Cette difficulté pourrait s'expliquer par des lacunes persistantes dans la formation à l'EBP (Blause et al., 2023a ; Blause et al., 2024 ; Chelune, 2010 ; Pagoto et al., 2007). Ensuite, il est envisageable que la transparence méthodologique de l'article demeure insuffisante pour permettre une mise en œuvre concrète en contexte clinique. Certains articles, bien que théoriquement réplicables, pourraient avant tout être rédigés à l'intention d'un public de chercheurs, au détriment de leur accessibilité pour les professionnels de terrain. Comme le soulignent Wilson et ses collègues (2009), un écart persiste entre les attentes des chercheurs et celles des praticiens, mettant en lumière la nécessité de mieux comprendre leurs perspectives respectives afin de favoriser un rapprochement entre recherche et pratique.

En résumé, alors que les chercheurs jugent l'étude comme étant globalement reproductible dans un contexte de recherche, les cliniciens ne partagent pas cet avis en ce qui concerne son applicabilité en pratique clinique. Cette divergence de perception pourrait refléter un écart persistant entre le monde de la recherche et celui de la clinique. Plusieurs obstacles peuvent en effet freiner l'intégration des données issues de la recherche dans la pratique quotidienne des cliniciens, notamment un manque de formation à l'EBP et/ou une qualité de *reporting* des ECR encore insuffisante pour permettre une transposition concrète des résultats à la réalité du terrain.

6. Limites

Plusieurs limites doivent être prises en compte dans l'interprétation des résultats de cette étude. Tout d'abord, bien que notre objectif initial était de recruter un maximum de participants, nous n'avons pas anticipé l'impact de la durée de passation sur l'engagement des participants. En effet, la longueur de l'étude a freiné la participation ; certains professionnels ont ainsi choisi de ne pas débuter l'étude, tandis que d'autres l'ont interrompue avant la fin. Par ailleurs, il est possible que les participants ayant complété l'intégralité de l'étude aient éprouvé une certaine fatigue cognitive, en particulier en fin de passation. Cette fatigue pourrait avoir entraîné une baisse d'attention ou un traitement plus superficiel des informations, affectant potentiellement la qualité de leurs réponses. Ce phénomène pourrait ainsi avoir biaisé certains résultats, en particulier dans les sections nécessitant une réflexion approfondie, comme l'analyse critique des articles.

Ensuite, la présentation des articles a pu influencer la manière dont les participants les ont analysés. En effet, le protocole expérimental reposait sur la lecture en ligne de deux articles présentés sous forme d'images, format qui s'éloigne des conditions habituelles de lecture d'articles scientifiques en contexte professionnel. Les participants ne pouvaient ni imprimer ni annoter les documents, ce qui constitue une contrainte notable, dans la mesure où ces stratégies peuvent être mobilisées pour permettre une lecture approfondie et favoriser la compréhension. Cette limitation découlait de contraintes logistiques, en particulier de la difficulté à organiser une passation en présentiel.

Par ailleurs, le recours à un dispositif en ligne ne nous a pas permis de contrôler précisément les conditions dans lesquelles les participants ont répondu. Il demeure donc possible que certains aient consulté les versions originales des articles sur un moteur de recherche sans le signaler, ou qu'ils aient effectué des captures d'écran afin de pouvoir relire les documents lors de la complétion du questionnaire. De telles pratiques, bien que compréhensibles, pourraient avoir introduit un biais dans nos résultats.

En ce qui concerne les articles utilisés, deux limites méritent d'être soulignées. Premièrement, ces articles ont été entièrement anonymisés : ni le nom des auteurs ni celui de la revue n'étaient mentionnés. Cette décision visait à prévenir toute reconnaissance de la part des participants, mais elle a pu, en contrepartie, restreindre leur capacité à évaluer la qualité de l'article, ne pouvant s'appuyer sur des éléments contextuels tels que la réputation scientifique de l'auteur

ou la politique éditoriale de la revue. Deuxièmement, l'évaluation de l'exhaustivité du *reporting* a été réalisée à l'aide de la grille CONSORT-SPI, bien que celle-ci ne soit pas spécifiquement conçue pour cet usage. Par ailleurs, l'attribution de scores binaires (0 ou 1) à certains *items*, bien que nécessaire pour l'analyse, a pu introduire une certaine rigidité, notamment lorsque les informations étaient rapportées de manière partielle.

Enfin, une dernière limite à souligner concerne l'absence de calcul préalable d'une taille d'échantillon. Faute de données issues de recherches antérieures sur ce sujet précis, il n'a pas été possible d'estimer une taille d'effet attendue, condition pourtant essentielle à la réalisation d'une analyse de puissance. Cette contrainte méthodologique restreint la généralisabilité de nos observations et doit inciter à la prudence dans l'interprétation des résultats.

7. Perspectives

Bien que notre étude présente certaines limites, les résultats obtenus offrent des pistes de réflexion intéressantes, tant sur le plan théorique que pratique. D'un point de vue théorique, ils contribuent à enrichir la compréhension des pratiques de lecture critique au sein d'un public de chercheurs. Alors que la plupart des études précédentes s'étaient centrées sur les cliniciens, en particulier via des questionnaires auto-rapportés (Blause et al., 2023a ; Blause et al., 2024), notre approche expérimentale permet d'apporter un éclairage plus direct sur la manière dont les chercheurs analysent des ECR en situation réelle de lecture.

Sur le plan pratique, ces résultats soulignent la nécessité de repenser les modalités de formation à la lecture critique, que ce soit en formation initiale ou continue. En effet, nos résultats indiquent que certains chercheurs n'ont identifié aucun biais, même dans l'article présentant un risque élevé. Cela pourrait indiquer qu'une partie d'entre eux ne maîtrise pas pleinement les outils et repères nécessaires à une évaluation rigoureuse. Or, si les chercheurs eux-mêmes peinent à identifier les biais méthodologiques, il devient difficile d'envisager une production scientifique exempte de telles lacunes. Ce constat plaide ainsi en faveur d'un renforcement des compétences critiques, y compris dans les parcours de formation à la recherche. Une formation systématique à l'utilisation d'outils d'aide au *reporting* ou d'analyse critique, tels que la déclaration CONSORT-SPI et la grille RoB2, devrait être encouragée. Ces outils, bien que disponibles, semblent encore peu utilisés de manière spontanée par les chercheurs. Ce constat est corroboré par l'absence d'outils mobilisés dans notre étude, ce qui rejoint les observations de l'étude de Blause et ses collaborateurs (2023b).

Pour encourager l'adoption de la grille CONSORT par les chercheurs, il serait pertinent que davantage de revues scientifiques rendent son utilisation obligatoire, à l'instar de certaines qui le pratiquent déjà (Shamseer et al., 2016). En outre, il est possible d'envisager que, dans le contexte technologique actuel, les chercheurs puissent tirer parti des avancées en apprentissage automatique (*machine learning*) pour automatiser l'évaluation de leurs articles selon des grilles telles que CONSORT, à l'image des travaux menés dans une méta-recherche récente (Kilicoglu et al., 2023). Un tel outil pourrait ainsi non seulement compenser le manque de temps dont disposent les auteurs, mais également les inciter à améliorer la qualité de leur *reporting* en vue d'obtenir des scores élevés calculés automatiquement. Il demeure néanmoins essentiel de considérer ce type de dispositif comme un outil d'accompagnement, sans pour autant qu'il se substitue à l'indispensable regard critique de l'être humain, qui reste central dans l'évaluation méthodologique.

De surcroît, dans l'objectif de réduire le fossé entre la recherche et la clinique, la mise en place d'espaces partagés de formation et d'échanges entre chercheurs et cliniciens pourrait s'avérer particulièrement bénéfique, notamment en ce qui concerne les pratiques de lecture critique. Il serait également pertinent de favoriser une collaboration accrue entre chercheurs, permettant de mutualiser leurs compétences et de renforcer la rigueur des approches méthodologiques. Une telle dynamique pourrait également favoriser la création de ressources pédagogiques, notamment des modules de formation en ligne axés sur des thématiques méthodologiques spécifiques, comme suggéré par Ioannidis et ses collègues (2015).

Quant aux perspectives de recherches futures, il serait tout d'abord pertinent de reproduire ce protocole auprès d'un échantillon plus large et plus diversifié, incluant notamment davantage de chercheurs issus d'autres pays francophones. Par ailleurs, le recours à une approche qualitative complémentaire (entretiens semi-directifs, méthode du *« think aloud »*, etc.) pourrait permettre d'explorer plus finement les pratiques de lecture et les processus d'analyse critique mobilisés. Enfin, l'évaluation de l'impact d'une formation courte à l'analyse critique outillée, auprès des chercheurs comme des cliniciens, représenterait une piste prometteuse pour mesurer l'effet de ce type d'intervention sur les performances d'analyse méthodologique.

Conclusion

L'objectif de cette étude consistait à analyser les pratiques de lecture critique adoptées par les chercheurs en neuropsychologie clinique, en psychologie cognitive et en neurosciences cognitives, en accordant une attention particulière à leur compétence d'analyse critique de deux ECR cliniques. À travers une approche expérimentale combinant un questionnaire et une tâche de lecture critique, nous avons recueilli des données sur leurs habitudes de lecture, leur niveau de confiance dans les résultats publiés, leur capacité à détecter les biais, leur compréhension et leur perception de ces articles ainsi que les critères qu'ils mobilisent pour juger de leur qualité méthodologique. Nos hypothèses initiales postulaient que les chercheurs adopteraient une lecture intégrale des articles, qu'ils considéreraient l'article *reporting* + comme supérieur sur plusieurs dimensions et détecteraient davantage de biais méthodologiques dans l'article *reporting* -, qu'ils mobiliseraient des critères statistiques dans leur évaluation critique et qu'ils surpasseraient les cliniciens en termes d'analyse méthodologique.

Les résultats obtenus permettent de confirmer une partie de nos hypothèses. Tout d'abord, une majorité de participants déclarent lire les articles dans leur intégralité, ce qui vient appuyer notre première hypothèse. En revanche, nos résultats montrent que la lecture d'ECR cliniques reste marginale au sein de notre échantillon et qu'aucun participant n'a rapporté recourir à un outil d'analyse critique. Concernant l'influence de la qualité méthodologique, les données recueillies confirment partiellement notre seconde hypothèse. En effet, l'article *reporting* + a été perçu comme supérieur à l'article *reporting* -, tant sur le plan de la qualité méthodologique que sur celui de la suffisance d'informations pour reproduire l'intervention, en apprécier l'efficacité et en juger l'utilité. Il a également suscité un niveau de confiance initial significativement plus élevé. De plus, conformément à nos attentes, une majorité de participants ont identifié davantage de biais dans l'article *reporting* - que dans l'article *reporting* +. En ce qui concerne les critères mobilisés pour évaluer la qualité méthodologique, une majorité d'entre eux ont effectivement été pris en compte par les participants. Toutefois, les éléments statistiques tels que la *p*-valeur et la taille d'effet ont été fréquemment considérés, contrairement à l'intervalle de confiance, ce qui ne permet donc pas de valider pleinement notre troisième hypothèse. En outre, il ressort de nos résultats que certains critères ont été considérés à tort comme satisfaisants, tandis que d'autres, à l'inverse, ont été jugés insatisfaisants alors qu'ils ne l'étaient pas. Enfin, la comparaison avec les cliniciens révèle peu de différences significatives entre les

deux groupes, hormis un sentiment plus affirmé chez les chercheurs d'être en mesure de reproduire l'intervention. Ces résultats ne permettent donc pas de soutenir pleinement notre dernière hypothèse.

Pour terminer, bien que les limites de cette étude appellent à la prudence quant à la généralisation des résultats, celle-ci ouvre des pistes concrètes pour améliorer les pratiques actuelles des chercheurs et des cliniciens. Une formation plus systématique à l'évaluation critique et aux outils comme la grille RoB2 ou la grille CONSORT-SPI, ainsi que la création d'espaces de dialogue entre professionnels, pourraient contribuer à renforcer l'articulation entre la recherche et la pratique. Le recours à des outils automatisés pourrait également constituer un levier facilitateur, à condition qu'il vienne en appui – et non en remplacement – de l'expertise humaine.

En définitive, ce travail souligne que la rigueur méthodologique et la lecture critique sont des compétences qui doivent être cultivées et partagées par l'ensemble des acteurs du domaine de la neuropsychologie. Pour que le pilier *recherche* de l'EBP puisse pleinement contribuer à l'amélioration des soins, il est essentiel que chercheurs et cliniciens unissent leurs efforts pour élever les standards de qualité, tant dans la production que dans l'utilisation des preuves scientifiques. C'est à cette condition que la neuropsychologie pourra continuer à progresser, en s'appuyant sur des bases méthodologiques solides et une pratique clinique éclairée.

Ressources bibliographiques

Aarons, G. A. (2004). Mental health provider attitudes toward adoption of evidence-based practice: The evidence-based practice attitude scale (EBPAS). *Mental Health Services Research, 6*(2), 61–74. https://doi.org/10.1023/b:mhsr.0000024351.12294.65.

Agoritsas, T., Vandvik, P., Neumann, I., Rochwerg, B., Jaeschke, R., Hayward, R., et al. (2015). Finding current best evidence. In Guyatt, G., Rennie, D., Meade, M. O., & Cook, D. J. (Eds.), *Users' guides to the medical literature: A manual for evidence-based clinical practice* (3rd ed.). McGraw Hill, Chicago.

APA Presidential Task Force on Evidence-Based Practice (2006). Evidence-based practice in psychology. *The American Psychologist, 61*(4), 271–285. https://doi.org/10.1037/0003-066X.61.4.271.

Atzeni, T., & Follenfant, A. (2013). Évolution des pratiques en neuropsychologie clinique : vers une pratique basée sur la preuve ? *Revue de Neuropsychologie, Neurosciences Cognitives et Cliniques, 5*(1), 28-37. https://doi.org/10.1684/nrp.2013.0249.

Babione, J. M. (2010). Evidence-Based Practice in Psychology: An Ethical Framework for Graduate Education, Clinical Training, and Maintaining Professional Competence. *Ethics & Behavior, 20*(6), 443-453. https://doi.org/10.1080/10508422.2010.521446.

Baker, T. B., McFall, R. M., & Shoham, V. (2008). Current Status and Future Prospects of Clinical Psychology. *Psychological Science In The Public Interest*, *9*(2), 67-103. https://doi.org/10.1111/j.1539-6053.2009.01036.x.

Barker, T. H., Stone, J. C., Sears, K., Klugar, M., Tufanaru, C., Leonardi-Bee, J., Aromataris, E., & Munn, Z. (2023). The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. *JBI Evidence Synthesis*. https://doi.org/10.11124/jbies-22-00430.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., et al. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *JAMA, 276*(8), 637–639. https://doi.org/10.1001/jama.1996.03540080059030.

Bhide, A., Shah, P. S., & Acharya, G. (2018). A simplified guide to randomized controlled trials. *Acta Obstetricia et Gynecologica Scandinavica*, *97*(4), 380-387. https://doi.org/10.1111/aogs.13309.

Blause, S., Durieux, N., Tirelli, E., & Willems, S. (2024). L'evidence-based practice en psychologie : une approche tripartite souvent mal comprise. *Pratiques Psychologiques*. https://doi.org/10.1016/j.prps.2024.03.005.

Blause, S., Léonard, F., Tirelli, E., & Willems, S. (2025). Can research findings be used in clinical neuropsychology? Analysis of randomized controlled trials of memory intervention for children. *Archives of Clinical of Neuropsychology*. https://doi.org/10.1093/arclin/acaf048.

Blause, S., Tirelli, E., Wauquiez, G., Raffard, S., Didone, V., & Willems, S. (2023a). What Information Do Neuropsychologists Use to Guide their Clinical Decisions? A Survey on Knowledge and Application of Evidence-Based Practice in a French-Speaking Population. *Archives Of Clinical Neuropsychology*. https://doi.org/10.1093/arclin/acad057.

Blause, S., Tirelli, E., & Willems, S. (2023b, December 1). Les résultats de la recherche clinique en neuropsychologie sont-ils utilisables ? Une analyse méta-scientifique [Poster presentation]. Journée d'Hiver, Paris, France.

Chelune, G. J. (2010). Evidence-based research and practice in clinical neuropsychology. *The Clinical Neuropsychologist,* *24*(3), 454–467. https://doi.org/10.1080/13854040802360574.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2e édition). New York, New York : Academic Press.

Dovis, S., Van Der Oord, S., Wiers, R. W., & Prins, P. J. M. (2015). Improving Executive Functioning in Children with ADHD : Training Multiple Executive Functions within the Context of a Computer Game. A Randomized Double-Blind Placebo Controlled Trial. *PLoS ONE, 10*(4), Article e0121651. https://doi.org/10.1371/journal.pone.0121651.

Durieux, N., Étienne, A.-M., & Willems, S. (2017). Introduction à l'evidence-based practice en psychologie. *Le Journal des Psychologues, 345*, 16-20. https://doi.org/10.3917/jdp.345.0016.

Durieux, N., Pasleau, F., Vandenput, P., Detroz, P., & Maillart, C. (2012). L'Evidence-Based Practice et les logopèdes en Communauté française de Belgique: Résultats préliminaires d'une enquête. *Cahiers de l'ASELF, 9*(4), 30-35.

Elbe, P., Bäcklund, C., Vega-Mendoza, M., Sörman, D., Gavelin, H. M., Nyberg, L., & Ljungberg, J. K. (2023). Computerized cognitive interventions for adults with ADHD : A systematic review and meta-analysis. *Neuropsychology*, *37*(5), 519-530. https://doi.org/10.1037/neu0000890.

Emwodew, D., Melese, T., Takele, A., Mesfin, N., & Tariku, B. (2021). Knowledge and Attitude Toward Evidence-Based Medicine and Associated Factors Among Medical Interns in Amhara Regional State Teaching Hospitals, Northwest Ethiopia : Cross-sectional Study. *JMIR Medical Education*, *7*(2), e28739. https://doi.org/10.2196/28739.

Falzon, L., Davidson, K. W., & Bruns, D. (2010). Evidence searching for evidence-based psychology practice. *Professional Psychology : Research And Practice, 41*(6), 550-557. https://doi.org/10.1037/a0021352.

Faulkner, C., Fidler, F., & Cumming, G. (2008). The value of RCT evidence depends on the quality of statistical analysis. *Behaviour Research And Therapy, 46*(2), 270-281. https://doi.org/10.1016/j.brat.2007.12.001.

Gates, N. J., & March, E. G. (2016). A Neuropsychologist's Guide To Undertaking a Systematic Review for Publication : Making the most of PRISMA Guidelines. *Neuropsychology Review*, *26*(2), 109-120. https://doi.org/10.1007/s11065-016-9318-0.

Grant, S., Mayo-Wilson, E., Montgomery, P., Macdonald, G., Michie, S., Hopewell, S., & Moher, D. (2018). CONSORT-SPI 2018 Explanation and Elaboration : guidance for reporting social and psychological intervention trials. *Trials, 19*(1). https://doi.org/10.1186/s13063-018-2735-z.

Gyani, A., Shafran, R., Myles, P., & Rose, S. (2014). The gap between science and practice: How therapists make their clinical decisions. *Behavior Therapy, 45*(2), 199–211. https://doi.org/10.1016/j.beth.2013.10.004.

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2021). Estimating the Prevalence of Transparency and Reproducibility-Related Research Practices in Psychology (2014–2017). *Perspectives On Psychological Science*, *17*(1), 239-251. https://doi.org/10.1177/1745691620979806.

Hardwicke, T. E., & Wagenmakers, E. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour, 7*(1), 15-26. https://doi.org/10.1038/s41562-022-01497-2.

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials – the gold standard for effectiveness research. *BJOG, 125*(13), 1716. https://doi.org/10.1111/1471-0528.15199.

Hopewell, S., Dutton, S., Yu, L., Chan, A., & Altman, D. G. (2010). The quality of reports of randomised trials in 2000 and 2006 : comparative study of articles indexed in PubMed. *BMJ*, *340*(mar23 1), c723. https://doi.org/10.1136/bmj.c723.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124.

Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research : Evaluation and Improvement of Research Methods and Practices. *PLoS Biology*, *13*(10), e1002264. https://doi.org/10.1371/journal.pbio.1002264.

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences : detection, prevalence, and prevention. *Trends In Cognitive Sciences*, *18*(5), 235-241. https://doi.org/10.1016/j.tics.2014.02.010.

Jin, Y., Sanger, N., Shams, I., Luo, C., Shahid, H., Li, G., Bhatt, M., Zielinski, L., Bantoto, B., Wang, M., Abbade, L. P., Nwosu, I., Leenus, A., Mbuagbaw, L., Maaz, M., Chang, Y., Sun, G., Levine, M. A., Adachi, J. D.,. . . Samaan, Z. (2018). Does the medical literature remain inadequately described despite having reporting guidelines for 21 years ? & ; ndash ; A systematic review of reviews : an update. *Journal Of Multidisciplinary Healthcare*, *Volume 11*, 495-510. https://doi.org/10.2147/jmdh.s155103.

Kilicoglu, H., Jiang, L., Hoang, L., Mayo-Wilson, E., Vinkers, C. H., & Otte, W. M. (2023). Methodology reporting improved over time in 176,469 randomized controlled trials. *Journal Of Clinical Epidemiology*, *162*, 19-28. https://doi.org/10.1016/j.jclinepi.2023.08.004.

Kray, J., Karbach, J., Haenig, S., & Freitag, C. (2012). Can Task-Switching Training Enhance Executive Control Functioning in Children with Attention Deficit/-Hyperactivity Disorder ? *Frontiers In Human Neuroscience, 5*. https://doi.org/10.3389/fnhum.2011.00180.

Leichsenring, F., Abbass, A., Hilsenroth, M. J., Leweke, F., Luyten, P., Keefe, J. R., Midgley, N., Rabung, S., Salzer, S., & Steinert, C. (2016). Biases in research : risk factors for non-replicability in psychotherapy and pharmacotherapy research. *Psychological Medicine*, *47*(6), 1000-1011. https://doi.org/10.1017/s003329171600324x.

Li, H., Xu, R., Gao, D., Fu, H., Yang, Q., Chen, X., Hou, C., & Gao, J. (2024). Evidence-based practice attitudes, knowledge and skills of nursing students and nurses, a systematic review and meta-analysis. *Nurse Education In Practice, 78*, 104024. https://doi.org/10.1016/j.nepr.2024.104024.

Lilienfeld, S. O. (2007). Psychological Treatments That Cause Harm. *Perspectives on Psychological Science, 2*(1), 53-70. https://doi.org/10.1111/j.1745-6916.2007.00029.x.

Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2014). Why Ineffective Psychotherapies Appear to Work. *Perspectives On Psychological Science*, *9*(4), 355-387. https://doi.org/10.1177/1745691614535216.

Luebbe, A. M., Radcliffe, A. M., Callands, T. A., Green, D., & Thorn, B. E. (2007). Evidence-based practice in psychology : Perceptions of graduate students in scientist–practitioner programs. *Journal Of Clinical Psychology, 63*(7), 643-655. https://doi.org/10.1002/jclp.20379.

Mahmoud, M. H., & Abdelrasol, Z. F. M. (2019). Obstacles in employing evidence-based practice by nurses in their clinical settings : a descriptive study. *Frontiers Of Nursing*, *6*(2), 123-133. https://doi.org/10.2478/fon-2019-0019.

Melchert, T. P., Halfond, R., Hamdi, N. R., Bufka, L. F., Hollon, S. D., & Cuttler, M. J. (2023). Evidence-based practice in psychology: Context, guidelines, and action. *American Psychologist*. https://doi.org/10.1037/amp0001253.

Miller, D. J., Spengler, E. S., & Spengler, P. M. (2015). A meta-analysis of confidence and judgment accuracy in clinical decision making. *Journal Of Counseling Psychology, 62*(4), 553-567. https://doi.org/10.1037/cou0000105.

Miskowiak, K. W., Carvalho, A. F., Vieta, E., & Kessing, L. V. (2016). Cognitive enhancement treatments for bipolar disorder : A systematic review and methodological recommendations. *European Neuropsychopharmacology*, *26*(10), 1541-1561. https://doi.org/10.1016/j.euroneuro.2016.08.011.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P., Elbourne, D., Egger, M., & Altman, D. G. (2010). CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *Journal Of Clinical Epidemiology, 63*(8), e1-e37. https://doi.org/10.1016/j.jclinepi.2010.03.004.

Moher, D., Jadad, A. R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. (1995). Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials, 16*(1), 62-73. https://doi.org/10.1016/0197-2456(94)00031-w.

