L'Evidence-Based Assessment en neuropsychologie clinique francophone : Revue des qualités psychométriques des outils d'évaluation et méta-analyse des indices de fidélité test-retest

Auteur : Dieu, Alix
Promoteur(s) : Willems, Sylvie; Burnay, Jonathan
Faculté : þÿ F a c u l t é d e P s y c h o l o g i e , L o g o p é d i e e t S c i e n c e s d e l E d u c a t i o n
Diplôme : Master en sciences psychologiques, à finalité spécialisée en psychologie clinique
Année académique : 2020-2021
URI/URL : http://hdl.handle.net/2268.2/13441

L'Evidence-Based Assessment en neuropsychologie clinique francophone

Revue des qualités psychométriques des outils d'évaluation et méta-analyse des indices de fidélité test-retest

Mémoire présenté en vue de l'obtention du diplôme de master en Psychologie clinique

Université de Liège

Faculté de Psychologie, Logopédie et Sciences de l'Éducation

Année académique 2020-2021

DIEU Alix

LIÈGE université

Promotrice : S. Willems

Co-promoteur : J. Burnay

Lecteurs : M. Wansard, X. Schmitz

Remerciements