Montgomery, P., Grant, S., Mayo-Wilson, E., Macdonald, G., Michie, S., Hopewell, S., & Moher, D. (2018). Reporting randomised trials of social and psychological interventions : the CONSORT-SPI 2018 Extension. *Trials, 19*(1). https://doi.org/10.1186/s13063-018-2733-1.

Morales, E., McKiernan, E. C., Niles, M. T., Schimanski, L., & Alperin, J. P. (2021). How faculty define quality, prestige, and impact of academic journals. *PloS One, 16*(10), e0257340. https://doi.org/10.1371/journal.pone.0257340.

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

Nakamura, B. J., Higa-McMillan, C. K., Okamura, K. H., & Shimabukuro, S. (2011). Knowledge of and Attitudes Towards Evidence-Based Practices in Community Child Mental Health Practitioners. *Administration And Policy In Mental Health And Mental Health Services Research*, *38*(4), 287-300. https://doi.org/10.1007/s10488-011-0351-2.

Nelson, T. D., Steele, R. G., & Mize, J. A. (2006). Practitioner Attitudes Toward Evidence-based Practice : Themes and Challenges. *Administration And Policy In Mental Health And Mental Health Services Research*, *33*(3), 398-409. https://doi.org/10.1007/s10488-006-0044-4.

OCEBM. (2011). OCEBM Levels of Evidence Working Group. *"The Oxford Levels of Evidence 2"*. Oxford Centre for Evidence-Based Medicine. https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence.

Pagoto, S. L., Spring, B., Coups, E. J., Mulvaney, S. A., Coutu, M., & Ozakinci, G. (2007). Barriers and facilitators of evidence-based practice perceived by behavioral science health professionals. *Journal Of Clinical Psychology, 63*(7), 695-705. https://doi.org/10.1002/jclp.20376.

Rivero, T. S., Nuñez, L. M. H., Pires, E. U., & Bueno, O. F. A. (2015). ADHD Rehabilitation through Video Gaming : A Systematic Review Using PRISMA Guidelines of the Current Findings and the Associated Risk of Bias. *Frontiers In Psychiatry*, *6*. https://doi.org/10.3389/fpsyt.2015.00151.

Rousseau, D. M., & Gunia, B. C. (2016). Evidence-based practice: The psychology of EBP implementation. *Annual Review of Psychology, 67*(1), 667–692. https://doi.org/10.1146/annurev-psych-122414-033336.

Satterfield, J. M., Spring, B., Brownson, R. C., Mullen, E. J., Newhouse, R. P., Walker, B. B., et al. (2009). Toward a transdisciplinary model of evidence-based practice. *The Milbank Quarterly, 87*(2), 368–390. https://doi.org/10.1111/j.1468-0009.2009.00561.x.

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An Excess of Positive Results : Comparing the Standard Psychology Literature With Registered Reports. *Advances In Methods And Practices In Psychological Science*, *4*(2). https://doi.org/10.1177/25152459211007467.

Schulz, K. F. (1996). Randomised trials, human nature, and reporting guidelines. *Lancet, 348*(9027), 596-598. https://doi.org/10.1016/s0140-6736(96)01201-9.

Shamseer, L., Hopewell, S., Altman, D. G., Moher, D., & Schulz, K. F. (2016). Update on the endorsement of CONSORT by high impact factor journals: a survey of journal "Instructions to Authors" in 2014. *Trials, 17*(1). https://doi.org/10.1186/s13063-016-1408-z.

Simon, S. D. (2001). Is the Randomized Clinical Trial the Gold Standard of Research? *Journal of Andrology, 22*(6), 938-943. https://doi.org/10.1002/j.1939-4640.2001.tb03433.x.

Sink, C. A., & Mvududu, N. H. (2010). Statistical Power, Sampling, and Effect Sizes: Three Keys to Research Relevancy. *Counseling Outcome Research and Evaluation, 1*(2), 1-18. https://doi.org/10.1177/2150137810373613.

Spring, B. (2007). Evidence-based practice in clinical psychology: What it is, why it matters; what you need to know. *Journal Of Clinical Psychology, 63*(7), 611-631. https://doi.org/10.1002/jclp.20373.

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H., Corbett, M., Eldridge, S., Emberson, J., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T. J., Li, T., … Higgins, J. P. T. (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ. British Medical Journal*, l4898. https://doi.org/10.1136/bmj.l4898.

Stewart, R. E., Stirman, S. W., & Chambless, D. L. (2012). A qualitative investigation of practicing psychologists' attitudes toward research-informed practice : Implications for dissemination strategies. *Professional Psychology Research And Practice, 43*(2), 100-109. https://doi.org/10.1037/a0025694.

Straus, S. E., Tetroe, J. M., & Graham, I. D. (2011). Knowledge translation is the use of knowledge in health care decision making. *Journal of Clinical Epidemiology, 64*(1), 6–10. https://doi.org/10.1016/j.jclinepi.2009.08.016.

Tan, A. C., Jiang, I., Askie, L., Hunter, K., Simes, R. J., & Seidler, A. L. (2019). Prevalence of trial registration varies by study characteristics and risk of bias. *Journal of Clinical Epidemiology, 113*, 64-74. https://doi.org/10.1016/j.jclinepi.2019.05.009.

Téllez, A., García, C. H., & Corral-Verdugo, V. (2015). Effect size, confidence intervals and statistical power in psychological research. *Psychology In Russia State Of Art*, *8*(3), 27-46. https://doi.org/10.11621/pir.2015.0303.

The jamovi project (2022). *jamovi* (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org.

Thyer, B. A., & Pignotti, M. (2011). Evidence-Based practices do not exist. *Clinical Social Work Journal*, *39*(4), 328-333. https://doi.org/10.1007/s10615-011-0358-x.

Tufanaru, C., Munn, Z., Aromataris, E., Campbell, J., & Hopp, L. (2020). Chapter 3: Systematic reviews of effectiveness. In E. Aromataris & Z. Munn (Eds.), *JBI Manual for Evidence Synthesis.* JBI. Available from https://synthesismanual.jbi.global.

Wilson, J., Armoutliev, E., Yakunina, E. S., & Werth, J. L. (2009). Practicing psychologists' reflections on evidence-based practice in psychology. *Professional Psychology: Research And Practice, 40*(4), 403-409. https://doi.org/10.1037/a0016247.

Wolfenden, L., Foy, R., Presseau, J., Grimshaw, J. M., Ivers, N. M., Powell, B. J., Taljaard, M., Wiggers, J., Sutherland, R., Nathan, N., Williams, C. M., Kingsland, M., Milat, A., Hodder, R. K., & Yoong, S. L. (2021). Designing and undertaking randomised implementation trials : guide for researchers. *BMJ*, m3721. https://doi.org/10.1136/bmj.m3721.

Annexes

Annexe 1

Analyse, selon la grille RoB2, du risque de biais dans les articles *reporting* + et *reporting* -.

Risk of Bias 2		**Article R+**	**Article R-**
Domain 1: Risk of bias arising from the randomization process			
1.1	Was the allocation sequence random?	1	NA
1.2	Was the allocation sequence concealed until participants were enrolled and assigned to interventions?	1	NI
1.2	Did baseline differences between intervention groups suggest a problem with the randomization process?	1	1
Risk-of-bias judgement domain 1		Low Risk	Some concerns
Domain 2: Risk of bias due to deviations from the intended interventions (*effect of assignment to intervention*)			
2.1	Were participants aware of their assigned intervention during the trial?	1	NI
2.2	Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?	1	NI
2.3	If Y/PY/NI to 2.1 or 2.2: Were there deviations from the intended intervention that arose because of the trial context?	NA	NI
2.4	If Y/PY to 2.3: Were these deviations likely to have affected the outcome?	NA	NA
2.5	If Y/PY/NI to 2.4: Were these deviations from intended intervention balanced between groups?	NA	NA
2.6	Was an appropriate analysis used to estimate the effect of assignment to intervention?	1	NI
2.7	If N/PN/NI to 2.6: Was there potential for a substantial impact (on the result) of the failure to analyse participants in the group to which they were randomized?	NA	NI
Risk-of-bias judgement domain 2		Low Risk	High Risk
Domain 3: Missing outcome data			
3.1	Were data for this outcome available for all, or nearly all, participants randomized?	1	0
3.2	If N/PN/NI to 3.1: Is there evidence that the result was not biased by missing outcome data?	NA	0
3.3	If N/PN to 3.2: Could missingness in the outcome depend on its true value?	NA	NI
3.4	If Y/PY/NI to 3.3: Is it likely that missingness in the outcome depended on its true value?	NA	1
Risk-of-bias judgement domain 3		Low Risk	Some concerns
Domain 4: Risk of bias in measurement of the outcome			
4.1	Was the method of measuring the outcome inappropriate?	1	1
4.2	Could measurement or ascertainment of the outcome have differed between intervention groups?	1	1
4.3	If N/PN/NI to 4.1 and 4.2: Were outcome assessors aware of the intervention received by study participants?	1	NI
4.4	If Y/PY/NI to 4.3: Could assessment of the outcome have been influenced by knowledge of intervention received?	NA	NI
4.5	If Y/PY/NI to 4.4: Is it likely that assessment of the outcome was influenced by knowledge of intervention received?	NA	1
Risk-of-bias judgement domain 4		Low Risk	High Risk
Domain 5: Risk of bias in selection of the reported result			
5.1	Were the data that produced this result analysed in accordance with a pre-specified analysis plan that was finalized before unblinded outcome data were available for analysis?	1	1
5.2	Is the numerical result being assessed likely to have been selected, on the basis of the results, from multiple eligible outcome measurements (e.g. scales, definitions, time points) within the outcome domain?	1	1
5.3	Is the numerical result being assessed likely to have been selected, on the basis of the results, from multiple eligible analyses of the data?	1	1
Risk-of-bias judgement domain 5		Low Risk	Low risk
Risk-of-bias judgement		Low Risk	High Risk

Annexe 2

Analyse, selon la grille CONSORT-SPI, de l'exhaustivité du *reporting* des articles.

CONSORT-SPI		Article R+	Article R-
Title and abstract			
1a	Identification as a randomised trial in the title	1	0
1b	Structured summary of trial design, methods, results, and conclusions	0	0
Introduction			
2a	Scientific background and explanation of rationale	1	1
2b	Specific objectives or hypotheses	0	0
2c	How the intervention was hypothesized to work	0	0
Methods			
3a	Describe of trial design, including allocation ratio	0	0
3b	Important changes to methods after trial commencement with reasons	1	0
4a	Eligibility criteria for participants	1	1
4b	Setting and locations of intervention delivery and where the data were collected	0	0
5a	The interventions for each group with sufficient details to allow replication	1	0
5b	Which interventions were actually delivered by providers and taken up by participants as planned	1	0
6a	Completely defined pre-specified outcomes	0	1
6b	Any changes to trial outcomes after the trial commenced, with reasons	1	0
7a	How sample size was determined	1	0
7b	When applicable, explanation of any interim analyses and stopping guidelines	0	0
Randomisation			
8a	Method used to generate the random allocation sequence	0	0
8b	Type of randomisation; detail of any restriction	0	0
9	Mechanism used to implement the random allocation sequence, describing any steps taken to conceal the sequence until interventions were assigned	0	0
10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	0	0
11a	Who was aware of intervention assignment after allocation and how any masking was done	0	0
11b	If relevant, description of the similarity of interventions	1	1
12a	Statistical methods used to compare group outcomes	0	0
12b	How missing data were handled, with details of any imputation method	1	0
12c	Methods for additional analyses, such as subgroup analyses, adjusted analyses, and process evaluations	0	0
Results			
13a	Where possible, the number approached, screened, and eligible prior to random assignment, with reasons for non-enrolment	0	0
13b	For each group, losses and exclusions after randomisation, together with reasons	1	0
14a	Dates defining the periods of recruitment and follow-up	0	0
14b	Why the trial ended or was stopped	0	0
15	Include socioeconomic variables where applicable	0	0
16	For each group, number included in each analysis and whether the analysis was by original assigned groups	0	0
17a	For each outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	0	0
17b	Indicate availability of trial data	1	0
17c	For binary outcomes, the presentation of both absolute and relative effect sizes is recommended	0	0
18	Results of any other analyses performed, including subgroup analyses, adjusted analyses, and process evaluations, distinguishing pre-specified from exploratory	0	0
19	All important harms or unintended effects in each group	0	0
Discussion			
20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	0	1
21	Generalisability of the trial findings	0	1
22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	0	0
Important informations			
23	Registration number and name of trial registry	0	0
24	Where the full trial protocol can be accessed, if available	1	0
25a	Sources of funding and other support; role of funders	1	0
25b	Declaration of any other potential interests	1	0
26a	Any involvement of the intervention developer in the design, conduct, analysis, or reporting of the trial	1	0
26b	Other stakeholder involvement in trial design, conduct, or analyses	0	0
26c	Incentives offered as part of the trial	1	1
Total		17	7

Annexe 3

Improving Executive Functioning in Children with ADHD: Training Multiple Executive Functions within the Context of a Computer Game. A Randomized Double-Blind Placebo Controlled Trial

Abstract

Introduction

Executive functions (EFs) training interventions aimed at ADHD-symptom reduction have yielded mixed results. Generally, these interventions focus on training a single cognitive domain (e.g., working memory [WM], inhibition, or cognitive-flexibility). However, evidence suggests that most children with ADHD show deficits on multiple EFs, and that these EFs are largely related to different brain regions. Therefore, training multiple EFs might be a potentially more effective strategy to reduce EF-related ADHD symptoms.

Methods

Eighty-nine children with a clinical diagnosis of ADHD (aged 8–12) were randomized to either a full-active-condition where visuospatial WM, inhibition and cognitive-flexibility were trained, a partially-active-condition where inhibition and cognitive-flexibility were trained and the WM-training task was presented in placebo-mode, or to a full placebo-condition. Short-term and long-term (3-months) effects of this gamified, 25-session, home-based computer-training were evaluated on multiple outcome domains.

Results

During training compliance was high (only 3% failed to meet compliance criteria). After training, only children in the full-active condition showed improvement on measures of visuospatial short-term-memory (STM) and WM. Inhibitory performance and interference control only improved in the full-active and the partially-active condition. No Treatment-condition x Time interactions were found for cognitive-flexibility, verbal WM, complex-reasoning, nor for any parent-, teacher-, or child-rated ADHD behaviors, EF-behaviors, motivational behaviors, or general problem behaviors. Nonetheless, almost all measures showed main Time-effects, including the teacher-ratings.

Conclusions

Improvements on inhibition and visuospatial STM and WM were specifically related to the type of treatment received. However, transfer to untrained EFs and behaviors was mostly nonspecific (i.e., only interference control improved exclusively in the two EF training conditions). As such, in this multiple EF-training, mainly nonspecific treatment factors – as opposed to the specific effects of training EFs – seem related to far transfer effects found on EF and behavior.

Trial Registration: trialregister.nl NTR2728. Registry name: improving executive functioning in children with ADHD: training executive functions within the context of a computer game; registry number: NTR2728.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors have no support or funding to report.

Competing Interests: P.J.M.P. is member of Stichting Gaming & Training, a nonprofit organization that facilitates the development and implementation of "Braingame Brian."; S.v.d.O. has been a paid consultant for Janssen Pharmaceuticals with regard to "Healseeker," a serious game for cognitive function training. S.D. and R.W.W. declare no competing interests exist. The statements in the competing interest section do not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

Introduction

Theories of ADHD suggest that deficits in executive functioning are at the core of the ADHD-syndrome, and play a pivotal role in explaining the problems children with ADHD encounter in daily life [1], [2], [3], [4]. Via dorsal frontostriatal brain circuits, executive functions (EF) allow individuals to regulate their behavior, thoughts and emotions, and thereby enable self-control [5]. Evidence indeed suggests that impairments in EF are related to deficits in attention, hyperactivity and impulsivity [6], [7], [8], [9], [10], [11], and with associated problems such as deficient academic functioning [12], [13]. Moreover, research suggests that EF-capacity and its associated levels of brain activity are not static, but may be altered by task-repetition or training [14]. Therefore, in the past few years, EF training interventions aimed at ADHD symptom reduction have received considerable interest.

Nonetheless, these EF interventions have yielded mixed results, especially on ADHD behavior (for an overview see [15], [16], [17], [18], [19], [20]; in addition see [21], [22], [23]). Generally, these interventions focus on training a single domain of cognitive functioning in children with ADHD, such as working memory (WM), inhibition, or cognitive flexibility. However, evidence suggests that most children with ADHD show deficits on multiple EFs [24], and that these EFs are largely related to different brain regions [25], [26], [27]. Therefore, training of multiple EFs might be a potentially more effective strategy to reduce EF related ADHD symptoms.

To date, evidence for multiple EF training interventions is limited. Few studies have investigated the effects of these interventions in children with ADHD [28], [29], [30], [31], [32], [33], and although these studies generally show promising results (e.g., improvement of ADHD behavior as rated by parents and/or a significant other [e.g., the teacher]; an increase of neural activity and gray matter volume in ADHD affected brain areas), none of these studies are placebo-controlled.

Besides EF deficits, children with ADHD have problems with motivation. Motivational models [34], [35], [36], [37], and subsequent research (for an overview see [38], [39]; also see [40], [41], [42], [43]) suggest that children with ADHD are less stimulated by reinforcement (i.e. reward) than typically developing children (probably due to a dopaminergic deficit), and therefore require higher amounts and frequencies of reward in order to perform optimally. This elevated need for reinforcement in children with ADHD may result in motivational problems during EF training: the child has to repeat the same responses over and over again for many trials, making most EF training programs tedious and boring for children with ADHD [44]. Research suggests that motivational problems can decrease the effects of EF training in children with ADHD [45]. However, gamification of an EF training or task (e.g., by using game mechanics and visuals) has been found to optimize both motivation and training-effects in children with ADHD [40], [45], [46]. Gaming increases the release of striatal dopamine [47], [48], promoting long-term potentiation of neural connections within the striatum [49], which is suggested to improve motivation and one's ability to learn [50] (e.g., during EF training).

In the current double-blind, placebo-controlled study, we investigated the efficacy of a gamified, 5-week, home-based, multiple EF training intervention titled Braingame Brian (BGB; [44]) in children with ADHD (combined-subtype). A previous waitlist-controlled study of BGB [31] showed promising results on reduction of symptoms of ADHD and improvement of EF. BGB targets multiple EFs that are commonly impaired in children with ADHD: visuospatial WM, response inhibition, and cognitive flexibility [51]. To date, most EF-training studies focus on the effects of WM training (e.g., see [15]), whereas very few studies investigate the unique effects (i.e. without WM training) of response inhibition- and/or cognitive flexibility training in children with ADHD. Only Kray et al. [23] investigated effects of a cognitive flexibility training in children with ADHD; they found placebo-controlled effects on untrained EF performance (i.e., interference control), but they did not investigate effects on behavior. Moreover, we are not aware of any studies investigating the unique effects of inhibition training in children with ADHD (for studies of combined WM and inhibition training see [28], [29], [30]). Therefore, participants in the current study were randomized to one of three treatment conditions: (1) a full-active-condition where visuospatial WM, response inhibition and cognitive flexibility were trained, (2) a partially-active-condition where only inhibition and cognitive flexibility were trained and the visuospatial WM training-task was presented in placebo-mode, or (3) to a full placebo-condition. Short-term and long-term (3-months) effects were evaluated across various outcome measures (including performance measures of WM, inhibition, cognitive-flexibility, interference control, and complex reasoning, and rating scales assessing parent- and teacher-rated ADHD behavior, parent-rated EF- and motivational behavior, and parent-, teacher- and child-rated general problem behavior).

We expected that: (1) improvement on outcome measures of WM, inhibition, and cognitive flexibility (i.e., performance measures and EF rating-scales) would be specifically related to the type of treatment received (e.g., greatest improvement on WM if WM was trained), (2) the (far-) transfer of treatment effects to other, untrained, domains of EF (such as interference control or parent-rated planning, organization of materials or self-monitoring) would be limited. We expected that spill-over effects to untrained domains of EF (far transfer) would be limited

because different EFs are largely related to different brain regions [25], [26], [27], and because most placebo-controlled EF training studies that investigate children with ADHD do not find such far transfer effects (e.g., see [19]), (3) children in the full-active condition would improve significantly more on ADHD behavior than children in either the partially-active condition or placebo condition, and (4) children in the partially-active condition would improve significantly more on ADHD behavior than children in the placebo condition. Finally, we also investigated other domains of impairment that are associated with ADHD (such as sensitivity to reward and punishment, oppositional defiant behavior, quality of life, and problems in daily situations). However, given the current knowledge-base in the field (e.g., there are no placebo- controlled EF training studies that investigate effects on sensitivity to reward and punishment, quality of life or problems in daily situations, and placebo-controlled studies investigating effects on oppositional defiant behavior show mixed results [21], [23], [84], [85], [86]), we refrained from presenting hypotheses regarding these domains of impairment.

Methods

Trial Design

This was a multicenter (14 sites), double-blind, placebo-controlled, multi-arm parallel-group study conducted in the Netherlands (trial register: http://www.trialregister.nl/trialreg/admin/rctview.asp?TC=2728; registry name: improving executive functioning in children with ADHD: training executive functions within the context of a computer game; registry number: NTR2728). No important changes to methods were made after trial commencement (the trial started April 2011 and ended January 2013). The protocol for this trial and CONSORT checklist are available as S1 Protocol and S1 CONSORT Checklist.

Participants

Study settings. Children were recruited from 14 outpatient mental-healthcare centers. This study was conducted in the Netherlands, within a predominantly urban type of community.

Eligibility criteria. Eligible participants were all children aged 8 to 12 years with (a) a prior DSM-IV-TR [52] diagnosis of ADHD combined-type and absence of any autism spectrum disorder according to a child psychologist or psychiatrist, (b) a score within the clinical range (95th to 100th percentile) on the ADHD scales of both the parent and teacher version of the Disruptive Behavior Disorder Rating Scale (DBDRS [53]; Dutch translation: [54]), (c) meeting criteria for ADHD combined-type on the ADHD section of the Diagnostic Interview Schedule for Children, parent version (PDISC-IV [55]). The PDISC-IV is a structured diagnostic interview based on the DSM-IV, with adequate psychometric properties, (d) absence of conduct disorder (CD) based on the CD sections of the PDISC-IV, (e) an IQ score >= 80 established by the short version of the Dutch Wechsler Intelligence Scale for Children (WISC-III; [56]). Two subtests, Vocabulary and Block Design, were administered to estimate Full Scale IQ (FSIQ). This composite score has satisfactory reliability and correlates highly with FSIQ [57], (f) absence of any neurological disorder, sensory (color blindness, vision) or motor impairment as stated by the parents, (g) not taking any medication other than Methylphenidate or

Dextroamphetamine. Participants discontinued their Methylphenidate at least 24 hours before each test-session, allowing a complete wash-out [58]. Participants taking Dextroamphetamine discontinued medication 48 hours before each test-session [59], finally, (h) parents had to agree to keep the dose of ADHD medication stable between the intake and the 3-months follow-up session, and had to consent not to initiate or participate in other psychosocial treatments. Group differences in baseline demographics and clinical characteristics are listed in Table 1.

Table 1. Baseline Demographics and Clinical Characteristics by Treatment Group.

Measure	Treatment Group							
	Full-Active		Partially-Active		Placebo			
	(n = 31) M	SD	(n = 28) M	SD	(n = 30) M	SD	F / χ^2	*Group Comparison*[a]
Gender (M:F)	25:6	-	22:6	-	24:6	-	.04	ns (p = .980)
Age (years)	10.6	1.4	10.3	1.3	10.5	1.3	.58	ns (p = .564)
FSIQ	101	11.5	101	11.4	101	11.6	.05	ns (p = .956)
DBDRS parent								
Inattention	22.0	3.6	21.3	4.1	21.9	4.6	.23	ns (p = .793)
Hyperactivity/Impulsivity	21.3	3.8	20.0	4.6	20.5	5.1	.69	ns (p = .504)
ODD	11.6	5.8	12.8	4.6	11.7	5.9	.40	ns (p = .674)
CD	2.9	3.1	2.7	2.9	3.2	2.9	.20	ns (p = .820)
DBDRS teacher								
Inattention	16.1	5.6	15.9	5.0	18.0	4.8	1.54	ns (p = .220)
Hyperactivity/Impulsivity	13.8	6.2	14.3	5.8	16.6	6.0	1.84	ns (p = .166)
ODD	7.4	6.0	7.1	5.0	8.6	6.6	.49	ns (p = .614)
CD	1.1	1.7	2.1	3.0	1.9	2.5	1.22	ns (p = .300)
PDISC-IV								
ODD diagnosis, *N* (%)	17 (55%)	-	18 (64%)	-	15 (50%)	-	1.24	ns (p = .539)
ADHD medication[b], *N* (%)	20 (65%)	-	19 (68%)	-	22 (73%)	-	.56	ns (p = .756)
Computergame experience (hours per week)	8.6	5.0	9.8	9.1	11.6	8.4	1.17	ns (p = .314)
Dyscalculia, *N* (%)	0 (0%)	-	0 (0%)	-	0 (0%)	-	-	-
Dyslexia, *N* (%)	2 (7%)	-	5 (18%)	-	5 (17%)	-	2.03	ns (p = .362)

Note. CD = conduct disorder; DBDRS = Disruptive Behavior Disorder Rating Scale; FSIQ = full scale IQ; M:F = Male:Female; ODD = oppositional defiant disorder; PDISC-IV = Diagnostic Interview Schedule for Children, parent version;

[a] Continuous data were investigated using ANOVAs. Nominal data were investigated using Pearson's chi-squared tests;

[b] Four children were taking Dextroamphetamine (two in the full-active condition, one in the partially-active condition, and one in the placebo condition).

doi:10.1371/journal.pone.0121651.t001

Treatment Conditions

General characteristics of the intervention. "Braingame Brian" (BGB [44]) is a computerized, home-based EF training, embedded in a game world and is named after its main character "Brian". Brian is a young inventor who, throughout the game, helps and befriends the game-worlds inhabitants by creating increasingly elaborate inventions (e.g., a delivery-rocket for the grocery-store owner). BGB consists of 25 training sessions. Within each session, the player can create inventions by completing two blocks of three training tasks. Within each block, the first training task is always a WM task (used for drawing a blueprint of the invention), the second and third task, a cognitive flexibility task and an inhibition task, are presented in changing order (and are used for sorting building-materials, and electrically-charging the invention). Each session takes about 35–50 minutes (30 minutes for completing the tasks and an optional amount of time for game-world exploration). An additional standardized external reward system – receiving game-related stickers, reward ribbons and medals for completing sessions (the same for all participants) – is used to even further raise the child's motivation to do the training (for more details see [44] and S1 Appendix). In the current study BGB was presented in three conditions:

Full-active condition. In this condition WM, inhibition and cognitive-flexibility were all in training-mode. Training-mode entailed that, after each block of training tasks, the difficulty level of the training task was automatically adjusted to the child's level of performance. Furthermore, in training-mode (a) the WM task [60] consisted of five training levels: the first level targeted visuospatial short-term memory (STM) only, whereas the other four levels targeted combinations of visuospatial STM, updating and manipulation of information (i.e. these four levels targeted both STM and the central executive). Each level was trained for 5 of the 25 sessions. The difficulty level was increased by increasing the amount of information that had to be remembered, updated and manipulated, (b) the inhibition task [61] was designed to decrease the time needed to inhibit a prepotent response (comparable with the stop signal reaction time measured by the STOP task [62]). On most trials the child had to respond to a go-stimulus by pressing left or right within a specific time-frame (a green colored response window between 550–850 ms). This created a prepotent response tendency. However, on 25% of the trials, somewhere after the go-stimulus and before the middle of the response window, a stop-signal was presented (a tone and a visual cue) and the child had to inhibit the prepotent response (stop-trials). The difficulty level was increased by shortening the time allowed to inhibit this response, (c) the cognitive-flexibility task [61] was designed to decrease the time a child needs to adapt his/her behavior when task-rules change (i.e. switch cost). Specifically, the child had to sort objects with different shapes and colors (e.g. blue or red colored plungers and wheels) to either the left or the right according to a rule. The rule was either to sort according to shape or to sort according to color. In 25% of the trials the rule switched (switch-trials). The difficulty level was increased by shortening the time allowed to switch between the two rules (for a more detailed description of the three training tasks see [31]).

Partially-active condition. In this condition the inhibition and cognitive-flexibility tasks were in training-mode, and the WM task was in placebo-mode. Placebo-mode entailed that only the first level of the WM task was presented (for all 25 sessions), and that the difficulty level was not adjusted to the child's level of performance (no more than two items had to be remembered). The amount of trials in placebo-mode was increased to match the training time in training-mode (10 minutes training per session for each EF domain).

Placebo condition. In this condition WM, inhibition and cognitive-flexibility were all in placebo-mode. In placebo-mode the inhibition task and the cognitive-flexibility task were presented the same way as in training-mode except that the stop-trials and switch-trials were replaced by go-trials and non-switch trials (i.e., no stop-trials and switch-trials were presented) and the difficulty level was not adjusted.

Process Measures

No important changes to outcome measures were made after trial commencement.

Compliance. Compliance was defined as completing all of the 25 training sessions within a 5-week period. Using this algorithm, each child was categorized as compliant or noncompliant to treatment.

Blinding. At post-test, parents were asked to report the condition they thought their child was assigned to (full-active, partially-active, or placebo).

Improvement index during training. To validate whether the training actually improved task performance on the designated EFs, the improvement on training performance from beginning to end of training was assessed. It was tested whether children improved during training with paired t-tests. For the inhibition training and the cognitive flexibility training the results of day 2 and 3 of training (the Start Index) were compared with the results of their two best training days (the Max Index). The WM training had five levels and each level covered only 5 of the 25 training days. Therefore, to measure improvement on the WM training, within each level, the results of day 2 of training (the Start Index) were compared with the results of the best training day (the Max Index).

Performance Measures

Stop task. The Stop task was used to measure the time needed to inhibit an ongoing response [62]. Two types of trials were presented: go-trials and stop-trials. During go-trials a go-stimulus (an arrow) that was either pointing right or left was presented. Participants were instructed to press a response button that corresponded to the direction of the stimulus as quickly and as accurately as possible. Stop-trials were identical to the go-trials but in addition a stop-signal was presented (a tone and a visual cue), which indicated that the participant had to withhold his/her ongoing response. The delay between the go- and stop-signal was dynamically varied (in steps of 50ms) so that inhibition was successful in 50% of the stop-trials. At this point, the go-process and stop-process are of equal duration, which makes it possible to estimate the latency of the stop-process: the stop signal reaction time (SSRT [62]). Aside from two practice blocks, four experimental blocks (of 64 trials each) were administered. The SSRT was used as outcome measure of inhibitory processing. Test retest reliability of the SSRT in children with ADHD is. 72 [63].

Stroop. The Stroop Color and Word Test [64] measures interference control and consists of three pages with words and/or colors. On the first page, word naming is measured by naming the words red, green, yellow, and blue, printed in black ink. On the second page, color naming is measured by naming the colors of small rectangles. The first and second page represent the congruent trials. On the third page, colors are then named when shown as nonmatching color words (incongruent trials). The interference score on the Stroop is the time needed for the third page minus the time needed for the second page, and was used as our outcome measure of interference control. The STROOP has adequate reliability [65].

Corsi Block Tapping Task (CBTT). The CBTT [66] assesses the capacity of visuospatial STM and WM. The task consists of nine cubes (blocks) that are positioned on a board. In the present study, the same test format (size of board and blocks, distances between blocks) was used as in Kessels, van Zandvoort, Postma, Kappelle, and de Haan [67] (also see [68]), and the same procedure was used as in Geurts, Verté, Oosterlaan, Roeyers, and Sergeant [69]. The experimenter tapped a sequence of blocks that a child then had to reproduce in the same (CBTT-forward) or in reversed order (CBTT-backward). The minimum sequence length was three and

the maximum was eight blocks, and each length was presented on three trials. The total amount of sequences that is correctly reproduced is the total score. The total score on the CBTT-forward (max. total score = 18) was used as an outcome measure for visuospatial STM and the total score on the CBTT-backward (max. total score = 18) was used as an outcome measure of visuospatial WM. The CBTT shows good reliability [70].

Digit span. The scaled score on the Digit-span subtest from the WISC-III testing battery [56] was used as a composite measure of verbal STM and WM. Participants were orally given sequences of numbers and were asked to repeat them, either in the same (i.e. STM) or in reversed order (i.e. WM). Digit span has adequate reliability [56].

Trail Making Test (TMT). The TMT of the Delis-Kaplan Executive Function System (D-KEFS [71]) measures cognitive flexibility and is a timed task that requires the individual to connect a series of letters and numbers in ascending order while alternating between numbers and letters. The scaled contrast score – the contrast between the scaled non-switch trials (number and letter sequencing) and the scaled switch trials (number-letter switching) – was used as outcome measure of cognitive flexibility (i.e., switch-cost). Test-retest reliabilities range from. 20 to. 77 [71].

Raven coloured progressive matrices. Raven's coloured progressive matrices [72] measures non-verbal reasoning ability. The test consists of 36 items. The total amount of items correct (total score; max. = 36) was used as outcome measure for non-verbal reasoning. Test-retest reliability ranges from. 68 to. 90 [73].

Questionnaires and Rating Scales

DBDRS (parent and teacher versions). The DBDRS contains four DSM-IV scales; Inattention, Hyperactivity/ Impulsivity, Oppositional Defiant Disorder (ODD), and CD. Parents and teachers rate the child's behavior on a 4-point Likert-type scale. Adequate psychometric properties have been reported [54]. The scores on the Inattention and Hyperactivity/Impulsivity scales were used as outcome measure of ADHD behavior. The scores on the ODD and CD scales were used as outcome measures of general problem behavior.

Behavior Rating Inventory of Executive Function questionnaire (BRIEF). [74]. The Dutch version of the BRIEF is used to assess parent-rated EF. The BRIEF consists of 75 questions and includes eight EF sub-domains: Inhibit, Shift, Emotional Control, Initiate, WM, Plan/Organize, Organization of Materials, and Monitor. The test has adequate psychometric properties [75]. T-scores on the EF sub-domains were used as outcome measures.

Sensitivity to Punishment and Sensitivity to Reward Questionnaire for children (SPSRQ-C). The SPSRQ-C measures parent-rated sensitivity to punishment and reward [76] (Dutch translation: [77]) and contains 33 items, divided in a Punishment Sensitivity scale, and three Reward Sensitivity scales: Reward Responsivity, Impulsivity/Fun-Seeking, and Drive. Each item is scored on a 5-point Likert scale. Adequate psychometric properties are reported [76]. Subscale scores were used as outcome measures.

Pediatric Quality of Life Inventory (PedsQL; parent and child versions). [78] (Dutch translation: [79]). The PedsQL consists of 23 items, scored on a five-point Likert-scale, and is divided in four subscales: Physical, Emotional, Social, and School Functioning. The Psychosocial Health Summary score (a composite of the Emotional, Social and School Functioning subscales) was used as outcome measure. Adequate psychometric properties are reported [79].

The Home Situations Questionnaire (HSQ). The HSQ [80] is designed to assess the impact of problem behavior at home and in public situations. Parents report whether each of 16 daily situations (e.g. getting dressed and going to bed) was a problem and rate their severity on a 9-point scale. The mean severity score was used as outcome measure. The HSQ has adequate psychometric properties [81].

Procedure

This study was approved by the faculty's IRB (the Ethics Review Board of the Faculty of Social and Behavioral Sciences of the university of Amsterdam). After obtaining written informed consent from the parents (on behalf of the participating children), parents and teachers completed the DBDRS. At this first screening the 6-month version of the DBDRS was administered (regarding the child's behavior over the past 6-months), whereas at the pre-test, post-test and follow-up a two-week version of the DBDRS was administered (regarding the child's behavior over the past two-weeks). If DBDRS inclusion criteria were met, children and parents were invited to the intake session. During this session questions regarding demographics were asked (see Table 1), and the PDISC-IV, and the short-form of the WISC-III were administered. The Chessboard WM task (for a detailed description see Dovis et al., 2013) was also administered during the intake session. However, this task was part of a different study and its results will therefore be reported elsewhere. If inclusion criteria were met, parent and child were invited to the pre-test session and the startup session, and were independently allocated to one of the three treatment conditions using the process of randomization by minimization [82] on the basis of age, gender, IQ, medication-use (yes/no), and parent- and teacher-rated inattention and hyperactivity/impulsivity symptoms (using the 6-months DBDRS). During the pre-test session the outcome measures were administered, and in the same week the teacher completed the two-week version of the DBDRS. The pre-test occurred approximately 1–2 weeks prior to the startup session (which was the start of the training). During the startup session parent and child were instructed about the training program, the computer, and the external reward system (see S1 Appendix), and a schedule for implementing the intervention and for weekly coaching calls was established. Once a research assistant completed a startup session with a particular family, he/she could not test or have further contact with that family or the teacher (to preserve blinding). During the 5-week, home-based training, a coach (a research assistant blind to the treatment condition) made weekly calls (of about 15 minutes; using a standardized telephone protocol) to the participating families to monitor progress, motivation and compliance, and to solve technical and game-related problems. Parents and children were explicitly instructed not to discuss the content of the training tasks with the coach. If a coach did receive information revealing the treatment condition, he/she was

replaced and could no longer have contact with the family or the teacher. 1–2 weeks after the final training session the post-test was scheduled and the teacher completed the DBDRS. 3-months after the final training session the follow-up was scheduled and the teacher completed the DBDRS. At each test-session experimenters were blind to condition.

Statistical Analyses

Sample size was determined by a prospective power analyses for univariate testing (using G*Power) based on the effect sizes of two previous EF-treatment studies [86], [45]. These studies suggested that the treatment effects on our primary outcome measures (i.e., EF measures, ADHD rating-scales) would be medium in size. Groups did not differ with respect to any of the baseline demographics or clinical characteristics (see Table 1). Also, including these baseline demographics and clinical characteristics (i.e., Gender, Age, FSIQ, DBDRS parent and teacher ratings, ODD diagnosis, ADHD medication use, Computergame experience, and Dyslexia) as covariates in the main analyses did not change the pattern of our results. Because repeated-measures were used, covariates were entered after mean centering (see [97]). Multinomial logistic regression was used for assessing the effectiveness of blinding.

An Intent-To-Treat (ITT) approach, using single imputations, was used to compare treatment effects of the three treatment conditions. That is, for each treatment group stochastic regression imputation was used to predict the missing posttest and follow-up values. The missing posttest values were based on the non-missing pretest and posttest scores of each treatment group. The missing follow-up values were based on the non-missing pretest scores, posttest scores, follow-up scores, and pretest-posttest difference scores of each treatment group (although the overall percentage of missing data was low – less than 5% was missing – it must be noted that stochastic regression imputation can increase the probability of making type I errors).

The dependent measures were subjected to four repeated measures MANOVAs (for the performance measures, for ADHD behavior, for EF and motivational behavior, and for general problem behavior; the covariance matrices were assumed to be unstructured), with Treatment condition (full-active, partially-active, placebo) as between-subject factor and Time (pre-test, post-test, follow-up) as within-subject factors. Bonferroni corrections for multiple testing were applied to these MANOVAs: only p-values < .0125 [.05/4] were considered significant. Trends and significant effects were further analyzed with simple contrasts. Bonferroni corrections for multiple testing were applied to these contrasts, in which the amount of dependent variables corrected for was defined per repeated measures MANOVA (7 performance-, 4 ADHD behavior-, 12 EF and motivational behavior-, and 7 general problem behavior variables were each analyzed in 3 pair-wise time comparisons [pre-test vs. post-test, post-test vs. follow-up, and pre-test vs. follow-up], resulting in a required significant level of p = .0024 [.05/21] for the performance measures contrasts, p = .0042 [.05/12] for the ADHD behavior contrasts, p = .0014 [.05/36] for the EF and motivational behavior contrasts, and p = .0024 [.05/21] for the general problem behavior contrasts). For additional within-group analyses paired t-tests were used (Bonferroni corrections were applied). Partial Eta squared effect sizes (η_p^2) are reported for all analyses: η_p^2 = .01 is regarded a small effect size, .06 a medium effect size, and .14 a large effect size [83].

Results

Process Measures

Compliance during training. Of the 31 participants assigned to the full-active condition, 30 (96.7%) met compliance criteria (25 training days within 5 weeks). All of the 28 participants assigned to the partially-active condition met compliance criteria. Of the 30 participants assigned to the placebo condition, 28 (93.3%) met compliance criteria. Overall, compliance to treatment was high, given that this was a home-based intervention that included a substantial portion of participants with ODD (see Fig 2).

Post-training dropout. Eight participants (9%) of our total sample (i.e., 3 children in the full-active condition, 2 children in the partially-active condition, and 3 children in the placebo condition) were lost to post-test and follow-up testing (see Fig 2). There were no significant differences on baseline demographics and clinical characteristics (i.e., gender, age, IQ, DBDRS parent and teacher ratings, ODD-diagnosis, medication use, computergame experience, Dyscalculia and Dyslexia) between the children lost to post-test and follow-up testing and the children who participated in these assessments (depending on the level of measurement a MANOVA or Pearson's chi-squared tests were used). But note that the sample size of the post-training drop out group was small.

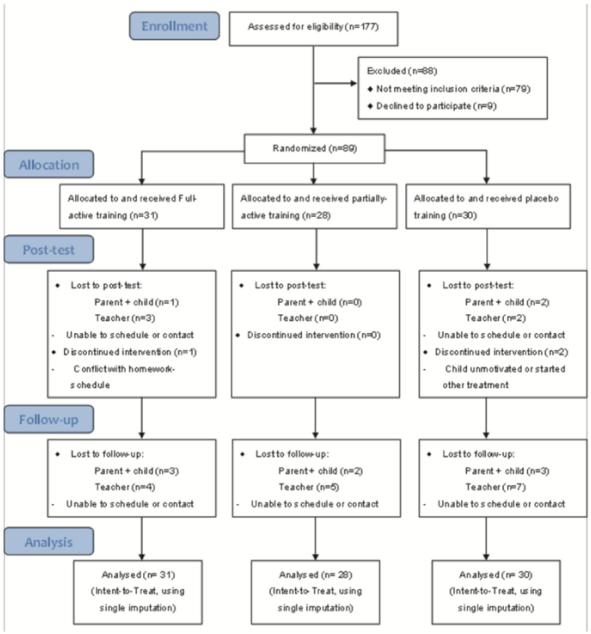

Fig 2. CONSORT flow diagram.

doi:10.1371/journal.pone.0121651.g002

Blinding. There was no significant association between the conditions wherein participants were actually included and the conditions whereof parents afterwards reported (guessed) that their child was assigned to (the multinomial logistic regression model indicated a non-significant model overall, $\chi 2(4) = 1.26$, $p = .868$, $-2LL = 18.004$). This suggests that, based upon their experience with the actual training condition, parents were not able to guess the condition wherein their child was included. Further, no participant (child, parent, teacher, experimenter, or coach) was unblinded at any point during the conduct of the trial.

Improvement index during training. It was tested whether children improved during training with paired t-tests (Bonferroni corrections for multiple testing were applied: only p-values < .0024 [.05/21] were considered significant). Within the full-active condition, paired t-tests showed a significant difference (improvement) between the Start Index and Max Index for the Inhibition training, t (30) = -18.66, p < .001, the Cognitive flexibility training, t (30) = -19.14, p < .001, and for all the levels of the WM training (level 1, t (30) = -7.25, p < .001; level 2,t(30)=-7.90,p<.001;level3,t(30)=-7.19,p<.001,level4,t(30)=-9.21,p<.001;level 5, t (30) = -7.72, p < .001). Within the partially-active condition (where WM was in placebo-mode), paired t-tests showed significant difference (improvement) between the Start Index and the Max Index for both the Inhibition training, t (27) = -15.86, p < .001, and the Cognitive flexibility training, t (27) = -22.89, p < .001.

Performance Measures

A 3x3 (Treatment condition x Time [pre-test, post-test, follow-up]) repeated measures MANOVA with the main scores of the EF tasks (Stoptask [SSRT], STROOP [interference score], CBTT-fwd [total score], CBTT-bkw [total score], Digit recall [scaled score], TMT [scaled con- trast score]) and the Raven (total score) as dependent variables (scores on all seven tasks were analyzed simultaneously), showed a main effect of Time, F (14,334) = 6.74, p < .001, η_p^2 = .22, no main effect of Treatment condition, F (14,162) = 1.41, p = .154, η_p^2 = .11, and a non-significant trend towards an interaction between Treatment condition and Time, F (28,676) = 1.59, p = .027, η_p^2 = .06 (after Bonferroni correction only p-values < .0125 were considered significant). To interpret these effects for each performance based measure, we used simple contrasts:

For each performance based measure, main Time effects and Treatment condition x Time interactions are presented per pair-wise time difference (i.e. pre- vs. post-test, post- vs. follow-up test; pre- vs. follow-up test) in Table 2.

After Bonferroni correction (p < .0024 [.05/21]) results indicate the following: Between the pre- and post-test there was a significant Treatment condition x Time interaction for the CBTT-fwd (p = .002), and a non-significant trend for the Stoptask (p = .037) and the CBTT-bkw (p = .039; see Table 2). Between the pre-test and follow-up there was a non-significant trend towards a Treatment condition x Time interaction for the CBTT-fwd (p = .013) and the STROOP (p = .07; see Table 2). Other pair-wise time differences in Treatment condition x Time interaction effects were non-significant both with- and without Bonferroni correction (investigating Digit recall forward and backward separately [using raw scores] did not change the results). Next, in order to obtain more insight into these two-way interactions, three follow-up repeated measures MANOVAs were performed: one for each combination of treatment conditions (Bonferroni corrections were applied: only p-values < .0167 [.05/3] were considered significant).

Full-active condition versus placebo condition. A 2x3 (Treatment condition x Time) repeated measures MANOVA with the main scores of the Stoptask, STROOP, CBTT-fwd, and CBTT-bkw as dependent variables, showed a main effect of Time, F (8,232) = 6.22, p < .001, η_p^2 = .18, a main effect of Treatment condition, F (4,56) = 5.06, p = .009, η_p^2 = .21, and a significant

interaction between Treatment condition and Time, F (8,232) = 3.90, p < .001, η_p^2 = .12. To further interpret this interaction for each relevant pair-wise time difference and each performance based measure, we used simple contrasts (in the previous contrast analyses no Treatment x Time interactions were found between post-test and follow-up; therefore, only the two-way interactions between pre-test and post-test and between pre-test and follow-up were further explored):

These contrasts are presented in Table 3 and indicate that, compared to pre-test performance, post-test- and/or follow-up performance on the Stoptask, the STROOP and the CBTT forward and backward improved more in the full-active condition than in the placebo condition (p-values ranged from. 002 to. 020; effect sizes ranged from medium to large; see Table 3 and Fig 3A–3D). However, after Bonferroni correction only the Treatment x Time interactions for the CBTT-fwd remained significant (as only p-values < .0063 [.05/8] were considered significant).

Table 2. Outcomes at Baseline, Post-test and Follow-up.

	Full-active Condition						Partially-active Condition						Placebo Condition						*Time constrasts. F*(1. 86)						*Treatment* * Time constrasts. F(2. 86)					
	Pre		Post		FU		Pre		Post		FU		Pre		Post		FU		Pre vs. Post		Post vs. FU		Pre vs. FU		Pre vs. Post		Post vs. FU		Pre vs. FU	
Domain and Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*	η_p^2	*F*	η_p^2	F	η_p^2	F	η_p^2	F	η_p^2	F	η_p^2
Performance Measures																														
Stoptask (SSRT)	189.6	43.7	151.4	39.7	147.4	40.9	197.0	65.2	158.8	31.6	150.2	48.7	200.1	73.8	197.7	69.0	189.0	68.6	16.50***	.16	1.18	.01	18.41***	.18	3.42*	.07	.06	.001	2.08	.05
STROOP (interfence score)	68.8	33.9	53.2	24.1	46.2	19.9	70.5	33.0	49.4	23.7	53.2	25.3	68.0	30.4	55.9	26.4	62.6	33.0	28.50***	.25	.17	.002	23.27***	.21	.71	.02	2.29	.05	2.69†	.06
CBTT-forward (total score)	9.6	2.4	11.7	2.3	11.0	2.4	9.5	2.2	10.1	2.3	9.7	2.5	9.9	2.9	10.1	1.9	9.6	2.1	18.08***	.17	10.13**	.11	2.72	.10	6.96**	.14	.21	.01	4.59*	.10
CBTT-backward (total score)	8.7	2.3	10.1	2.9	10.0	2.3	9.1	2.5	9.4	2.0	9.3	2.3	9.2	2.5	9.2	2.3	9.4	2.3	5.33*	.06	.004	<.001	4.87*	.05	3.36*	.07	.23	.01	2.53	.06
Digit Recall (scaled score)	9.3	2.9	9.9	3.3	9.9	3.1	9.1	3.2	9.9	3.3	9.6	3.3	9.8	2.6	9.9	2.9	9.5	3.1	3.12	.04	.45	.01	.98	.01	.47	.01	.21	.01	.95	.02
TMT (scaled contrast score)	8.9	3.1	8.6	2.8	8.6	2.8	8.7	3.1	8.9	3.5	9.0	2.3	8.7	2.6	8.6	2.4	8.7	2.1	.02	<.001	.002	<.001	.01	<.001	.16	<.001	.02	<.001	.23	.01
Raven (total score)	33.0	2.7	33.5	2.1	34.2	1.7	32.1	3.5	32.9	3.3	33.6	2.5	33.2	2.2	33.7	2.0	33.7	1.9	6.86*	.07	5.21*	.06	25.81***	.23	.11	.002	1.50	.03	2.04	.05
ADHD Behavior																														
P-DBDRS Inattention	17.5	4.3	11.9	5.7	12.9	4.1	17.7	5.5	12.6	4.5	14.6	5.3	18.2	4.4	13.6	5.2	14.1	4.7	66.21***	.44	5.31*	.06	48.37***	.36	.23	.01	.92	.02	.69	.02
P-DBDRS Hyp/ Imp	17.0	5.3	12.2	6.6	12.6	6.4	16.5	4.5	12.0	5.3	13.0	5.1	17.0	4.7	12.9	6.5	12.5	5.7	63.01***	.42	.55	.01	55.70***	.39	.11	.003	.85	.02	.37	.01
T-DBDRS Inattention	14.0	4.5	11.7	5.6	12.2	5.8	14.7	5.0	11.0	5.6	13.3	6.6	13.8	5.4	12.3	5.0	11.3	5.1	19.27***	.18	.89	.01	10.11**	.11	1.20	.03	2.20	.05	.30	.01
T-DBDRS Hyp/ Imp	12.7	6.4	11.1	5.5	9.3	4.9	13.1	6.6	10.0	6.2	11.5	7.0	12.4	5.1	11.6	6.0	9.1	4.0	13.29***	.13	2.65	.03	20.94***	.20	1.66	.04	4.77*	.10	.93	.02
EF- & Motiv. Behavior																														
P-BRIEF Inhibit	70.5	11.5	63.8	10.9	63.1	10.9	69.8	7.6	65.6	10.1	65.8	10.4	71.5	9.4	63.8	10.2	62.7	10.4	37.42***	.30	.51	.01	41.24***	.32	1.05	.02	.25	.01	1.85	.04
P-BRIEF Working Memory	68.0	6.8	59.1	9.6	60.9	8.9	67.1	7.7	60.3	9.3	62.4	8.3	67.8	6.8	59.5	6.6	61.1	8.1	94.93***	.53	5.82*	.06	81.63***	.49	.61	.01	.04	.001	1.21	.03
P-BRIEF Shift	58.8	10.5	53.1	8.1	51.8	10.2	58.8	11.1	54.3	11.8	53.1	10.1	64.6	9.6	56.7	8.2	54.8	8.5	40.34***	.32	4.08*	.05	60.10***	.41	1.09	.03	.09	.002	1.55	.04
P-BRIEF Emotional Control	59.4	11.0	53.1	10.5	54.0	11.0	59.2	11.1	56.7	12.9	56.5	12.2	63.4	9.3	57.2	10.6	55.7	11.4	27.45***	.24	.07	.001	25.84***	.23	1.71	.04	.56	.01	1.97	.04
P-BRIEF Initiate	58.9	9.2	53.3	10.1	53.3	10.5	58.3	6.2	53.1	9.8	54.4	8.2	62.4	9.0	54.7	9.3	56.3	10.8	45.23***	.35	1.05	.012	38.56***	.31	.73	.02	.26	.01	.66	.02
P-BRIEF Plan/ Organize	61.5	8.7	56.4	9.0	55.9	8.2	61.6	7.7	56.8	9.2	57.8	9.5	63.1	7.3	58.1	7.6	59.4	7.0	32.26***	.27	.54	.01	29.66***	.26	.01	<.001	.44	.01	.56	.01
P-BRIEF Organiz. Materials	54.5	10.0	51.8	12.4	52.0	11.1	58.5	6.2	56.5	8.2	55.1	10.4	55.8	9.5	52.6	9.9	54.5	9.8	9.87**	.10	.08	.001	6.23*	.07	.18	.004	1.19	.02	.42	.01
P-BRIEF Monitor	63.5	5.6	58.2	7.1	60.6	8.3	63.1	7.5	59.4	10.3	59.5	9.3	65.4	5.5	58.5	7.9	60.3	8.2	34.30***	.29	2.83	.03	19.72***	.19	.99	.02	.64	.02	.58	.01
P-SPSRQ Punish. Sens.	2.6	0.6	2.4	0.5	2.4	0.6	2.3	0.5	2.2	0.5	2.4	0.6	2.8	0.6	2.6	0.9	2.6	0.7	5.46*	.06	1.33	.02	2.08	.02	.14	.003	.73	.02	1.52	.03
P-SPSRQ Imp/ Fun Seeking	3.2	0.6	3.0	0.5	3.0	0.4	3.3	0.5	3.1	0.5	3.3	0.6	3.4	0.6	3.2	0.6	3.3	0.7	10.95**	.11	3.39†	.04	3.19	.04	.28	.01	1.21	.03	.36	.01
P-SPSRQ Reward Respons.	3.7	0.6	3.6	0.6	3.4	0.6	3.8	0.6	3.6	0.7	3.7	0.6	3.6	0.7	3.6	0.8	3.6	0.7	3.78†	.04	.09	.001	7.65**	.08	.66	.02	2.06	.05	1.01	.02
P-SPSRQ Drive	3.1	0.9	3.2	0.9	3.0	0.8	3.5	0.8	3.5	0.7	3.6	1.0	3.4	0.9	3.1	1.1	3.2	0.8	1.06	.01	.17	.002	1.94	.02	2.45	.06	1.30	.03	1.45	.03
Gen. Problem Behavior																														
P-DBDRS ODD	8.3	4.8	6.0	5.1	7.0	5.3	10.0	5.1	8.3	5.3	8.6	4.6	9.4	4.2	6.9	4.1	6.9	4.7	25.35***	.23	1.93	.02	20.90***	.20	.25	.01	.65	.02	1.02	.02
P-DBDRS CD	1.1	1.6	0.7	1.0	1.0	1.7	1.5	1.5	0.8	1.3	1.3	1.4	1.6	1.9	1.2	1.6	0.8	1.5	13.29***	.13	1.44	.02	5.85*	.06	.45	.01	2.63	.06	1.72	.04
T-DBDRS ODD	6.6	5.0	5.9	4.9	5.3	4.6	6.0	4.7	4.5	4.3	5.8	5.7	6.8	6.0	5.1	5.4	4.3	4.7	10.02**	.10	.01	<.001	7.44**	.08	.66	.02	1.89	.04	1.68	.04
T-DBDRS CD	1.2	1.6	1.1	1.9	1.5	2.4	1.6	2.3	1.5	2.4	1.2	2.1	1.9	2.6	1.1	1.8	1.0	1.6	3.55†	.04	.01	.001	2.73	.03	1.96	.04	.89	.02	2.70	.06
P-PEDsQL Psy. soc. Hlth.	61.8	12.1	73.1	13.9	72.6	9.1	61.0	14.1	69.0	14.3	65.3	12.9	51.3	14.5	63.8	14.9	62.2	15.6	53.48***	.38	1.84	.02	47.39***	.36	.81	.02	.87	.01	2.77†	.06
C-PEDsQL Psy. soc. Hlth.	67.2	17.3	67.8	13.5	66.9	15.1	68.3	14.1	70.0	16.2	70.7	13.5	63.7	11.7	67.2	15.3	66.3	15.4	2.88	.03	.18	.001	2.06	.02	.54	.01	.18	.004	.71	.02
P-HSQ Mean Severity Score	4.3	1.8	3.6	1.8	3.8	1.9	4.2	1.8	3.5	1.6	3.7	2.0	4.7	1.5	3.7	1.5	3.5	1.5	15.71***	.15	.12	.001	12.69**	.13	.28	.01	.55	.02	1.36	.03

Note. BRIEF = Behavior Rating Inventory of Executive Function; C- = Child-rated; CBTT = Corsi Block Tapping Task; CD = conduct disorder; DBDRS = Disruptive Behavior Disorder Rating Scale; FU = Follow-up-test (after 3 months); HSQ = Home Situations Questionnaire; Imp/Fun Seeking = Impulsivity/Fun Seeking; ODD = oppositional defiant disorder; Organiz. Materials = Organization of Materials; P- = Parent-rated; PEDsQL = Pediatric Quality of Life Inventory; Post = Post-test; Pre = Pre-test; Psy.soc. Hlth. = Psychosocial Health Summary Score; Punish. Sens. = Punishment Sensitivity; Reward Respons. = Reward Responsiveness; SPSRQ = Sensitivity to Punishment and Sensitivity to Reward Questionnaire for children; SSRT = Stop Signal Reaction Time; T- = Teacher-rated; TMT = Trail Making Task;

* $p < .05$;

** $p < .01$;

*** $p < .001$;

† $p < .075$.

doi:10.1371/journal.pone.0121651.t002

Partially-active condition versus placebo condition

A 2x3 (Treatment condition x Time) repeated measures MANOVA with the main scores of the Stoptask, STROOP, CBTT-fwd, and CBTT-bkw as dependent variables, showed a main effect of Time, F (8,220) = 3.49, p = .001, η_p^2 = .11, no main effect of Treatment condition, F (4,53) = 1.44, p = .235, η_p^2 = .10, and no significant interaction between Treatment condition and Time, F (8,220) = 1.07, p = .388, η_p^2 = .04. However, since we had specific expectations regarding the Treatment condition x Time interactions – we only expected this interaction for the Stoptask and the STROOP, not for the CBTT forward and backward (as WM was not trained in either condition) – simple contrasts were used to further explore the non-significant interaction effect:

These contrasts are presented in Table 3 and indicate that, compared to pre-test performance, post-test performance on the Stoptask improved more in the partially-active condition than in the placebo condition (p = .045; medium effect size; see Table 3 and Fig 3A). However, this difference was no longer significant after Bonferroni correction: as only p-values < .0063 (.05/8) were considered significant.

Table 3. Outcome of repeated measures MANOVAs contrasts for task performance in each combination of treatment conditions.

	Full-active vs. Placebo				Partially-active vs. Placebo				Full-active vs. Partially-active			
	Treatment*Time constrasts, *F*(1,59)				Treatment*Time constrasts, *F*(1,56)				Treatment*Time constrasts, *F*(1,57)			
	Pre vs. Post		Pre vs. FU		Pre vs. Post		Pre vs. FU		Pre vs. Post		Pre vs. FU	
Measure	***F***	η_p^2	***F***	η_p^2	***F***	η_p^2	***F***	η_p^2	***F***	η_p^2	***F***	η_p^2
Stoptask (SSRT)	5.73*	.09	2.63	.04	4.22*	.07	2.68	.05	<.001	<.001	.09	.002
STROOP (interference)	.25	.004	6.53*	.10	1.16	.02	2.26	.04	.60	.01	.39	.01
CBTT-forward	11.03**	.16	8.35**	.12	.83	.02	.77	.01	6.92*	.11	4.15*	.07
CBTT-backward	5.98*	.09	2.91	.05	.19	.003	.02	<.001	3.71†	.06	5.76*	.09

Note. CBTT = Corsi Block Tapping Task; FU = Follow-up-test (after 3 months); Post = Post-test; Pre = Pre-test; SSRT = Stop Signal Reaction Time;

* $p < .05$;

** $p < .01$;

† $p < .06$.

doi:10.1371/journal.pone.0121651.t003

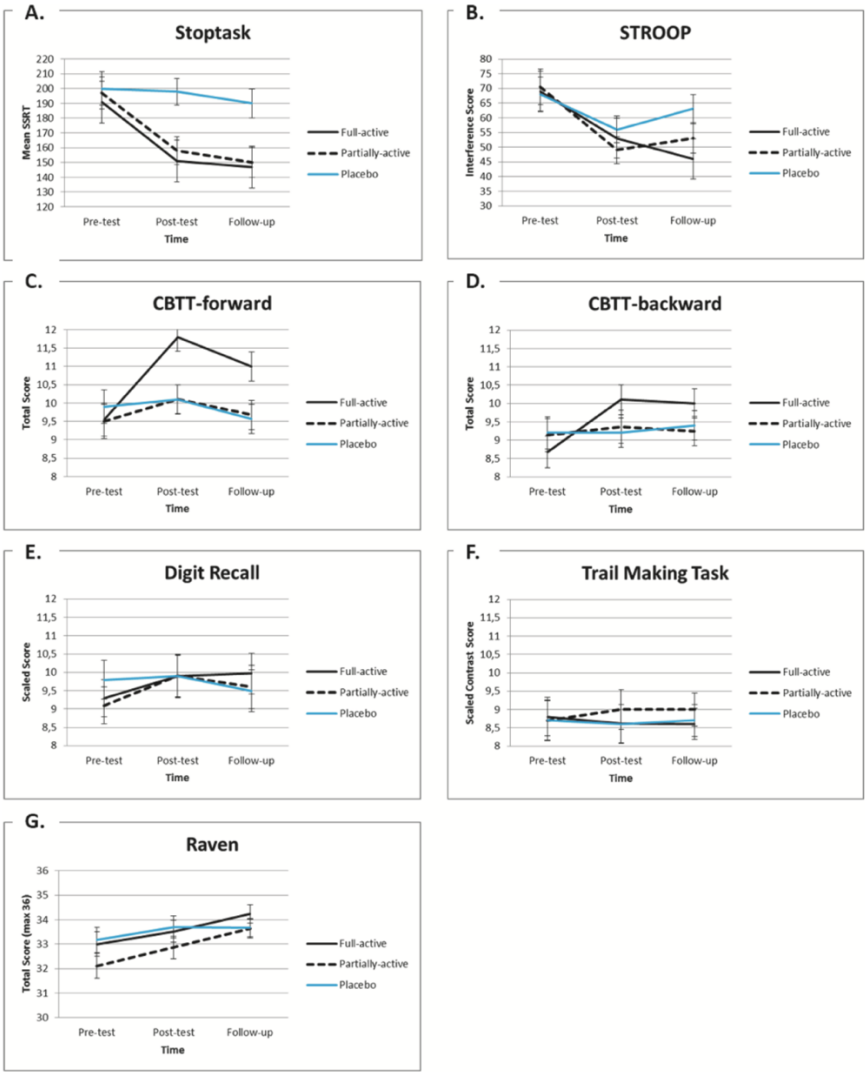

Fig 3. Mean values and standard errors of the executive functioning tasks (A—F) and the Raven (G) on the pre-test, post-test and (3 month) follow-up of children in the three treatment groups. *Note*: SSRT = Stop Signal Reaction Time; CBTT = Corsi Block Tapping Task.

doi:10.1371/journal.pone.0121651.g003

Full-active condition versus partially-active condition

A 2x3 (Treatment condition x Time) repeated measures MANOVA with the main scores of the Stoptask, STROOP, CBTT-fwd, and CBTT-bkw as dependent variables, showed a main effect of Time, $F(8,224) = 9.79$, $p < .001$, $\eta_p^2 = .26$, no main effect of Treatment condition, $F(4,54) = 1.76$, $p = .151$, $\eta_p^2 = .12$, and a non-significant trend towards an interaction between Treatment condition and Time, $F(8,224) = 2.00$, $p = .048$, $\eta_p^2 = .07$ (after Bonferroni correction only p-values < .0167 [.05/3] were considered significant). To further interpret this interaction, we used simple contrasts:

These contrasts are presented in Table 3 and indicate that, compared to pre-test performance, post-test and/or follow-up performance on the CBTT (forward and backward) improved more in the full-active condition than in the partially-active condition (p-values ranged from. 011 to. 046, effect sizes were medium; see Table 3 and Fig 3C and 3D). However, these differences were no longer significant after Bonferroni correction: as only p-values < .0063 (.05/8) were considered significant.

Within-group analyses. For each EF task where a Treatment condition x Time interaction was significant with or without Bonferroni correction (Stoptask, STROOP, CBTT-fwd, CBTT-bkw), differences within each treatment group between the pre- and post-test and between the pre-test and follow-up were tested with additional paired t-tests (Bonferroni corrections were applied: only p-values < .0021 [.05/24] were considered significant).

Results are presented in Table 4. After Bonferroni correction we found that: in the full-active condition performance on the Stoptask, the STROOP, the CBTT-fwd and the CBTT-bkw significantly improved between pre- and post-test. Performance on the Stoptask, the STROOP and the CBTT-bkw also significantly improved between pre-test and follow-up (there was a trend for performance on the CBTT-fwd, p = .003). In the partially-active condition performance on the STROOP significantly improved between pre- and post-test (there was a trend for performance on the Stoptask, p = .005), and performance on the Stoptask significantly improved between pre-test and follow-up (there was a trend for performance on the STROOP, p = .016). In the placebo condition none of the differences were significant (although there was a trend for STROOP performance between pre- and post-test, p = .043; see Table 4).

Table 4. Within-group comparisons of pair-wise time differences in task performance (using paired t-tests).

	Full-active		Partially-active		Placebo	
	Paired t-tests, t (30)		Paired t-tests, t (27)		Paired t-tests, t (29)	
Measure	Pre vs Post t	Pre vs FU t	Pre vs Post t	Pre vs FU t	Pre vs Post t	Pre vs FU t
Stoptask (SSRT)	4.29***	4.64***	3.03**	3.53**	.20	.65
STROOP (interference)	3.91***	4.47***	3.49**	2.57*	2.12*	1.22
CBTT-forward	-4.70***	-3.25**	-1.88	-.41	-.43	.83
CBTT-backward	-3.39**	-3.49**	-.59	-.34	<.001	-.38

Note. CBTT = Corsi Block Tapping Task; FU = Follow-up-test (after 3 months); Post = Post-test; Pre = Pre-test; SSRT = Stop Signal Reaction Time;
* $p < .05$;
** $p < .01$;
*** $p < .001$.

doi:10.1371/journal.pone.0121651.t004

Questionnaires and Rating Scales

ADHD behavior (parent and teacher DBDRS). A 3x3 (Treatment condition x Time) repeated measures MANOVA with mean scores on the Inattention and Hyperactivity/Impulsivity scales of the parent and the teacher version of the DBDRS as dependent variables, showed a main effect of Time, F (8,340) = 13.32, p < .001, η_p^2 = .24, no main effect of Treatment condition, F (8,166) = .33, p = .953, η_p^2 = .02, and no significant interaction between Treatment condition and Time, F (16,688) = .77, p = .718, η_p^2 = .02. The significant Time effect was further explored using simple contrasts:

For each ADHD scale, main Time effects are presented per pair-wise time difference in Table 2. After Bonferroni correction (p < .0042 [.05/12]) results indicate that: compared to the pre-test, both parents and teachers reported a significant decrease in ADHD symptoms at the post-test and at the follow-up (effect sizes of parent-ratings were large; effect sizes of teacher- ratings

ranged from medium to large). However, the non-significant Treatment x Time interaction indicates that this decrease did not differ between the Treatment conditions (in addition see Table 2 & Fig 4).

Parent-rated EF- and motivational behavior (BRIEF and SPSRQ-C). A 3x3 (Treatment condition x Time) repeated measures MANOVA with mean t-scores on the Inhibit-, Working Memory-, Shift-, Emotional Control-, Initiate-, Plan/Organize-, Organization of Materials-, and Monitor scales of the BRIEF, and the mean scores on the Sensitivity to Punishment-, Impulsivity/Fun Seeking-, Reward Responsiveness-, and Drive scales of the SPSRQ-C as dependent variables, showed a main effect of Time, $F(30,318) = 4.91$, $p < .001$, $\eta_p^2 = .32$, no main effect of Treatment condition, $F(30,146) = 1.08$, $p = .368$, $\eta_p^2 = .18$, and no significant interaction between Treatment condition and Time, $F(60,644) = .72$, $p = .942$, $\eta_p^2 = .06$. The significant Time effect was further explored using simple contrasts:

These contrasts are presented in Table 2. After Bonferroni correction ($p < .0014$ [.05/36]) results indicate the following: after training, parents reported a significant improvement (with large effect sizes) on almost all scales of the BRIEF (EF behavior; only improvement on the Organization of Materials scale was no longer significant after Bonferroni correction) and on the Impulsivity/Fun Seeking scale of the SPSRQ-C (motivational behavior; medium effect size; improvement on the Punishment Sensitivity scale [$p = .022$] and the Reward Responsiveness scale [$p = .007$] was no longer significant after Bonferroni correction). However, the non-significant Treatment x Time interaction indicates that these improvements did not differ between the Treatment conditions (in addition see the Treatment x Time contrasts in Table 2).

General problem behavior (DBDRS, PEDsQL, and HSQ). A 3x3 (Treatment condition x Time) repeated measures MANOVA with mean scores on the ODD and the CD scales of the parent and the teacher version of the DBDRS, the Psychosocial Health Summary score of the parent and the child version of the PEDsQL, and the mean severity score of the parent-rated HSQ as dependent variables, showed a main effect of Time, $F(14,334) = 5.15$, $p < .001$, $\eta_p^2 = .18$, a non-significant trend towards a main effect of Treatment condition, $F(14,162) = 1.83$, $p = .038$, $\eta_p^2 = .14$ (after Bonferroni correction only p-values $< .0125$ were considered significant), and no significant interaction between Treatment condition and Time, $F(28,676) = 1.10$, $p = .337$, $\eta_p^2 = .04$. The significant Time effect was further explored using simple contrasts:

These contrasts are presented in Table 2. After Bonferroni correction ($p < .0024$ [.05/21]) results indicate the following: after training, parents reported a significant improvement on all general problem behavior indices (effect sizes ranged from medium to large), and teachers reported a significant improvement on the ODD scale of the DBDRS (medium effect size). However, the non-significant Treatment x Time interaction indicates that these improvements did not differ between the Treatment conditions (in addition see the Treatment x Time contrasts in Table 2). In contrast to their parents, children reported no significant difference in their Psychosocial Health Summary Score after training.

Treatment Responders

In addition to the overall means, the percentage of children who benefitted from training was calculated for each measure that showed significant (with or without Bonferroni correction) main Time effects and/or Treatment condition x Time interactions on the pairwise comparisons of pre- and post-test scores and/or pre- and follow-up test scores (see Table 2). On each of these measures children were either classified as responders or non-responders by using reliable change indices [98], [99]. Based on classification guidelines by Wise [99], a participant was classified as responder when both the following criteria were met: (1) a reliable change index (RCI) of at least 1.28 (RCI was based on the method of [98]), and (2) an improvement of scores of at least 1 standard deviation [99]. Results for each treatment condition are presented in Table 5. The pattern of these results strongly resembles the pattern of the mean results (see Table 5; in addition see Table A in S2 Appendix).

Table 5. Proportion of treatment groups showing improvement on performance measures and rating-scales (i.e., responders).

	Full-active Condition		Partially-active Condition		Placebo Condition	
Domain and Measure	**Pre vs. Post** *% responders*	**Pre vs. FU** *% responders*	**Pre vs. Post** *% responders*	**Pre vs. FU** *% responders*	**Pre vs. Post** *% responders*	**Pre vs. FU** *% responders*
Performance Measures						
Stoptask (SSRT)	***41.9***	**54.8**	***35.7***	***35.7***	16.7	20.0
STROOP (interfence score)	9.7	22.6	21.4	10.7	14.6	6.7
CBTT-forward (total score)	***48.4***	***41.9***	17.9	17.9	13.3	13.3
CBTT-backward (total score)	***38.7***	29.0	17.9	3.6	10.0	16.7
Raven (total score)	12.9	19.4	10.7	14.3	16.7	16.7
ADHD Behavior						
P-DBDRS Inattention	**51.6**	***48.4***	**53.6**	***32.1***	***50.0***	***50.0***
P-DBDRS Hyp/Imp	***45.2***	***38.7***	***50.0***	***42.9***	***40.0***	***46.7***
T-DBDRS Inattention	29.0	***32.2***	***42.9***	28.6	23.3	30.0
T-DBDRS Hyp/Imp	16.1	***38.7***	25.0	25.0	20.0	30.0
EF- & Motiv. Behavior						
P-BRIEF Inhibit	***35.5***	***35.5***	25.0	25.0	26.7	***43.3***
P-BRIEF Working Memory	**54.8**	***41.9***	***39.3***	***32.1***	**56.7**	***50.0***
P-BRIEF Shift	25.8	***32.3***	17.9	25.0	***40.0***	***50.0***
P-BRIEF Emotional Control	***32.3***	***32.3***	14.3	21.4	***40.0***	***36.7***
P-BRIEF Initiate	***35.5***	***38.7***	***50.0***	***39.3***	***43.3***	30.0
P-BRIEF Plan/Organize	29.0	29.0	28.6	***35.7***	***33.3***	***33.3***
P-BRIEF Organiz. Materials	22.6	19.4	25.0	28.6	23.3	16.7
P-BRIEF Monitor	***48.4***	***48.4***	25.0	***32.1***	**53.3**	***46.7***
P-SPSRQ Punish. Sens.	12.9	9.7	3.6	3.6	23.3	13.3
P-SPSRQ Imp/Fun Seeking	22.6	16.1	25.0	14.3	10.0	13.3
P-SPSRQ Reward Respons.	16.1	22.6	21.4	10.7	10.0	3.3
Gen. Problem Behavior						
P-DBDRS ODD	***32.3***	19.4	10.7	10.7	30.0	20.0
P-DBDRS CD	6.5	3.2	25.0	14.3	10.0	13.3
T-DBDRS ODD	6.5	***32.3***	21.4	17.9	16.7	20.0
P-PEDsQL Psy.soc. Hlth.	**51.6**	***48.4***	25.0	21.4	***40.0***	***36.7***
P-HSQ Mean Severity Score	22.6	25.8	14.3	21.4	30.0	30.0

Note. BRIEF = Behavior Rating Inventory of Executive Function; CBTT = Corsi Block Tapping Task; CD = conduct disorder; DBDRS = Disruptive Behavior Disorder Rating Scale; FU = Follow-up-test (after 3 months); HSQ = Home Situations Questionnaire; Imp/Fun Seeking = Impulsivity/Fun Seeking; ODD = oppositional defiant disorder; Organiz. Materials = Organization of Materials; P- = Parent-rated; PEDsQL = Pediatric Quality of Life Inventory; Post = Post-test; Pre = Pre-test; Psy.soc. Hlth. = Psychosocial Health Summary Score; Punish. Sens. = Punishment Sensitivity; Reward Respons. = Reward Responsiveness; SPSRQ = Sensitivity to Punishment and Sensitivity to Reward Questionnaire for children; SSRT = Stop Signal Reaction Time; T- = Teacher-rated; **Bold + italic formatted number** = more than 30% responders; **Bold formatted number** = more than 50% responders; Children were classified as responders based on reliable change indices [98], [99].

doi:10.1371/journal.pone.0121651.t005

Discussion

The aim of this study was to determine the short- and long-term effects of a gamified training intervention (BGB) that targets multiple EFs (visuospatial WM, response inhibition and cognitive flexibility) compared to a placebo version of the intervention on various outcome measures in children with ADHD combined-type. In addition, to determine the unique effect of the inhibition and cognitive flexibility training tasks, we compared a full-active condition (where WM, inhibition, and cognitive flexibility were all in training-mode) to a partially-active condition (where only inhibition and cognitive flexibility were in training-mode).

Results indicated that only children in the full-active condition showed improvement on measures of visuospatial STM and WM. Inhibitory performance and interference control only improved in the full-active condition and the partially-active condition. However, no Treatment-condition x Time interactions (with or without Bonferroni corrections) were found for cognitive flexibility, verbal STM and WM, non-verbal complex reasoning, or child-rated psychosocial health, nor for any parent- or teacher-rated ADHD symptoms, EF behaviors, motivational behaviors, or general problem behaviors. Nonetheless, almost all measures showed significant Time-effects, including the teacher-ratings (effect sizes ranged from medium to large).

These findings suggest that improvements on inhibition and visuospatial STM and WM were specifically related to the type of treatment received. However, improvements on untrained EFs and behavior (far transfer effects) were mostly nonspecific (i.e., only interference control improved exclusively in the two conditions where EFs were trained). As such, in this multiple EF training, mainly nonspecific treatment factors – as opposed to the specific effects of training EFs – seem related to the far transfer effects on EF and behavior.

In many ways our findings are similar to those of previous placebo controlled (single) EF training studies in children with ADHD [21], [84], [85], [86], [23] (but note that only one of these studies [21] corrected for multiple testing). Most of these studies find differential treatment effects on outcome measures of trained EFs (although Kray et al. [23], as in the present study, found no significant differences on cognitive flexibility). However, such near transfer effects may not be surprising since many of these outcome measures are very similar to the training tasks themselves and improvement may be the result of a learned strategy instead of improved cognitive capacity [87]. Further, in most placebo controlled studies differential far transfer to untrained EF tasks has been limited, and differential effects on parent- or teacher-rated behavior (e.g., ADHD or EF) are generally not found. Only Klingberg et al. [86] found a differential effect of WM training on parent-rated ADHD. However, the placebo condition used in Klingberg et al. was considerably shorter in time than the training condition. This suggests a difference in parent involvement between the conditions, which may have interacted with the outcome of parent-rated ADHD behavior (e.g., through expectancy effects or inequality of parent-child interactions; see [15]). Another notable feature of the study of Klingberg et al. is that they did not include children with comorbid ODD. However, including ODD diagnosis as a covariate did not change the pattern of our main results. Therefore, the absence of comorbid ODD in the Klingberg et al. study seems an unlikely explanation for their distinctive findings on parent-rated ADHD. This assumption is further substantiated by the findings we presented in Table A (see S2 Appendix): Irrespective of treatment condition, children with comorbid ODD were at least as likely to improve on parent-rated ADHD behavior as children without comorbid ODD.

There are also several important differences between our findings and the findings of previous placebo controlled EF training studies. Although we used more stringent compliance criteria than most previous studies (i.e., completing 100% of the training sessions versus completing 80% of the training sessions), in our study only 3% of the participants failed to meet compliance

criteria, whereas in previous studies 15–23% failed to meet compliance criteria. Since most previous studies also used an external reward system, a structured schedule for implementing the intervention, weekly contact with a coach, and performance feedback during training, the most obvious reason for this difference in compliance is the relatively strong gamification of BGB. This hypothesis is consistent with previous findings of increased time-on- training when EF training was gamified [45] (also see [40]), and with the finding that gaming increases the release of striatal dopamine [47], [48], which is associated with increased motivation to continue playing and performing [50].

Moreover, in contrast to the previous placebo-controlled studies, we found a significant improvement on teacher-rated ADHD behavior (effect sizes ranged from medium to large). Although this improvement was unrelated to specific effects of the EF training (as it was also found in the placebo condition), it is still a remarkable finding. Some have argued that EF training studies only find Time effects on parent-ratings but not on teacher-ratings because teachers, in contrast to parents, are only minimally involved in training and thus may be less biased than parents (e.g., by their expectancies of the training outcome) [31]. This suggests that generalization of improvement to teacher-ratings might represent relatively unbiased evidence of treatment induced changes in the child's behavior. Nonetheless, it is unclear what caused this improvement. It seems unrelated to specific EF training effects, and the only nonspecific treatment factor that clearly distinguishes our study from previous studies appears to be the use of relatively strong gamification (i.e., teachers were not more involved than in previous studies). Is it possible that gamification somehow improved classroom behavior? For example, there is evidence that video game playing can enhance various cognitive skills (e.g., attention; see [88]). However, if playing video games by itself would be sufficient to improve classroom functioning in children with ADHD, it seems illogical that the participants in our study, who play commercial video games for 10 hours per week (see Table 1), did not improve before. Nonetheless, it may be that parents' positive attitude towards this particular game enhanced its positive effects. For example, sharing the joy of achievement in the game with his/her parents could have enhanced the child's appraisal of the game's positive feedback and its effect on his/her self-esteem beyond that of commercial video games (as many parents don't encourage children to indulge in commercial gaming). Although there is a link between parental praise and children's self-esteem [89], and self-esteem has been found to mediate the relationship between ADHD and classroom functioning [90], future research should investigate this further. Furthermore, the gamification of BGB may also have impacted classroom functioning by enhancing children's motivation to comply with treatment. If children were more motivated to comply with treatment than in other EF training studies, which is consistent with the relatively high compliance rate in our study, there may have been less need for parents to discipline their children during training. Evidence suggests that decreased negative parental discipline mediates the effect of ADHD treatment (e.g., medication and behavior therapy) on teacher-rated ADHD behavior [91]. Future EF training studies should use larger samples and appropriate process measures to further investigate these potential mechanisms of mediation.

Although some previous EF training studies in children with ADHD have found differential effects on interference control [23], [30] ([85] and [86] also found differential effects on the

STROOP, but they only used the incongruent trials as outcome measure; baseline response times to congruent trials were not controlled for, making it impossible to calculate the interference score), our study is the first to find differential effects on response inhibition. In contrast to the placebo condition, response inhibition was improved in both the full-active condition and the partially-active condition, but no differences were found between these two experimental conditions. This suggests that a combined inhibition and cognitive flexibility training by itself (i.e., without WM) is sufficient to improve response inhibition in children with ADHD. Possibly, previous EF training studies investigating effects on measures of response inhibition in children with ADHD [29], [30], [32] found no improvements because their intervention did not include an inhibition training task (i.e., Hoekzema et al. [32] trained WM, cognitive flexibility, attention, planning and problem solving), or because their inhibition training task was based on a less appropriate response inhibition paradigm; the go/no-go task instead of the stop task [29], [30]. In contrast to the stoptask, the go/no-go task has been criticized as not function- ally isolating inhibition (e.g., because of its interaction with selective attention and decision making, and the confounding effects of its prepotent response processes; see [2], [92], [93]). Nonetheless, since we did not investigate effects of the inhibition- and cognitive flexibility training separately, we can only speculate that the improvement on response inhibition was the result of our stop-task-based inhibition training. Additional research is needed to investigate this in more detail.

In contrast to our findings on other near transfer measures, no differential effects of EF training were found on the cognitive flexibility measure (neither with or without Bonferroni correction). This may be the result of the difference between the switch-cost (the index of cognitive flexibility) that was trained, and the switch-cost that was used as outcome measure of cognitive flexibility. Our outcome measure (the scaled contrast score on the TMT) measures global switch-cost (i.e., the difference between a block of switch-trials and blocks of non-switch trials), whereas the cognitive flexibility training focused on training local switch-cost (i.e., the difference between switch-trials and non-switch trials within a block of trials). Although, both types of switch-cost are considered valid measures of cognitive flexibility, evidence suggests that they tap somewhat different cognitive processes and can be differentiated on a neural level [94], [95]. Therefore, it could be argued that our outcome measure of cognitive flexibility was in fact a measure of far transfer. Future studies should investigate this further using more varied measures of cognitive flexibility.

The fact that far transfer was also found in the placebo condition might not (only) be explained by nonspecific treatment effects (e.g., effects of expectancies, self-fulfilling prophecies, attribution, gamification, or improved parent-child interactions), but may be the result of actual cognitive training in the placebo condition. Although the cognitive load in our placebo condition was very low, it could be argued that the requisite of the placebo tasks to focus attention for a substantial amount of time was sufficient to improve cognitive control (e.g., attention) and the behavior of our participants. However, this appears inconsistent with the very limited improvement on EF performance in the placebo condition, and the lack of effects resulting from other activities that require prolonged focused attention (e.g., paying attention in school, playing [educational] video games).

Because no wait-list control condition was utilized, it is not possible to determine to which extent our findings relate to effects of multiple testing, the passage of time, or (nonspecific) treatment factors. However, a previous study investigating BGB [31] found no improvement on parent- and teacher-rated ADHD and EF behavior in a wait-list control group, whilst they did find improvement in the group that was trained. This suggests that the current findings on ADHD and EF behavior are probably not attributable to mere passage of time or multiple testing (for a study of children with autism spectrum disorder see [100]).

In this study different EFs were trained simultaneously within the same training session. However, based on the current state of the literature it is unclear if this is indeed the best strategy for multiple EF training (i.e., there are no studies that directly investigate this). One could assume that training different EFs simultaneously is more effective (especially for transfer to daily life) than training one EF at a time (i.e., training each EF in separate sessions), because functioning in daily life also requires the use of multiple EFs at once. However, our results do not suggest that training three EFs per session (i.e., the full-active condition) has more effect on daily functioning than training two EFs per session (i.e. the partially-active condition). Future studies should further investigate this.

In the current study, far transfer effects were mostly nonspecific. However, we only investigated overall group differences (i.e., disregarding potential subpopulations that show differential responses to treatment), and children were allocated to treatment conditions irrespective of their individual EF deficits. Therefore, before discarding EF training as potential treatment for children with ADHD, future studies should examine moderators (e.g., severity of EF deficits; teacher expectancies) and mediators of treatment success (e.g., improvement on EF performance; parental praise), and should investigate effects of individually tailored EF training (i.e., to make optimal use of the available training-time future studies should match training focus to the specific EF problems of each individual child). Furthermore, to increase chances of finding far transfer that results from EF training specifically, training tasks should be made more ecologically valid (e.g., by using EF training tasks that resemble the complexity of problematic situations in daily-life) and should be intertwined with relevant real-life EF-taxing activities (e.g., completing chores in daily-life could be an additional goal in the EF training; for more suggestions see [96]). Finally, the domains of far transfer that were investigated in this study were limited to direct measures of performance and indirect measures of behavior (e.g., behavior as rated by parents, teachers or children). Future studies should also include direct measures of behavior. For example, a recent placebo-controlled WM training study [84] found no specific treatment effects on parent-rated behavior (teacher-rated behavior was not investigated), but found specific effects on aspects of experimenter-observed off-task behavior during an academic task.

In conclusion, our findings suggest that improvements on inhibition and visuospatial STM and WM were specifically related to the type of treatment received. However, improvements on untrained EFs and behavior were mostly nonspecific. As such, in this multiple EF training (BGB), mainly nonspecific treatment factors—as opposed to the specific effects of training EFs—seem related to the far transfer effects on EF and behavior.

Annexe 4

Can task-switching training enhance executive control functioning in children with attention deficit/-hyperactivity disorder?

Abstract

Introduction

The key cognitive impairments of children with attention deficit/-hyperactivity disorder (ADHD) include executive control functions such as inhibitory control, task-switching, and working memory (WM). In this training study we examined whether task-switching training leads to improvements in these functions.

Methods

Twenty children with combined type ADHD and stable methylphenidate medication performed a single-task and a task-switching training in a crossover training design. The children were randomly assigned to one of two groups. One group started with the single-task training and then performed the task-switching training and the other group vice versa. The effectiveness of the task-switching training was measured as performance improvements (relative to the single-task training) on a structurally similar but new switching task and on other executive control tasks measuring inhibitory control and verbal WM as well as on fluid intelligence (reasoning).

Results

The children in both groups showed improvements in task-switching, that is, a reduction of switching costs, but not in performing the single-tasks across four training sessions. Moreover, the task-switching training lead to selective enhancements in task-switching performance, that is, the reduction of task-switching costs was found to be larger after task-switching than after single-task training. Similar selective improvements were observed for inhibitory control and verbal WM, but not for reasoning.

Conclusions

Results of this study suggest that task-switching training is an effective cognitive intervention that helps to enhance executive control functioning in children with ADHD.

Introduction

The main goal of the present study was to determine the range of plasticity in executive control functioning in children with attention deficit/-hyperactivity disorder (ADHD). Executive control can be defined as a set of higher-order cognitive functions that organize and regulate goal-directed behavior including processes of planning, interference control, working memory (WM), task-switching, and task coordination (e.g., [1]). Behavioral deficits observed in children with ADHD are characterized by inattention, impulsivity, and hyperactivity ([2]), and it has been suggested that those deficits are primarily related to executive control impairments, such as inhibitory control and WM ([3], [4]).

One experimental task that has frequently been applied in recent years to examine executive control functioning is the task-switching paradigm (for a recent review; [5]). The advantage of this paradigm is that it allows the separation of different components of executive control, such as task-set selection and maintenance, task-set switching, and interference control ([6]). In this type of task, the participants are usually instructed to switch between two simple cognitive tasks. For example, the participants are presented ambiguous stimuli, such as a series of digits varying in number and value (1, 3, 111, 333). In one task (task A), they have to decide whether the value of digits is one or three, and in the other task (task B), whether the number of digits is one or three. Performance can be measured in mixed-task blocks, in which the participants have to switch between both tasks A and B on every second trial, and in single-task blocks, in which only one of the tasks (A or B) has to be performed ([7], [8]). This allows the determination of two types of performance costs associated with the switching situation: mixing costs are defined as the difference in mean performance between mixed-task and single-task blocks and are assumed to refer to the ability to maintain two task sets and select between them. Switching costs are defined as the difference in mean performance between switch and non-switch trials within mixed-task blocks and they are assumed to measure the ability to flexibly switch between tasks (cf. [7], [8]). Finally, the efficiency of interference control can be measured by comparing the performance on congruent trials (in which the number and value decisions are not conflicting, i.e., 1, 333) with performance on incongruent trials (in which the number and value decisions are conflicting, i.e., 3, 111), that is, interference costs can be defined as the difference in mean performance between incongruent trials and congruent trials. Cepeda et al. (2000) examined switching and interference costs in ADHD children (6–12 years old), on and off medication, in comparison to children without ADHD that were matched by age and

IQ. Results of this study revealed that only ADHD children off medication showed larger switching costs and interference costs than healthy controls but there were no performance differences in these costs between ADHD children on medication and the control children. Moreover, switching costs in ADHD children off medication were only larger on incongruent trials, suggesting that children with ADHD particularly had problems to inhibit irrelevant task information when switching from one task to the other one ([9]).

Given that children diagnosed with ADHD usually achieve lower academic degrees compared to equally cognitively able children without ADHD, and also have major problems in everyday functioning until adulthood ([10]), the question of effective treatments, such as cognitive training interventions that help to improve executive control functioning, is of high relevance. One desirable feature of cognitive training interventions is that the training program does not only result in an improvement on the trained task, but that it also transfers to tasks that were not part of the training intervention ([11]). To determine the scope of transfer, we distinguish between near and far transfer. Near transfer refers to a generalization of training-related improvements to a new but structurally similar transfer task (e.g., transfer of task-switching training to another switching task, [12]; [13]), while far transfer refers to dissimilar theoretical constructs (e.g., transfer of task-switching training to a WM task; cf. [14]).

In a recent lifespan study, we investigated near and far transfer of task-switching training in children, younger, and older adults with a pretest–training–posttest design ([14]). Pretest and posttest consisted of a cognitive test battery including several tests measuring task-switching (near transfer), interference control, verbal and visual WM, and fluid intelligence (far transfer). Importantly, we included an active control group in this study. Transfer was defined as relative performance improvements at posttest in the treatment group (task-switching training) as compared to the control group (single-task training). Note that both groups performed the identical number and type of A and B tasks, but the control group performed them in separate blocks (single-task training), while the training group switched between both tasks on every second trial (task-switching training). Results indicated that (a) all three age groups showed near transfer effects, that is, a larger reduction of mixing and switching costs from pretest to posttest in the training group than in the control group; (b) near transfer effects were more pronounced in children and older adults than in younger adults; and (c) far transfer effects were observed in all age groups, that is, performance improvements in interference control, verbal and visual WM, and fluid intelligence. The effect sizes for the group of children were between $d' = 1.2–2.1$ for near transfer and $d' = 0.5–0.9$ for far transfer of task-switching training. Given

these promising effects of the cognitive training intervention in healthy children, the specific aim of the present study was to examine whether the training is of similar effectiveness in a group of children with substantial impairments in executive control.

There are a few studies demonstrating that training of executive control in children with ADHD leads to near as well as far transfer effects. Klingberg et al. (2002, 2005) used an adaptive training procedure including visuospatial and verbal WM tasks. They found performance improvements not only on the trained visuospatial WM task but also on non-trained tasks assessing visual–spatial memory, fluid intelligence (the Raven's), and interference control. More recently, Shalev et al. (2007) applied an attentional training program in order to improve school performance (e.g., math exercises, reading comprehension) and behavior (parents' self-reports of ADHD symptoms) in ADHD children (6–13 years old). The attentional training included the practice of sustained and selective attention, orienting of attention, and executive attention. The authors found training-related improvements in school performance as well as a reduction of inattention symptoms reported by the parents. Although these far transfer effects are impressive, it should be noted that the authors did not report the improvements on the trained tasks and they did not measure near transfer effects. Kerns et al. (1999) used a similar attentional training including seven ADHD children (7–11 years old) and reported far transfer effects to a number of attentional tasks that were not trained during the intervention.

The main goal of the present study was to determine the transfer scope after task-switching training in ADHD children. Therefore, we investigated near and far transfer effects of this training, similar to a previous study ([14]). For ethical reasons (see also Procedure), we applied a crossover training design so that all ADHD children performed the cognitive intervention (i.e., the task-switching training) that has already been shown to enhance executive control functioning in healthy young children. However, they received the treatment at different times during the training protocol. That is, after performing the pretest, the children were randomly assigned to one of two groups: group 1 first performed the single-task training followed by posttest 1 and then the task-switching training followed by posttest 2 (see Table 1). Group 2 first performed the task-switching training as well as the first posttest and then the single-task training and the second posttest.

On the basis of previous results showing near and far transfer effects of WM and attentional control training in children with ADHD ([15a]; [15b]; [16]) as well as near and far transfer effects of task-switching training in healthy young children ([14]), we expected treatment-

specific effects in this study. In particular, we predicted a larger reduction of mixing and switching costs after the treatment (task-switching training) than after the single-task training (near transfer) as well as larger improvements in executive control and fluid intelligence measures (far transfer). That is, group 2 should show larger performance improvements from pretest to posttest 1 as compared to group 1, and group 1 should show larger improvements from posttest 1 to posttest 2 than group 2. Given that far transfer effects are usually the smaller the less similar the training and the transfer tasks are, we also expected larger effect sizes for near than for far transfer.

Materials and methods

Participants

Thirty children were recruited for this study. Ten participants had to be excluded from the analysis because they were not willing to finish the training study (n = 7) or went off medication during the study (n = 3). Given that ADHD is more common in boys than girls ([17]), we included only male children. The final sample consisted of 20 boys that were randomly assigned to one of the two training conditions (group 1: n = 10, group 2: n = 10). Both groups were comparable in terms of age (p= 1.00; group 1: range= 8.7–12.1 years; group 2: range= 7.7–11.6 years) and IQ (p= 0.44). The severity of the ADHD-related symptoms was assessed by means of the German parent rating scale FBB–HKS ([18]). The questionnaire is based on the DSM-IV and ICD-10 criteria for ADHD and hyperkinetic disorders and allows the assessment of behavioral symptoms on the four scales (1) severity of inattention, (2) severity of hyperactivity/impulsivity, (3) generalized inattention problems, and (4) generalized hyperactivity/impulsivity problems. We found no between-group differences on any of the four scales (all *ps* > 0.53). Means and SD for age, IQ, and parent ratings are provided in Table 2.

All participants were enrolled in mainstream elementary and secondary schools. Prior to the inclusion into the study, they had been diagnosed according to the guidelines of DSM-IV ([2]) at the Department of Child and Adolescent Psychiatry, Saarland University Hospital, Germany.

The diagnosis was based on a structured interview (K-DIPS, [19]), an intelligence assessment (WISC-IV, [20]), and standard rating scales (such as the FBB–HKS, [18]) administered by expert physicians and psychologists.

After being diagnosed, the children had been medicated with methylphenidate. Although individual doses varied as a function of body weight and severity of the symptoms, most of the boys (n = 18) were prescribed 10–20 mg/day and two older children (10–11 years of age) up to 40 mg/day. Prior to the inclusion into the study, an independent physician assessed the effectiveness of the medication.

In sum, we applied the following inclusion criteria: (a) diagnosis of ADHD combined subtype, (b) age between 7 and 12 years, (c) stable long-term medication (methylphenidate), and (d) an IQ > 80 as measured with the Kaufmann Assessment Battery for Children (K-ABC; [21]). Exclusion criteria were (a) maternal drug abuse in pregnancy, (b) premature birth (<32 weeks) and low birth weight (<2000 g), (c) enrollment in special education settings, (d) neurological or chronic internal diseases, (e) Autism Spectrum, psychotic, bipolar, severe anxiety, and depressive disorder, and (f) any treatment with psychotropic drugs besides methylphenidate. The ethics review board of the Saarland Medical Association approved this training study. Written informed consent was given by one of the parents for all participating children. Subjects were paid 60 EUR for participating in the study.

Table 1 | Outline of the training protocol.

Pretest session 1	Training sessions 2–5	Posttest 1 session 6	Training sessions 7–10	Posttest 2 session 11
BOTH GROUPS	**GROUP 1**	**BOTH GROUPS**	**GROUP 1**	**BOTH GROUPS**
Single-tasks (tasks A and B)	Single-task training	Single-tasks (tasks A and B)	Task-switching training	Single-tasks (tasks A and B)
Task-switching (tasks A and B)	(tasks C and D)	Task-switching (tasks A and B)	(tasks C and D)	Task-switching (tasks A and B)
COGNITIVE BATTERY	**GROUP 2**	**COGNITIVE BATTERY**	**GROUP 2**	**COGNITIVE BATTERY**
Stroop task	Task-switching training	Stroop task	Single-task training	Stroop task
Verbal working memory	(tasks C and D)	Verbal working memory	(tasks C and D)	Verbal working memory
Fluid intelligence		Fluid intelligence		Fluid intelligence
Control measures		Control measures		Control measures
Demographic questionnaire				

Table 2 | Mean (SD) age, IQ, and sum scores on the FBB–HKS parent rating scale as a function of group at pretest.

	Group 1 (single-task training first)		Group 2 (task-switching training first)	
	M	**SD**	***M***	**SD**
Age	10.1	1.2	10.1	1.3
IQ	107	14	103	11
FBB–HKS: severity of inattention	13.8	4.6	14.4	7.7
FBB–HKS: severity of hyperactivity/impulsivity	15.2	7.4	13.7	5.4
FBB–HKS: generalized inattention problems	14.3	5.5	12.8	6.2
FBB–HKS: generalized hyperactivity/impulsivity problems	9.9	5.0	9.0	5.5

FBB–HKS scores are based on 20 items describing behavioral problems associated with ADHD and its subjective experienced severity. Parents were to rate the statements on a scale from 0 (not at all) to 3 (very much). Higher values correspond to more severe symptoms.

Procedure

For ethical reasons, all children in this study performed the training intervention (i.e., the task-switching training) but at different times during the training protocol. Therefore, transfer of task-switching training was assessed by means of a pretest–training–posttest 1–training–posttest 2 design (see Table 1). To determine the transfer scope, the pretest and posttest sessions consisted of a structurally similar, but new switching task and a battery of several cognitive tasks that are assumed to measure executive control as well as fluid intelligence. All training conditions included four sessions, each of them lasting about 30–40 min. The training protocol was carried out over a time period of 11 weeks, that is, the children performed approximately one session per week, similar to the training protocol of our previous training study ([14]). Three expert experimenters (one psychologist and two research assistants) administered the tests and experimental tasks. They were randomly assigned to the test sessions.

Pretest and posttest assessment

Task-switching. We used the same task-switching paradigm as in one of our previous training studies (cf. [14]). In this type of paradigm, the participants worked through single-task blocks (i.e., performing task A or B only) and through mixed-task blocks requiring the switching between both tasks A and B on every second trial. Participants received no task cues and had to keep track of the task sequence. In task A, participants were to decide whether a picture showed a fruit or a vegetable ("food" task), and to respond by pressing a left or right response key, respectively. In task B, they were to decide whether the picture was small or large ("size" task) and they also responded with a left or right response key. The same two response keys were used for both tasks and all stimuli were ambiguous. Stimuli consisted of 16 fruit and 16 vegetable pictures and each one of them was presented in a large and a small version.

Children first performed two single-task practice blocks (each consisting of 17 trials) and then worked through 20 experimental blocks (8 single-task and 12 mixed-task blocks, each consisting of 17 trials). The order of blocks was random with the constraint that two single and two mixed-task blocks were grouped together. At the beginning of each trial, a fixation cross appeared for 1400 ms, followed by the target that was presented until the subject responded. After 25 ms, the next fixation cross appeared. The children were instructed to respond as fast and as accurately as possible. After each block, subjects received feedback about their mean speed and accuracy of responding.

Cognitive test battery. The cognitive test battery included several experimental tasks and tests measuring executive control (inhibitory control, verbal working memory) and fluid intelligence. The pre- and post-test assessment took about 60–70 min. We applied a modified version of the "Color-Stroop Task" ([22]). In this version, children were shown words (e.g., "red," "tree") presented in red, blue, green, or yellow font successively on the computer screen. The color words were presented either in the congruent color or in an incongruent color. Children were to indicate the color of the words as quickly as possible by pressing one of four response buttons. Participants first performed two practice blocks (12 trials) and then four experimental blocks (24 trials). Stimuli were presented until the subject responded or for a maximum of 2000 ms. The time window between the response and the next stimulus was 700 ms. The Stroop interference effect was defined as the difference in mean performance between incongruent and congruent trials. Verbal WM was assessed with the test "Digit Sorting" (cf. [7]). In this test, the experimenter read aloud a series of digits ranging in value from 1 to 20. The participants were to repeat the digits by sorting them in numerical order. The number of digits in each series varied between three and seven. Children first performed three practice series à three digits. The test started with three series à three digits, and then the number of digits per series was increased by one after each third series. The task was aborted after three consecutive erroneous responses. The test score was the number of correctly solved items.

We applied the matrix reasoning test from the German version of the Wechsler Intelligence Scale for Children (WISC-IV; [20]). In this test, the children were presented with a partially filled grid and asked to select the item that properly completed the matrix. Participants first completed three practice items, followed by up to 35 test items. The task was aborted after four consecutive erroneous responses or if four out of five consecutive items were not successfully completed. The test score was the number of correct responses.

In addition, we included two control measures on which we expected no positive transfer of the switching training. As a measure of perceptual speed of processing, we applied the Digit–Symbol Substitution test ([23]). Children saw a template containing nine digit–symbol mappings on the top of the page. Below, they saw 100 digits without the corresponding symbols. They were instructed to fill in the correct symbols as fast as possible. The test score was the number of correctly completed symbols after 90s. As a measure of semantic knowledge, we used the Spot-a-Word test ([24]). In this test, 35 items are presented successively on the computer screen. Each item contains one correct word and four non-words. The participants were asked to find the one correct word. The test score was the number of correctly identified

words. The order of cognitive tasks and tests was constant during pre- and post-test assessment and were applied in the following order: Digit–Symbol Substitution Test, Task-Switching, Color-Stroop Task, Digit Sorting, Wechsler Intelligence Scale, and Spot-a-Word Test.

Training intervention. In the single-task training sessions, the children performed single-task blocks including either task C or task D. In the task-switching training sessions, the participants performed mixed-task blocks, that is, they were instructed to switch between both tasks C and D on every second trial. The experimental procedure during the training intervention was identical to the one applied at pretest and posttest except that children performed different tasks (tasks C and D). In task C ("transportation" task), subjects were to decide whether the pictures showed planes or cars, and in task D, ("number" task) whether one or two planes/cars were presented. They started with two practice blocks (eight trials) followed by 24 experimental blocks (17 trials). In single-task training sessions, the children also started with two practice blocks (eight trials) and then performed 24 experimental blocks (à 17 trials; 12 blocks of task C; and 12 blocks of task D in an alternating sequence). Overall all children worked through 1696 training trials in each training condition.

Data analysis

Analyses for the switching and the Color-Stroop tasks were focused on mean RT for correct responses. We also analyzed response accuracy (% errors) but consistent with previous data, we found no transfer of training ([14]). Practice blocks and the first trial in each block were excluded from data analyses. For all remaining tasks of this study, the analyses were based on accuracy (number of correct responses). In order to test for between-group training effects, we run analyses of variance (ANOVA) with the between-subjects factor Group (group 1: single-task first, group 2: switching first). For the evaluation of transfer effects, we also calculated Cohen's *d* ([25]), or the standardized mean difference in performance between pretest and posttests ([26]). That is, the pretest–posttest differences (for each of the two groups) were divided by the pooled SD for test occasions. We then corrected all *d* values for small sample bias using the Hedges and Olkin correction factor (*d'*; [27]).

Results

Training effects

To test for between-group differences in training-related benefits, we ran two ANOVAs, the first one for the single-task conditions and the second one for the task-switching conditions. Figures 1A,B show the latencies a function of Training Session (1–4) and experimental Group (group 1, group 2).

Single-task training. Training-related changes were analyzed with Group (group 1: single-task first, group 2: switching first) as between-subjects factor and Session (S1, S2, S3, S4) as within-subjects factor. Results showed a main effect of Session, $F(3, 54) = 5.65$, $p < 0.01$, $\eta_p^2 = 0.24$, with a quadratic slope, $F(1, 54) = 16.99$, $p < 0.001$, $\eta_p^2 = 0.49$, indicating that latencies increased from session 1 to sessions 2 and 3 but decreased again in session 4 (see Figure 1A). Neither the main effect of Group nor the interaction with Session was significant.

Task-switching training. The ANOVA including the between-subjects factor Group (group 1: single-task first, group 2: switching first) and the two within-subjects factors Session (S1, S2, S3, S4) and Trial Type (non-switch, switch) showed a main effect of Session, $F(3, 54) = 5.05$, $p < 0.01$, $\eta_p^2 = 0.22$, indicating that latencies decreased as a function of training, and a main effect of Trial Type, $F(1, 18) = 23.84$, $p < 0.001$, $\eta_p^2 = 0.57$ (switching costs). An interaction between Session and Trial Type indicated that switching costs were reduced as a function of training, $F(3, 18) = 4.21$, $p = 0.01$, $\eta_p^2 = 0.19$ (see Figure 1B). Neither the main effect nor the interactions with the factor Group reached significance (all $p > 0.46$).

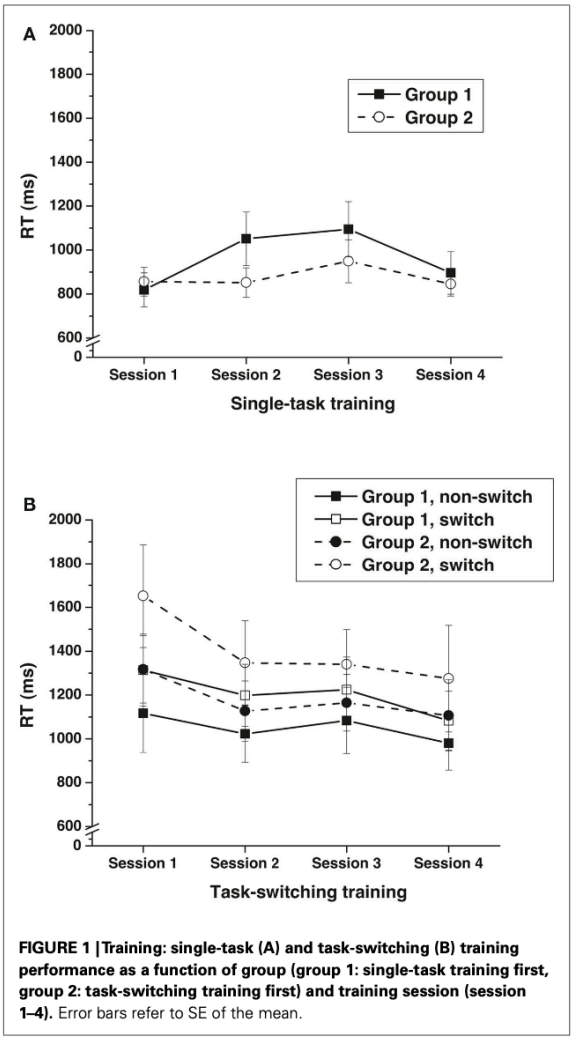

FIGURE 1 | Training: single-task (A) and task-switching (B) training performance as a function of group (group 1: single-task training first, group 2: task-switching training first) and training session (session 1–4). Error bars refer to SE of the mean.

Analysis of pretest data

In order to make sure that transfer effects were not confounded with pre-existing differences in baseline performance, we tested for between-group differences at pretest before analyzing near and far transfer. ANOVAs with the between-subjects factor Group (group 1, group 2) showed no significant group differences on any of the tasks (task-switching: $p = 0.65$, interference control: $p = 0.79$, verbal WM: $p = 0.66$, fluid intelligence: $p = 0.54$, perceptual speed: $p = 0.82$, semantic knowledge: $p = 0.65$).

Near transfer effects

To investigate near transfer effects, we ran an ANOVA including the between-subjects factors Group (group 1: single-task first, group 2: switching first), and the within-subjects factors Trial Type (single, non-switch, switch), and Testing Time (pretest, posttest 1, posttest 2). As in previous studies (e.g., [7]), mixing and switching costs were defined as two orthogonal

contrasts. In the first contrast, the mean performance in single trials was tested against the mean performance on non-switch and switch trials (mixing costs), and in the second contrast mean performance on non-switch trials was tested against the mean performance on switch trials (switching costs). Training-specific effects were assessed by computing two contrasts for the factor Testing Time (pretest, posttest 1, posttest 2): The first contrast compared mean performance at pretest and posttest 1, and the second one compared mean performance at posttest 1 and posttest 2. The means and SD of all experimental conditions are shown in Table 3. Mixing and switching costs as a function of testing time for both groups are displayed in Figures 2A,B.

The overall ANOVA showed a main effect of Trial Type, $F(2, 36) = 41.55$, $p < 0.001$, $\eta_p^2 = 0.70$, revealing significant mixing costs and switching costs ($F(1, 18) = 35.42$, $p < 0.001$, $\eta_p^2 = 0.66$, and $F(1, 18) = 68.38$, $p < 0.001$, $\eta_p^2 = 0.79$, respectively). In addition, we found interactions between Trial Type and Testing Time, $F(4, 18) = 6.49$, $p < 0.001$, $\eta_p^2 = 0.27$, and Trial Type, Testing Time, and Group, $F(4, 18) = 3.51$, $p = 0.01$, $\eta_p^2 = 0.16$. Mixing costs were reduced from pretest to posttest 1, $F(1, 18) = 14.57$, $p < 0.001$, $\eta_p^2 = 0.45$. This reduction was somewhat larger for group 2 (task-switching training; $d' = 1.4$) than for group 1 (single-task training; $d' = 0.6$), $F(1, 18) = 3.40$, $p = 0.08$, $\eta_p^2 = 0.16$. Consistently, there also was a reduction of mixing costs from posttest 1 to posttest 2 in group 1 (task-switching training; $d' = 1.2$) but increased costs in group 2 (single-task training; $d' = -0.7$), $F(1, 18) = 6.64$, $p < 0.05$, $\eta_p^2 = 0.27$ (see Figure 2A).

Switching costs were also reduced from pretest to posttest 1, $F(1, 18) = 21.97$, $p < 0.001$, $\eta_p^2 = 0.55$. Although this effect was larger for group 2 (task-switching training; $d' = 2.6$) than for group 1 (single-task training; $d' = 1.0$), the interaction with group failed to reach significance ($p = 0.17$). The contrast between posttest 1 and posttest 2 showed a reduction of switching costs from posttest 1 to posttest 2 in group 1 (task-switching training; $d' = 0.4$) but an increase in group 2 (single-task training; $d' = -1.0$), $F(1, 18) = 5.02$, $p < 0.05$, $\eta_p^2 = 0.22$ (see Figure 2B).

Far transfer effects

We used a similar ANOVA design to examine far transfer effects of the training intervention. We first report the results on far transfer to other executive control tasks, that is, to interference control and verbal WM, followed by the findings on fluid intelligence (reasoning), and finally to the two control measures, speed of processing, and semantic knowledge. Data of all far transfer measures are shown in Table 3.

Table 3 | Mean performance (SD) for the near transfer (task-switching) and far transfer (inhibition, working memory, fluid intelligence) as a function of testing time (pretest, posttest 1, posttest 2) and group (group 1, group 2).

	Group 1 (single-task training first)						Group 2 (task-switching training first)					
	Pretest		Posttest 1		Posttest 2		Pretest		Posttest 1		Posttest 2	
	M	SD	*M*	SD	*M*	SD	*M*	SD	*M*	SD	*M*	SD
TASK-SWITCHING (NEAR TRANSFER)												
Single trials	1026	304	1106	414	1109	485	1080	272	1021	377	1027	397
Non-switch trials	1261	432	1303	559	1146	512	1355	480	1106	336	1255	722
Switch trials	1607	594	1487	641	1271	514	1715	480	1161	380	1503	869
Mixing costs	409	274	290	208	100	157	455	250	113	246	352	483
Switching costs	346	185	184	186	125	227	360	161	55	150	248	247
STROOP TASK (FAR TRANSFER)												
Congruent trials	899	205	843	163	788	99	859	235	836	221	840	230
Incongruent trials	952	219	900	186	825	127	939	214	823	170	836	182
Interference costs	53	46	57	67	38	61	80	48	−13	82	−4	86
WORKING MEMORY (FAR TRANSFER)												
Working memory	7.0	3.1	8.1	3.1	10.7	2.6	7.6	2.9	9.2	2.4	8.8	1.6
FLUID INTELLIGENCE (FAR TRANSFER)												
Fluid intelligence	21.9	4.9	23.3	2.8	23.3	4.8	20.4	5.7	20.7	4.7	21.4	3.1
CONTROL MEASURES												
Perceptual speed	32.4	8.7	37.4	9.4	38.6	10.2	31.3	12.0	36.5	11.6	36.6	11.2
Semantic knowledge	9.2	2.5	10.4	3.3	11.0	3.9	9.8	3.3	8.8	2.9	9.7	3.5

Interference control. Data were submitted to a three-way ANOVA with the factors Group (group 1: single-task first, group 2: switching first), Testing Time (pretest, posttest 1, posttest 2), and Trial Type (congruent, incongruent). We found a main effect of Testing Time, $F(2, 36)= 5.68$, $p < 0.01$, $\eta_p^2 = 0.24$, indicating that the participants responded faster at posttest 1 than at pretest ($p < 0.05$). The main effect of Trial Type pointed to reliable interference costs $F(1, 18)= 13.69$, $p < 0.01$, $\eta_p^2 = 0.43$, while the main effect of Group failed to reach significance ($p= 0.88$). An interaction between Testing Time and Trial Type, $F(2, 36)= 3.80$, $p < 0.05$, $\eta 2\ p = 0.17$, revealed that interference costs were reduced from pretest to posttest 1, $F(1, 18)= 7.07$, $p < 0.05$, $\eta_p^2 = 0.28$. Importantly, we also found a marginally significant interaction between Testing Time, Trial Type, and Group, $F(2, 36)= 3.12$, $p= 0.06$, $\eta_p^2 = 0.15$. The contrast between pretest and posttest 1 showed that the reduction of interference costs was larger in group 2 (task-switching training; $d' = 1.6$) than in group 1 (single-task training; $d' = 0.1$), $F(1, 18)= 8.33$, $p= 0.01$, $\eta_p^2 = 0.32$ (see Figure 3A), but we obtained no larger reduction of interference costs in group 1 (task-switching training; $d' = 0.4$) than in group 2 (single-task training; $d' = 0.2$) from posttest 1 to posttest 2 ($p= 0.58$).

Verbal working memory. The ANOVA with the factors Group and Testing Time revealed a main effect of Testing Time, $F(2, 36)= 19.62$, $p < 0.001$, $\eta_p^2 = 0.52$, indicating that WM performance improved from pretest to posttest 1 and also from posttest 1 to posttest 2 (both $ps < 0.01$). The main effect of Group was not significant ($p= 0.95$). An interaction between Group and Testing Time, $F(2, 36)= 8.41$, $p < 0.001$, $\eta_p^2 = 0.32$, showed larger performance improvements from

posttest 1 to posttest 2 in group 1 (task-switching training; $d' = 0.9$) than in group 2 (single-task training; $d' = -0.2$; see Figure 3B). However, no training-specific improvements were found from pretest to posttest 1 (single-task training: $d' = 0.3$; task-switching training: $d' = 0.6$; p= 0.44).

Fluid intelligence (reasoning). The ANOVA with the factors Group and Testing Time neither revealed significant main effects nor an interaction (all *ps* > 0.26).

Control variables. The ANOVA based for perceptual speed of processing showed a main effect of Testing Time, F (2, 36)= 15.61, p < 0.001, $\eta_p^2 = 0.46$, with performance improvements from pretest to posttest 1 (p < 0.001), but neither the main effect for Group nor the interaction with Testing Time reached significance (both *ps* > 0.77). The analysis of the semantic knowledge task showed no significant effects (all *ps* > 0.31).

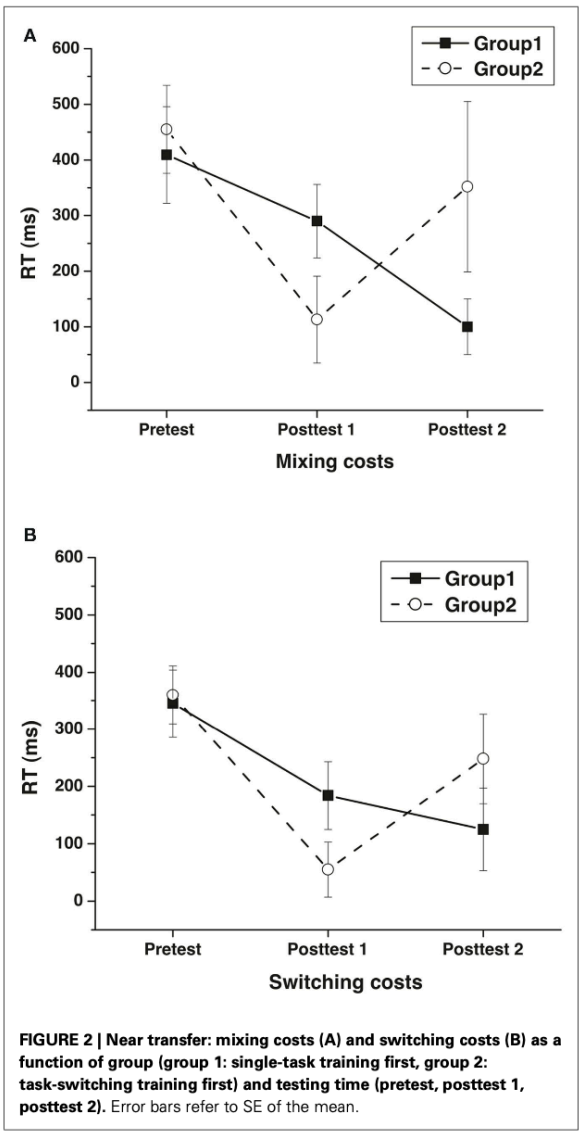

FIGURE 2 | Near transfer: mixing costs (A) and switching costs (B) as a function of group (group 1: single-task training first, group 2: task-switching training first) and testing time (pretest, posttest 1, posttest 2). Error bars refer to SE of the mean.

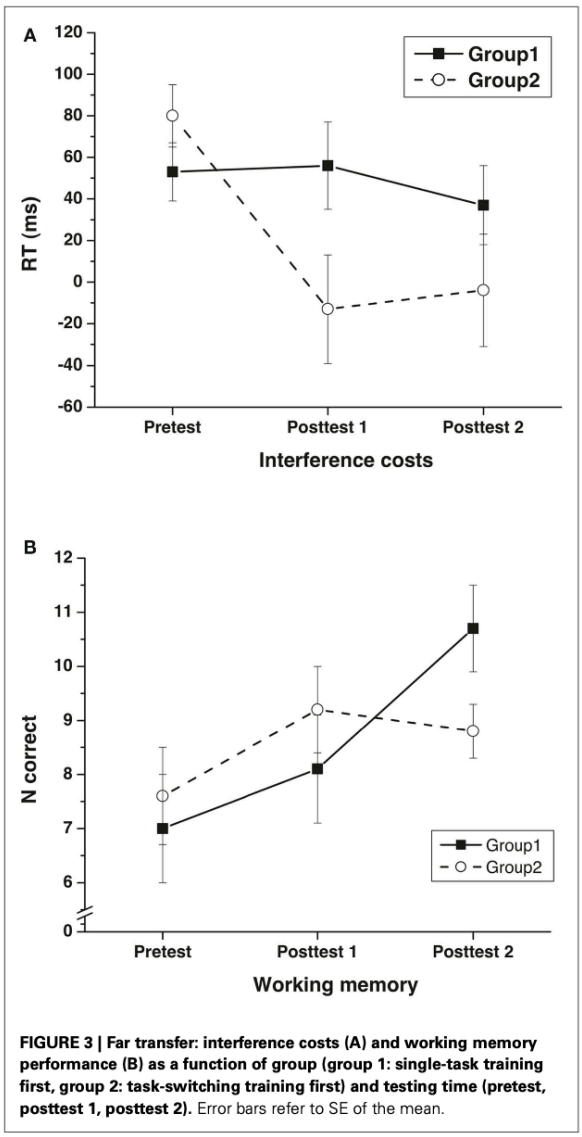

FIGURE 3 | Far transfer: interference costs (A) and working memory performance (B) as a function of group (group 1: single-task training first, group 2: task-switching training first) and testing time (pretest, posttest 1, posttest 2). Error bars refer to SE of the mean.

Discussion

Children with ADHD showed a reduction of switching costs throughout the task-switching training, suggesting that they already benefited from a relatively short intervention of four training sessions. Even more important than the training-related improvements in switching performance are the near and far transfer effects observed in this study. As illustrated in Figure 2A, the task-switching training led to a substantial reduction of mixing costs in a similar switching task with similar effect sizes for the two groups (group1: $d' = 1.4$; group 2: $d' = 1.2$), which can be considered as large effects ([26]). Interestingly, the treatment effect was about the same independently of whether the subjects had already performed the single-task training or not. In contrast, the task-switching training resulted in a large reduction of switching costs in the group that performed the task-switching training first (group 2: $d' = 2.6$), but the reduction in the group that had already performed single-task training was only very small (group 1: $d' = 0.4$), probably because there was not much room for improvement in task-switching (see Figure 2B). A similar pattern of findings occurs for the training-related changes in inhibitory control. While the group that performed the task-switching training first showed a substantial reduction of interference costs (group 2: $d' = 1.6$), this reduction was, however, only of small size for the group that had already performed the single-task training (group 1: $d' = 0.4$). We obtained a large increase in verbal WM in the group that performed the task-switching training first (group 2: $d' = 0.9$) while the effect size was only medium for the group that had already performed the single-task training (group 1: $d' = 0.6$). In contrast to our previous study with young children ([14]), we did not find transfer of task-switching training to performance on a fluid intelligence test in children suffering from ADHD. However, it should be noted that we used different tests in both studies, which might explain the difference in findings.

In sum, the present study provided the first evidence for near and far transfer of task-switching training in children suffering from ADHD. It therefore is of major interest to examine whether the training was as effective in children with ADHD as it has previously been in healthy children. Comparing the results from the present study with our previous one ([14]) showed that the effect sizes for the near transfer of task-switching were higher in healthy children than in the ADHD sample in terms of mixing costs (mean $d'_{\text{healthy group}} = 2.1$, mean $d'_{\text{ADHD group}} = 1.3$) but similar in terms of switching costs (mean $d'_{\text{healthy group}} = 1.2$, mean $d'_{\text{ADHD group}} = 1.5$). Regarding the far transfer to interference control, we even found slightly higher effects sizes in the ADHD group than in the healthy sample (mean $d'_{\text{healthy group}} = 0.5$, mean $d'_{\text{ADHD group}} = 0.8$),

while the transfer to WM was comparable across studies (mean $d'_{\text{healthy group}} = 0.9$, mean $d'_{\text{ADHD group}} = 0.8$). Thus, the general pattern of results across both groups showed the typical finding of larger effect sizes on near compared to far transfer tasks. In addition, the size of these effects was similar (with the exception of mixing costs), indicating that results of the ADHD children seem to be within the range of what has been reported for healthy children. Although this finding has to be replicated within a single study, it points to the potential for the application of relatively short cognitive interventions in clinically relevant populations.

Although there was evidence for training-specific improvements of the task-switching intervention, it should be noted that we also obtained transfer effects of medium sizes after the single-task training. One possible explanation of this finding is that ADHD children have major deficits in the control of attention and interference. Given that the stimuli in this study were ambiguous, even the single-task training may have resulted in a certain amount of training in executive control. This means that although the ADHD children were not trained in task-switching, they may have been trained in focusing their attention on relevant information while ignoring irrelevant task features. Although we found large effect sizes for near and far transfer of task-switching training, this study has some limitations. First, the sample was relatively small so that some interactions of the expected training-specific effects were only marginally significant. Second, the fact that we only investigated male children limits the generalizability of our findings. Third, given our training design, we also observed a decrease in task-switching performance between the posttest 1 and posttest 2 for the group that performed the single-task training after the task-switching training (group 2), as illustrated in Figures 2A,B. One possible explanation for this finding is that the ADHD children suffered from a loss of motivation across the four easier single-task training sessions and were therefore also less motivated to perform the switching tasks at posttest 2. Another explanation would be that the decrease in performance reflects negative transfer in the sense that the intensive training in performing single-tasks interferes with the coordination of control processes required for the switching tasks. Unclear is, however, why this negative effect does not occur for group 1. Either way future research is needed to clarify the nature of this carryover effect. If training order effects influence motivation, future studies could additionally control for individual differences in motivation and self-regulatory strategies such as self-efficacy or active engagement in the training. Such individual characteristics have recently been found to moderate memory training and transfer effects in elderly subjects (e.g., West and Hastings, 2011). As children with ADHD have impairments in regulating and maintaining engagement in an activity for a longer period of

time, these motivational factors might also contribute to differential training and transfer effects in this clinical group.

The present training study extended our knowledge regarding useful cognitive training interventions for children with combined ADHD who were on stable methylphenidate medication. Previous studies found that executive control functioning as well as academic skills and behavioral deficits can be improved by WM training and attentional control training ([28]; [29]; [30]). The intensity of the training was quite high in these studies [e.g., at least 25 training sessions in the Klingberg et al. (2005) study]. Results of our study suggest that performance improvements in executive control functioning can be achieved after a relatively short training intervention of four sessions in task-switching. However, whether even larger training effects can be achieved with adaptive or more intensive training procedures ([15a]; [15b]) and whether training effects can be maintained over a longer period of time has to be clarified in future studies. Another interesting question for future research with important clinical implications is to directly compare the effectiveness of the already existing training programs or to combine them in order to achieve an optimal cognitive intervention for children with ADHD.

Annexe 5

PREMIÈRE VERSION DU QUESTIONNAIRE

Avant de débuter le questionnaire, veuillez noter qu'il se divise en deux parties : la première concerne vos informations personnelles et vos habitudes de lecture d'articles scientifiques, tandis que la seconde vous invite à lire deux articles scientifiques puis à répondre à des questions s'y rapportant.

Vous êtes libre d'interrompre l'enquête à tout moment et de la reprendre plus tard. Cependant, nous vous conseillons de la compléter en une seule fois ou, si besoin, de faire une pause avant d'entamer la lecture de chacun des deux articles, car il est impossible de revenir en arrière dans le questionnaire.

SECTION 1 : INFORMATIONS GENERALES

1. **Confirmez-vous être psychologue spécialisé(e) en neuropsychologie et exercer l'activité de chercheur dans un pays francophone au moins à temps partiel actuellement ou dans la dernière année ?**
   - o Oui
   - o Non

2. **Vous êtes :**
   - o Un homme
   - o Une femme
   - o Autre

3. **Depuis combien d'années exercez-vous en tant que chercheur (doctorat inclus) :**
   - o Moins d'1 an
   - o De 1 à 5 ans
   - o De 6 à 10 ans
   - o Plus de 10 ans

4. **Dans quel(s) pays exercez-vous ?**
   - o France
   - o Belgique
   - o Canada
   - o Suisse
   - o Grand-Duché de Luxembourg
   - o Autre (QO)

5. Vous êtes titulaire :

- o D'un Master (ou DESS) de psychologie spécialisé(e) en neuropsychologie ou MAS en neuropsychologie
- o D'un Master (ou DESS) de psychologie sans spécialisation en neuropsychologie
- o D'un DESS en psychologie obtenu avant 2000, et vous avez une pratique spécialisée en neuropsychologie depuis plus de 10 ans
- o D'un Doctorat
- o D'un post-doctorat
- o D'un Diplôme Universitaire (DU)
- o D'un Certificat Universitaire (CU)
- o Autres (QO)

SECTION 2 : CONTEXTE PROFESSIONNEL

6. Quelle est votre situation professionnelle actuelle ?

- o Chercheur à temps plein
- o Chercheur à temps partiel – Plus d'un mi-temps mais pas à temps plein
- o Chercheur à temps partiel – A mi-temps
- o Chercheur à temps partiel – Moins d'un mi-temps
- o Sans activité de chercheur

7. Quel statut occupez-vous actuellement ?

- o Doctorant
- o Post-doctorant
- o Chercheur permanent/ enseignant à l'université
- o Autre (QO)

8. Dans le cadre de votre travail de chercheur avez-vous des tâches d'enseignement ?

- o Non
- o Oui, des cours théoriques
- o Oui, des travaux pratiques
- o Autres (QO)

9. Avez-vous une activité en neuropsychologie clinique ?

- o Oui
- o Non

9.1. Si oui, combien de temps par semaine consacrez-vous à la clinique (pourcentage d'un temps plein) ?

- o Moins de 10 %
- o 10 à 25 %
- o 25 % à 50 %
- o Plus de 50 %

9.2. Si oui, quelle(s) population(s) rencontrez-vous actuellement dans le cadre de vos activités de neuropsychologue clinicien ?

- o Enfants
- o Adolescents
- o Adultes
- o Adultes âgés (> 60 ans)

9.3. Si oui, dans quel(s) type(s) de service(s) exercez-vous votre activité de neuropsychologue clinicien actuellement ?

- o Neurologie, neuropédiatrie
- o Gériatrie (service de gériatrie, EHPAD, ...)
- o Consultation mémoire
- o Neurochirurgie
- o Oncologie
- o Psychiatrie, pédo-psychiatrie, géronto-psychiatrie, addictologie
- o Rééducation, réhabilitation
- o Médico-social ou médico-éducatif (FAM, UEROS, SESSAD, ...)
- o Recherche, université
- o Libéral
- o Association
- o Organisme de formation
- o Autres (QO)

10. Avez-vous déjà publié des articles scientifiques dans une revue internationale avec peer-reviewing (En 1e auteur ou non) ?

- o Oui
- o Non

10.1. Si oui, combien en avez-vous publié (en premier auteur ou non) ?

- o 1-5
- o 5-10
- o 10-50
- o Plus de 50

11. Avez-vous déjà participé à la réalisation d'un essai contrôlé randomisé ?

- o Oui
- o Non

12. Avez-vous déjà publié des articles scientifiques portant sur des prises en charge/ revalidations en neuropsychologie ou des essais contrôlés randomisés dans une revue internationale avec peer-reviewing (en 1e auteur ou non) ?

- o Oui
- o Non

12.1. Si oui, combien d'articles prises en charge/ revalidations avez-vous publiés en tant que premier auteur, en tant que co-auteur et en tant que dernier auteur ? (QO)

12.2. Si oui, combien d'essais contrôlés randomisés avez-vous publiés en tant que premier auteur, en tant que co-auteur et en tant que dernier auteur ? (QO)

12.3. Si oui, combien d'essais contrôlés randomisés prises en charge/ revalidations avez-vous publiés en tant que premier auteur, en tant que co-auteur et en tant que dernier auteur ? (QO)

13. Si vous avez déjà rédigé un article portant sur un essai contrôlé randomisé, avez-vous utilisé un outil de *reporting* pour vous aider dans votre rédaction ?

- Oui
- Non

13.1. Si oui, lequel ? (QO)

13.2. Si non, pourquoi ? (QO)

SECTION 3 : HABITUDES DE LECTURE

14. Vous arrive-t'il de lire des études scientifiques portant sur des prises en charges ?

- Non jamais
- Oui, mais vraiment exceptionnellement
- Oui, en moyenne une fois par trimestre
- Oui, en moyenne une ou deux fois par mois
- Oui, une fois par semaine ou plus

15. Quelles sont les informations auxquelles vous prêtez attention afin de savoir si l'article que vous êtes en train de lire est de bonne qualité ? (QO)

Vous allez à présent lire deux essais contrôlés randomisés visant à évaluer l'efficacité de deux prises en charge en neuropsychologie. Nous vous demandons de lire l'article de la même manière et dans les mêmes conditions que vous le ferez habituellement dans le cadre de votre travail de chercheur. Après chaque article, vous devrez répondre à une série de questions.

16. Êtes-vous d'accord de lire ces articles et de continuer cette enquête ?

- Oui
- Non

16.1. Si non, pourquoi ? (QO)

__Attention !__ Nous vous rappelons qu'il est préférable de ne pas interrompre le questionnaire entre la lecture de l'article et les questions qui y sont associées. En cliquant sur « Suivant », vous commencerez la lecture du premier article.

SECTION 4 : PRATIQUES DE LECTURE ET QUESTIONS DE CONTROLE

17. Avez-vous lu l'article dans son entièreté ?

- o Oui
- o Non

17.1. Si non, quelles parties avez-vous lues ?

- o Le résumé (*« abstract »*)
- o L'introduction
- o Les méthodes et procédures
- o Les instruments de mesure et d'évaluation
- o Les résultats (texte)
- o Les résultats (tableaux et graphiques avec leurs légendes)
- o La discussion
- o La conclusion finale

18. Aviez-vous déjà lu cet article auparavant ?

- o Oui
- o Non

19. Êtes-vous familier avec le sujet traité dans l'article ?

- o Oui
- o Non

20. Dans l'article que vous venez de lire, l'intervention a été testée sur :

- o Des enfants avec trouble déficitaire de l'attention
- o Des adultes avec trouble du spectre autistique
- o Des enfants tout-venants

21. Dans l'article que vous venez de lire, l'intervention administrée dans cette étude était :

- o *Cogmed*
- o *RehaCom*
- o *Braingame Brian*

22. Dans l'article que vous venez de lire, combien y avait-il de groupes d'intervention ?

- o 2
- o 3
- o 4

SECTION 5 : CONFIANCE DONNÉE AUX RÉSULTATS ET IDENTIFICATION DES BIAIS

23. Quel est votre niveau de confiance en ce qui concerne les résultats de cet article ? *De (0) = « Pas du tout confiance » à (10) = « Entièrement confiance ».*

24. Avez-vous repéré des biais méthodologiques dans l'article que vous venez de lire pouvant altérer la confiance que vous avez envers ses résultats ?

- Oui
- Non

24.1. Si oui, lesquels ? (QO)

25. Indiquez dans quelle mesure vous êtes en accord avec chaque phrase en utilisant l'échelle suivante. *De (1) = « Pas du tout d'accord » à (4) = « Parfaitement d'accord ».*

- Vous avez compris la partie introduction
- Vous avez compris la méthodologie
- Vous avez compris la partie résultat
- Vous avez compris la partie discussion
- L'article que vous venez de lire a de bonnes qualités méthodologiques
- Si vous souhaitiez reproduire cette intervention, vous pensez avoir assez d'information
- Si vous deviez vous positionner quant à l'efficacité de l'intervention, vous pensez avoir assez d'information
- Si vous deviez vous positionner quant à l'utilité de cette intervention, vous pensez avoir assez d'information

26. Avez-vous été rechercher cet article sur un moteur de recherche afin d'obtenir plus d'informations ?

- Oui
- Non

27. Avez-vous utilisé un outil d'aide à la lecture (traducteur, grille de lecture, etc.) pour lire cet article ?

- Oui
- Non

27.1. Si oui, lequel ? (QO)

SECTION 6 : ÉVALUATION CRITIQUE DE L'ARTICLE

28. Afin de juger de la qualité de l'article, à quoi avez-vous prêté attention ?
Pour chacun des items, répondre par : (1) = « Je n'y ai pas prêté attention », (2) = « J'y ai prêté attention et j'ai eu l'impression que c'était réalisé de façon satisfaisante », (3) = « J'y ai prêté attention et j'ai eu l'impression que ce n'était pas réalisé ou réalisé de façon insatisfaisante », (4) = « J'y ai prêté attention mais je ne m'en souviens plus », (5) = « J'y ai prêté attention mais je ne peux me positionner quant à la qualité ».

- o Au calcul de la taille de l'échantillon
- o Au processus de randomisation entre les groupes
- o A la similarité entre les groupes avant que l'étude ne commence
- o Au processus d'aveuglement (*blinding*) des participants
- o Au processus d'aveuglement (*blinding*) des expérimentateurs responsables de l'intervention
- o Au processus d'aveuglement (*blinding*) des parents
- o Au processus d'aveuglement (*blinding*) des évaluateurs (pré et post- intervention)
- o A la similarité des caractéristiques des participants
- o Aux similarités des interventions entre le groupe expérimental et les groupes contrôles
- o A la description détaillée du matériel de l'intervention
- o A la description détaillée du contenu de chaque séance de l'intervention (longueur, type d'exercice)
- o A la description détaillée de l'agenda et de l'organisation de l'intervention
- o Aux raisons d'abandon des participants entre le début et la fin de l'étude
- o A la pertinence des résultats récoltés (*outcomes*)
- o A la façon dont les résultats récoltés (*outcomes*) ont été mesurés
- o A la pertinence des tests statistiques utilisés
- o A la présence d'une *p*-valeur pour chacun des résultats (*outcomes*)
- o A la présence d'une taille d'effet pour chacun des résultats (*outcomes*)
- o A la présence d'un intervalle de confiance autour de la taille d'effet dans les résultats
- o A la pertinence du *design* choisi (ECR) en fonction de l'objectif de l'étude
- o A la présence d'un protocole publié avant le début de l'étude
- o A l'absence de conflit d'intérêt
- o A la disponibilité des données de l'étude

29. Quel est votre niveau de confiance en ce qui concerne les résultats de cet article ? *De (0) = « Pas du tout confiance » à (10) = « Entièrement confiance ».*

__Attention !__ Nous vous rappelons qu'il est préférable de ne pas interrompre le questionnaire entre la lecture de l'article et les questions qui y sont associées. En cliquant sur « Suivant », vous commencerez la lecture du second article.

SECTION 7 : PRATIQUES DE LECTURE ET QUESTIONS DE CONTROLE

30. Avez-vous lu l'article dans son entièreté ?

- o Oui
- o Non

30.1. Si non, quelles parties avez-vous lues ?

- o Le résumé (*« abstract »*)
- o L'introduction
- o Les méthodes et procédures
- o Les instruments de mesure et d'évaluation
- o Les résultats (texte)
- o Les résultats (tableaux et graphiques avec leurs légendes)
- o La discussion
- o La conclusion finale

31. Aviez-vous déjà lu cet article auparavant ?

- o Oui
- o Non

32. Êtes-vous familier avec le sujet traité dans l'article ?

- o Oui
- o Non

33. Dans l'article que vous venez de lire, l'intervention a été testée sur :

- o Des enfants avec déficience intellectuelle
- o Des enfants avec trouble déficitaire de l'attention avec hyperactivité
- o Des enfants avec dyslexie

34. Dans l'article que vous venez de lire, en quoi consistait l'intervention ?

- o Sur l'écran d'un ordinateur, des étoiles et des nuages apparaissaient. Les enfants devaient appuyer sur la touche de droite lorsqu'une étoile s'affichait et sur la touche de gauche lorsqu'un nuage apparaissait
- o Les enfants alternaient entre deux tâches. Dans la première tâche, les enfants devaient dire si ce qu'ils voyaient à l'écran était un fruit ou un légume et, dans la deuxième tâche, ils devaient dire si ce qu'ils voyaient était grand ou petit
- o Des symboles *« + »* et *« - »* apparaissaient à l'écran. Les enfants devaient nommer le symbole affiché lorsqu'il était noir et indiquer le symbole opposé lorsqu'il était rouge

35. Dans l'article que vous venez de lire, combien y avait-il de groupes d'intervention ?

- 2
- 3
- 4

SECTION 8 : CONFIANCE DONNÉE AUX RÉSULTATS ET IDENTIFICATION DES BIAIS

Idem que section 5.

SECTION 9 : ÉVALUATION CRITIQUE DE L'ARTICLE

Idem que section 6.

SECTION 10 : FIN DU QUESTIONNAIRE

36. Selon vous, lequel des deux articles que vous venez de lire possédait la meilleure qualité méthodologique ?

- L'article 1
- L'article 2

37. Avez-vous interrompu le questionnaire à un moment donné ?

- Oui
- Non

37.1. Si oui, à quel moment ? (QO)

38. Avez-vous des commentaires ou suggestions concernant le thème de cette enquête ? (QO)

Merci d'avoir complété ce questionnaire, nous tâcherons d'en valoriser les données autour de communications relatives au sujet abordé. Pour recevoir un retour sur l'enquête ainsi que des informations concernant les articles lus, veuillez contacter Sacha Blause (Sacha.Blause@uliege.be).

Annexe 6

Caractéristiques démographiques des participants.

Données démographiques	***N***	**%**
Genre		
Homme	4	19
Femme	17	81
Autre	0	0
Pays d'exercice		
France	4	19
Belgique	17	81
Canada	0	0
Suisse	1	5
Grand-Duché de Luxembourg	1	5
Autre	0	0
Années d'activités		
Moins d'1 an	5	24
De 1 à 5 ans	7	33
De 6 à 10 ans	5	24
Plus de 10 ans	4	19
Diplômes		
Master (ou DESS) de psychologie spécialisée en neuropsychologie ou MAS en neuropsychologie	14	67
Master (ou DESS) de psychologie sans spécialisation en neuropsychologie	3	14
DESS en psychologie obtenu avant 2000 et pratique spécialisée en neuropsychologie depuis plus de 10 ans	0	0
Doctorat	8	38
Post-doctorat	5	24
Diplôme Universitaire (DU)	2	10
Certificat Universitaire (CU)	0	0
Autre	0	0

Notes. N = nombre de participants concernés ; % = pourcentage calculé sur l'échantillon total (N = 21)

Annexe 7

Contexte professionnel des participants.

Contexte professionnel	***N***	**%**
Situation professionnelle actuelle		
Chercheur à temps plein	13	62
Chercheur à temps partiel – Plus d'un mi-temps mais pas à temps plein	1	5
Chercheur à temps partiel – Mi-temps	4	19
Chercheur à temps partiel – Moins d'un mi-temps	3	14
Sans activité de chercheur	0	0
Statut actuel		
Doctorant	9	43
Post-doctorant	2	10
Chercheur permanent/ enseignant à l'université	3	14
Autre	7	33
Tâches d'enseignement		
Aucune	7	33
Cours théoriques	10	48
Travaux pratiques	11	52
Autre	0	0
Activité de neuropsychologue clinicien		
Oui	7	33
Non	14	67
Temps consacré à la clinique par semaine (*N* = 7)		
Moins de 10 %	1	14
10 à 25 %	0	0
25 à 50 %	4	57
Plus de 50 %	2	29
Population clinique rencontrée (*N* = 7)		
Enfants	3	43
Adolescents	2	29
Adultes	6	86
Adultes âgés (> 60 ans)	5	71
Secteur d'exercice actuel (*N* = 7)		
Neurologie, neuropédiatrie	0	0
Gériatrie (service de gériatrie, EHPAD, etc.)	1	14
Consultation mémoire	0	0
Neurochirurgie	0	0
Oncologie	0	0
Psychiatrie, pédo-psychiatrie, géronto-psychiatrie, addictologie	0	0
Rééducation, réhabilitation	3	43
Médico-social ou médico-éducatif (FAM, UEROS, SESSAD, etc.)	0	0
Recherche, université	1	14
Libéral	3	43
Association	0	0
Organisme de formation	0	0
Autre	1	14
Publication d'articles scientifiques dans une revue internationale avec *peer-reviewing*		
Oui	12	57
Non	9	43
Nombre d'articles scientifiques publiés dans une revue internationale avec *peer-reviewing* (*N* = 12)		
1-5	5	42
5-10	2	16
10-50	5	42
Plus de 50	0	0
Participation à la réalisation d'un essai contrôlé randomisé		
Oui	9	43
Non	12	57
Publication d'articles scientifiques portant sur des prises en charge/ revalidations en neuropsychologie ou des essais contrôlés randomisés dans une revue internationale avec *peer-reviewing* (en 1e auteur ou non)		
Oui	2	10
Non	19	90
Utilisation d'un outil de *reporting* lors de la rédaction d'un essai contrôlé randomisé (*N* = 2)		
Oui	0	0
Non	2	100

Notes. N = nombre de participants concernés ; % = pourcentage calculé sur l'échantillon total (N = 21) sauf mention spécifique

Annexe 8

Réponses des participants à la question ouverte concernant les critères utilisés pour évaluer la qualité d'un article scientifique.

1	La problématique/question de recherche est-elle clairement définie ? Le projet de recherche aborde-t-il un problème important ou un obstacle critique au progrès dans le domaine ? Le *design* est-il adapté ? Quels sont les risques de biais (selon le *design* ; mise en aveugle ou non ; calcul de taille d'échantillon ou non ; les auteurs reconnaissent-ils les limites de l'étude, etc.) ?
2	Le nombre de participants, les tests utilisés, la cohérence entre les hypothèses, la méthodo, les statistiques et la discussion.
3	La qualité de la revue et le type d'étude (ECR ou méta-analyses).
4	Le site de publication et la récence.
5	La revue dans laquelle l'article est publié, la description de la méthode et les conflits d'intérêt.
6	La méthode en globalité, la taille de l'échantillon et les statistiques.
7	Le cadre théorique, l'objectif de la recherche, la méthodologie et les statistiques utilisées.
8	L'introduction, la méthode, les résultats et la discussion.
9	Les sources et la méthodologie.
10	Je m'intéresse à la taille de l'échantillon (voir si la puissance est assez importante), à la taille d'effet (si l'effet de la prise en charge est réel, est-il important ?), à la manière dont l'amélioration des performances est évaluée (tâches neuropsychologiques ou/et contexte écologique). Pour avoir l'effet complet d'une prise en charge, c'est intéressant d'avoir les deux, et si seulement un des deux, les performances écologiques sont évidemment plus importantes. Je m'intéresse au recrutement des participants et à la formation des groupes : que fait-on avec le groupe contrôle ? Une prise en charge (autre) est-elle tout de même prévue ? Dans le cas contraire, il y aura des variables confondantes. En d'autres termes, pour que l'étude soit bien réalisée, il faut que la seule différence entre le groupe contrôle et le groupe expérimental soit la prise en charge spécifique (tous les autres paramètres doivent être similaires dans les deux groupes).
11	La question théorique abordée.
12	La question de recherche est claire, innovante, *« impactante »*. Si données préalables, calcul de taille d'échantillon + puissance statistique suffisante. Le *design* d'étude et les statistiques sont adaptés. Les auteurs et la revue dans laquelle l'article a été publié. La méthodologie est bien détaillée et le *reporting* des résultats est clair/ complet. La transparence des auteurs quant aux biais et limites potentiels.
13	Les dates des références (anciennes ou non), en fonction de l'article, le nombre de personnes/ articles sur lesquels se basent les analyses / l'intervention (petit ou grand échantillon), la méthodologie et descriptions des outils (si c'est détaillé ou non), les limitations (sont-ils critiques sur la méthodo ou autre), si juste en lisant je pourrais reproduire l'étude / l'analyse / l'intervention.
14	Si l'étude est significative ou non et quelle est la taille d'effet. Je prête également attention aux ingrédients actifs et j'essaie de repérer les éventuelles limites ou biais.
15	Le nom de la revue, s'assurer que l'article est bien *peer-reviewed*, l'identité de l'auteur, la qualité de la méthodologie et la taille de l'échantillon testée.
16	Le type de revue dans lequel l'article est publié, la qualité du texte (grammaire, enchaînement des idées), la méthodologie employée.
17	Type de revue, facteur d'impact, nombre de citations, abstract clair et pertinent et enfin moins souvent, indice statistique comme la puissance, etc.
18	Formulation claire des hypothèses, justification de la taille d'échantillon et procédure utilisée.
19	Le journal, les auteurs, l'agencement des idées, l'abstract, le contenu global, la description de la méthodologie et les détails mentionnés.
20	Premièrement dans les méthodes, je regarde le type de tests utilisés, la composition des groupes (présence de données démographiques, équivalence des groupes), l'allocation des participants et le nombre de participants. Ensuite, je regarde les analyses prévues. Puis, si par rapport aux méthodes, tous les résultats sont bien reportés et si ces résultats incluent également une taille d'effet. Enfin, dans les résultats et la discussion, si les résultats sont interprétés de façon cohérente et si les chercheurs répondent correctement à l'hypothèse principale. L'open data entre aussi en considération.
21	Dans quelle revue a été publié l'article.

Annexe 9

Résultats du test de normalité de Shapiro-Wilk.

Test de normalité – Shapiro Wilk	**Statistique**	***p***
Confiance *reporting* + vs. *reporting* -	0.92	.251

Annexe 10

Résultats des tests de normalité de Shapiro-Wilk.

Test de normalité – Shapiro-Wilk	**Statistique**	***p***
Confiance avant vs. après – *reporting* +	0.77	.003*
Confiance avant vs. après – *reporting* -	0.65	< .001*

*Notes. * p < .05*

Annexe 11

Réponses des participants à la question ouverte quant aux biais méthodologiques identifiés dans l'article *reporting* + ($N = 5$).

1	Il aurait fallu deux groupes contrôles en plus : un avec des enfants qui seraient invités simplement à jouer à un jeu vidéo et un deuxième avec des enfants qui seraient invités à réaliser des tâches papier-crayon. De cette manière, les auteurs pourraient répondre aux questions qu'ils se posent dans la discussion. Au niveau des mesures prises, il manque des mesures écologiques telles que de réelles performances académiques par exemple.
2	Si j'ai bien compris, ils font repasser les mêmes tests en pré- et en post-test, sans tenir compte de l'effet test-retest. Ils utilisent également des tests un peu dépassés comme les critères du DSM-IV et la WISC-III (cela dit, je ne connais pas bien cette version, donc elle est peut-être intéressante dans ce cas-ci).
3	Les critères d'inclusion et d'exclusion me semblent peu précis… Notamment la prise de médication (même si homogène entre les groupes).
4	Justification de la taille d'échantillon sur base d'études précédentes (pas de test de puissance).
5	Très peu, à part celles évoquées par les auteurs, si ce n'est de ne pas avoir mentionné la possibilité que l'arrêt du traitement pendant les entraînements ait pu avoir un effet sur la capacité de ceux qui sont sous traitement à se concentrer durant la tâche expérimentale, et de les avoir mis dans le même groupe que ceux qui ne sont pas sous traitement. Ça reste un détail.

Annexe 12

Réponses des participants à la question ouverte quant aux biais méthodologiques identifiés dans l'article *reporting* - (N = 10).

1	Pas de test de puissance pour le calcul de la taille de l'échantillon et pas de mesures écologiques.
2	Oui, comme le fait qu'il n'y ait que des garçons ou que les enfants TDAH soient sous médication, mais ils mentionnent certaines de leurs limites dans la discussion.
3	Absence de randomisation de l'administration des tâches du pré- et du post-test.
4	Pas d'analyse de puissance. Petit échantillon. Utilisation d'un score de différence pour calculer le *task switching*. Pas de correction particulière sur les temps de réponse.
5	Idéalement, je ne serais pas partie sur ce *design* me semble-t-il + pas de périodes de *wash-out* entre les différentes sessions d'entraînement ? Si j'ai bien retenu, ils ont retiré les données relatives aux *drop-out*, sans analyser en intention de traiter. Je ne me souviens d'ailleurs pas avoir lu comment se répartissent les raisons des abandons dans chacun des groupes + pas lu d'analyse pour comparer ces sujets à ceux toujours dans l'étude. Pas de mise en aveugle, ni d'infos sur ce que les participants connaissent des objectifs de l'étude et des tâches administrées. Limites et manque de généralisation avec le très petit échantillon et la population de garçons exclusivement.
6	Alors, il faudrait du temps pour décrire ce qui peut biaiser les résultats, mais on peut déjà dire de manière générale un manque de clarté concernant : - Les participants et les *screening* qui ont été fait, on a des informations trop vagues concernant certaines choses, par exemple la médication, ou même les outils utilisés pour le diagnostic, ou encore la justification de ne prendre que des garçons car il y a plus de garçon TDAH, ou la récompense financière pour des enfants. - Les tâches, avec des choix de tâches pas toujours justifiés, et parfois pas très bien construite (tâches entrainement ou évaluation) ou encore ordre des tâches pas suffisamment bien établi avec le manque d'un vrai groupe contrôle. - Les descriptions de la procédure ou des résultats qui comportent des lacunes et des imprécisions. Enfin, discussion peu précise et peu rigoureuse limitant l'impact des résultats.
7	Pas de mention du *blinding*, donc quelle consigne a été donnée aux enfants et aux examinateurs ? Pas de justification du nombre de participants. Qu'en est-il de la puissance ? Les auteurs ne cherchent que des garçons pour leur étude, citant que les filles ont peu souvent un trouble *ADHD*. Certains auteurs pensent que le n de garçons avec un ADHD est le même que celui des filles, mais que les filles sont diagnostiquées beaucoup plus tard que les garçons. A mon sens ce n'est donc pas une bonne justification, leur échantillon ne correspond pas à la population. Toujours concernant les critères d'exclusion, ils excluent les enfants avec un diagnostic de *ASD*. Or, les deux troubles seraient assez liés et nombreux enfants auraient les deux troubles. Les résultats ne seront donc pas applicables pour beaucoup d'enfant. Les auteurs, si j'ai bien compris, utilisent une intervention quasi identique à leur test pour le *task switching*. Pour avoir confiance en leur résultats, j'aurais préféré qu'ils effectuent aussi une mesure de ce construct avec un autre test. Je ne peux pas être certaine que l'intervention améliore le *task switching*, il est possible que les enfants soient juste entrainés à mieux répondre à l'épreuve. Il n'y a pas de mesure sur le long terme. Est-ce que les effets de l'intervention seront toujours présents dans quelques mois/années ? Les auteurs ne mentionnent pas modifier l'ordre de passation des tests ; quid de la fatigabilité ? Quid de l'effet test-retest, si les tests sont similaires de semaine en semaine ?
8	La taille d'échantillon est très petite.
9	Faible effectif.
10	Je trouve que la sélection et la répartition des sujets ne sont pas vraiment bien détaillées. Idem pour les tâches et questionnaires, ce qui ne permet pas d'avoir toutes les informations pour juger la pertinence.

Annexe 13

Tableau de contingence entre la fréquence de lecture d'ECR cliniques et l'identification de biais dans l'article *reporting* -.

Fréquence de lecture	Biais identifiés article R-		Total
	Non	Oui	
Jamais	0	3	3
Exceptionnellement	1	5	6
Une/ deux fois par mois	1	2	3
Une fois semaine ou plus	1	0	1
Total	3	10	13

Annexe 14

Tableau de contingence entre l'identification de biais dans l'article *reporting* - et le choix de l'article jugé de meilleure qualité.

Biais identifiés R-	Identification du meilleur article		Total
	Non	Oui	
Non	0	3	3
Oui	2	8	10
Total	2	11	13

Annexe 15

Résultats des tests de normalité et synthèse des tests statistiques appliqués.

Test de normalité – Shapiro-Wilk	**Statistique**	***p***
Compréhension de l'introduction	0.75	.002*
Compréhension de la méthodologie	0.81	.009*
Compréhension des résultats	0.87	.048*
Compréhension de la discussion	0.85	.025*
Évaluation de la qualité méthodologique	0.81	.008*
Suffisamment d'informations pour reproduire l'intervention	0.93	.311
Suffisamment d'informations pour juger l'efficacité	0.81	.010*
Suffisamment d'informations pour juger l'utilité	0.78	.004*

*Notes. * $p < .05$*

Variables	**Normalité**	**Choix du test**
Compréhension de l'introduction	Non respectée	Rangs signés de Wilcoxon
Compréhension de la méthodologie	Non respectée	Rangs signés de Wilcoxon
Compréhension des résultats	Non respectée	Rangs signés de Wilcoxon
Compréhension de la discussion	Non respectée	Rangs signés de Wilcoxon
Évaluation de la qualité méthodologie	Non respectée	Rangs signés de Wilcoxon
Suffisamment d'informations pour reproduire l'intervention	Respectée	Test *t* de Student pour échantillons appariés
Suffisamment d'informations pour juger l'efficacité	Non respectée	Rangs signés de Wilcoxon
Suffisamment d'informations pour juger l'utilité	Non respectée	Rangs signés de Wilcoxon

Annexe 16

Répartition des critères pris en compte ou ignorés par les participants.

Critères examinés pour évaluer la qualité de l'article	Article R+		Article R-	
	N	%	*N*	%
Calcul de la taille de l'échantillon				
Critère non pris en compte ou évaluation non claire	4	31	2	15
Critère pris en compte	9	69	11	85
Processus de randomisation entre les groupes				
Critère non pris en compte ou évaluation non claire	2	15	6	46
Critère pris en compte	11	85	7	54
Similarité entre les groupes avant que l'étude ne commence				
Critère non pris en compte ou évaluation non claire	5	38	3	23
Critère pris en compte	8	62	10	77
Processus d'aveuglement (*blinding)* des participants				
Critère non pris en compte ou évaluation non claire	1	8	7	54
Critère pris en compte	12	92	6	46
Processus d'aveuglement (*blinding*) des expérimentateurs responsables de l'intervention				
Critère non pris en compte ou évaluation non claire	5	38	5	38
Critère pris en compte	8	62	8	62
Processus d'aveuglement (*blinding*) des parents				
Critère non pris en compte ou évaluation non claire	2	15	8	62
Critère pris en compte	11	85	5	38
Processus d'aveuglement (*blinding*) des évaluateurs (pré et post- intervention)				
Critère non pris en compte ou évaluation non claire	6	46	5	38
Critère pris en compte	7	54	8	62
Similarité des caractéristiques des participants				
Critère non pris en compte ou évaluation non claire	5	38	4	31
Critère pris en compte	8	62	9	69
Similarités entre le groupe expérimental et les groupes contrôles				
Critère non pris en compte ou évaluation non claire	4	31	3	23
Critère pris en compte	9	69	10	77
Description détaillée du matériel de l'intervention				
Critère non pris en compte ou évaluation non claire	3	23	5	38
Critère pris en compte	10	77	8	62
Description détaillée du contenu de chaque séance de l'intervention (longueur, type d'exercice)				
Critère non pris en compte ou évaluation non claire	2	15	4	31
Critère pris en compte	11	85	9	69
Description détaillée de l'agenda et de l'organisation de l'intervention				
Critère non pris en compte ou évaluation non claire	8	62	3	23
Critère pris en compte	5	38	10	77
Raisons d'abandon des participants entre le début et la fin de l'étude				
Critère non pris en compte ou évaluation non claire	6	46	2	15
Critère pris en compte	7	54	11	85
Pertinence des résultats récoltés (*outcomes*)				
Critère non pris en compte ou évaluation non claire	4	31	4	31
Critère pris en compte	9	69	9	69
Manière dont les résultats récoltés (*outcomes*) ont été mesurés				
Critère non pris en compte ou évaluation non claire	3	23	5	38
Critère pris en compte	10	77	8	62
Pertinence des tests statistiques utilisés				
Critère non pris en compte ou évaluation non claire	3	23	6	46
Critère pris en compte	10	77	7	54
Présence d'une *p*-valeur pour chacun des résultats (*outcomes*)				
Critère non pris en compte ou évaluation non claire	3	23	2	15
Critère pris en compte	10	77	11	85
Présence d'une taille d'effet pour chacun des résultats (*outcomes*)				
Critère non pris en compte ou évaluation non claire	4	31	2	15
Critère pris en compte	9	69	11	85
Présence d'un intervalle de confiance autour de la taille d'effet dans les résultats				
Critère non pris en compte ou évaluation non claire	7	54	8	62
Critère pris en compte	6	46	5	38
Pertinence du *design* choisi en fonction de l'objectif de l'étude				
Critère non pris en compte ou évaluation non claire	2	15	6	46
Critère pris en compte	11	85	7	54
Présence d'un protocole publié avant le début de l'étude				
Critère non pris en compte ou évaluation non claire	3	23	4	31
Critère pris en compte	10	77	9	69
Absence de conflit d'intérêt				
Critère non pris en compte ou évaluation non claire	11	85	9	69
Critère pris en compte	2	15	4	31
Disponibilité des données de l'étude				
Critère non pris en compte ou évaluation non claire	8	62	6	46
Critère pris en compte	5	38	7	54

Notes. N = nombre de participants concernés ; % = pourcentage calculé sur l'échantillon total (N = 13)

Annexe 17

Répartition des réponses correctes et incorrectes des participants.

Critères considérés pour évaluer la qualité de l'article	Article R+		Article R-	
	N	%	*N*	%
Calcul de la taille de l'échantillon	Satisfaisant		Insatisfaisant	
Mauvaise réponse	0	0	0	0
Bonne réponse	9	69	11	85
Processus de randomisation entre les groupes	Satisfaisant		Insatisfaisant	
Mauvaise réponse	0	0	3	23
Bonne réponse	10	77	4	31
Similarité entre les groupes avant que l'étude ne commence	Satisfaisant		Satisfaisant	
Mauvaise réponse	1	8	2	15
Bonne réponse	7	54	8	62
Processus d'aveuglement (*blinding*) des participants	Satisfaisant		Insatisfaisant	
Mauvaise réponse	0	0	1	8
Bonne réponse	12	92	5	38
Processus d'aveuglement (*blinding*) des expérimentateurs responsables de l'intervention	Satisfaisant		Insatisfaisant	
Mauvaise réponse	0	0	1	8
Bonne réponse	8	62	7	54
Processus d'aveuglement (*blinding*) des parents	Satisfaisant		Insatisfaisant	
Mauvaise réponse	0	0	0	0
Bonne réponse	11	85	5	38
Processus d'aveuglement (*blinding*) des évaluateurs (pré et post- intervention)	Satisfaisant		Insatisfaisant	
Mauvaise réponse	0	0	1	8
Bonne réponse	7	54	7	54
Similarité des caractéristiques des participants	Satisfaisant		Insatisfaisant	
Mauvaise réponse	1	8	5	38
Bonne réponse	7	54	4	31
Similarités entre le groupe expérimental et les groupes contrôles	Satisfaisant		Insatisfaisant	
Mauvaise réponse	1	8	8	62
Bonne réponse	8	62	2	15
Description détaillée du matériel de l'intervention	Satisfaisant		Insatisfaisant	
Mauvaise réponse	1	8	5	38
Bonne réponse	9	69	3	23
Description détaillée du contenu de chaque séance de l'intervention (longueur, type d'exercice)	Satisfaisant		Insatisfaisant	
Mauvaise réponse	3	23	6	46
Bonne réponse	8	62	3	23
Description détaillée de l'agenda et de l'organisation de l'intervention	Satisfaisant		Insatisfaisant	
Mauvaise réponse	1	8	8	62
Bonne réponse	4	31	2	15
Raisons d'abandon des participants entre le début et la fin de l'étude	Satisfaisant		Insatisfaisant	
Mauvaise réponse	2	15	10	77
Bonne réponse	5	38	1	8
Pertinence des résultats récoltés (*outcomes*)	Satisfaisant		Insatisfaisant	
Mauvaise réponse	1	8	5	38
Bonne réponse	9	69	4	31
Manière dont les résultats récoltés (*outcomes*) ont été mesurés	Satisfaisant		Insatisfaisant	
Mauvaise réponse	1	8	6	46
Bonne réponse	9	69	2	15
Pertinence des tests statistiques utilisés	Satisfaisant		Insatisfaisant	
Mauvaise réponse	1	8	4	57
Bonne réponse	10	77	3	43
Présence d'une *p*-valeur pour chacun des résultats (*outcomes*)	Satisfaisant		Satisfaisant	
Mauvaise réponse	0	0	1	8
Bonne réponse	10	77	10	77
Présence d'une taille d'effet pour chacun des résultats (*outcomes*)	Satisfaisant		Satisfaisant	
Mauvaise réponse	2	15	0	0
Bonne réponse	7	54	11	85
Présence d'un intervalle de confiance autour de la taille d'effet dans les résultats	Satisfaisant		Insatisfaisant	
Mauvaise réponse	5	38	0	0
Bonne réponse	1	8	5	38
Pertinence du *design* choisi en fonction de l'objectif de l'étude	Satisfaisant		Insatisfaisant	
Mauvaise réponse	1	8	6	46
Bonne réponse	10	77	1	8
Présence d'un protocole publié avant le début de l'étude	Satisfaisant		Insatisfaisant	
Mauvaise réponse	0	0	0	0
Bonne réponse	9	69	9	69
Absence de conflit d'intérêt	Satisfaisant		Insatisfaisant	
Mauvaise réponse	0	0	0	0
Bonne réponse	2	15	4	31
Disponibilité des données de l'étude	Satisfaisant		Insatisfaisant	
Mauvaise réponse	1	8	1	8
Bonne réponse	4	31	6	46

Notes. N = nombre de participants concernés ; % = pourcentage calculé sur l'échantillon total (N = 13)

Annexe 18

Analyse de l'association entre la profession et l'identification de biais méthodologiques dans l'article *reporting* +.

Profession	Biais identifiés article R+		Total
	Non	Oui	
Chercheurs	7	6	13
Cliniciens	5	2	7
Total	12	8	20

	p	φ
Test exact de Fisher	.642	0.171

Notes. φ = coefficient Phi

Annexe 19

Résultats des tests de normalité et d'homogénéité des variances, et synthèse des tests statistiques appliqués.

Test de normalité – Shapiro-Wilk	**Statistique**	***p***
Compréhension de l'introduction	0.66	<.001*
Compréhension de la méthodologie	0.90	.043*
Compréhension des résultats	0.76	<.001*
Compréhension de la discussion	0.62	< .001*
Évaluation de la qualité méthodologique	0.65	< .001*
Suffisamment d'informations pour reproduire l'intervention	0.76	< .001*
Suffisamment d'informations pour juger l'efficacité	0.84	.004*
Suffisamment d'informations pour juger l'utilité	0.89	.024*

*Notes. * p < .05*

Homogénéité des variances – Levene	***F***	***p***
Compréhension de l'introduction	0.11	.743
Compréhension de la méthodologie	1.98	.176
Compréhension des résultats	0.28	.604
Compréhension de la discussion	1.67	.213
Évaluation de la qualité méthodologique	15.43	<.001*
Suffisamment d'informations pour reproduire l'intervention	0.004	.953
Suffisamment d'informations pour juger l'efficacité	0.89	.357
Suffisamment d'informations pour juger l'utilité	0.21	.652

*Notes. * p < .05*

Variables	**Normalité**	**Homogénéité**	**Choix du test**
Compréhension de l'introduction	Non respectée	Respectée	Test de Mann-Whitney
Compréhension de la méthodologie	Non respectée	Respectée	Test de Mann-Whitney
Compréhension des résultats	Non respectée	Respectée	Test de Mann-Whitney
Compréhension de la discussion	Non respectée	Respectée	Test de Mann-Whitney
Évaluation de la qualité méthodologique	Non respectée	Non respectée	Test de Mann-Whitney
Suffisamment d'informations pour reproduire l'intervention	Non respectée	Respectée	Test de Mann-Whitney
Suffisamment d'informations pour juger l'efficacité	Non respectée	Respectée	Test de Mann-Whitney
Suffisamment d'informations pour juger l'utilité	Non respectée	Respectée	Test de Mann-Whitney

Résumé

Ce mémoire, réalisé en binôme partiel, s'intéresse aux défis rencontrés par les professionnels de la neuropsychologie dans l'intégration de l'*Evidence-Based Practice* (EBP), et plus spécifiquement à l'intégration du pilier *recherche*. Tandis que mon binôme s'est intéressé aux neuropsychologues cliniciens, ce travail porte spécifiquement sur les chercheurs, explorant leurs attitudes à l'égard de la recherche scientifique ainsi que leurs pratiques en matière de lecture critique.

Si l'EBP est aujourd'hui reconnue comme un modèle de référence en santé, son application concrète en neuropsychologie clinique reste entravée par de nombreuses difficultés, notamment en ce qui concerne la lecture et l'évaluation des articles scientifiques. Plusieurs études ont mis en évidence des lacunes méthodologiques importantes dans la littérature actuelle, y compris les Essais Contrôlés Randomisés (ECR), pourtant considérés comme le *« gold standard »* en matière de recherche. Face à ce constat, il devient essentiel de s'interroger non seulement sur la capacité des cliniciens à mobiliser leur esprit critique, mais également sur celle des chercheurs eux-mêmes. Dans cette optique, notre étude propose une approche combinant un questionnaire à une tâche concrète de lecture critique de deux ECR rigoureusement sélectionnés pour leur différence de qualité méthodologique (évaluée sur base des grilles CONSORT-SPI et RoB2). Les chercheurs participants ont été invités à lire ces articles, à évaluer leur confiance dans les résultats, à identifier d'éventuels biais, à juger la qualité méthodologique et à indiquer les critères sur lesquels ils s'étaient basés pour leur évaluation.

Nos résultats suggèrent que, même si les chercheurs sont en mesure d'identifier des biais méthodologiques et de déterminer quel article est méthodologiquement supérieur, leur lecture critique s'appuie principalement sur des jugements intuitifs, et non sur l'utilisation d'outils d'analyse critique. Par ailleurs, nos résultats soulignent l'écart persistant entre la recherche scientifique et la pratique clinique dans le sens où les chercheurs jugent l'intervention décrite comme reproductible, tandis que les cliniciens peinent à en percevoir l'applicabilité concrète dans leur pratique quotidienne. En définitive, ce mémoire souligne la nécessité de renforcer la formation à la lecture critique tant pour les cliniciens que pour les chercheurs, et plaide pour un rapprochement entre ces deux acteurs. Favoriser une collaboration plus étroite offrirait l'opportunité d'un partage mutuel des compétences et des attentes, ce qui faciliterait une intégration plus pertinente et efficace des données probantes en pratique clinique.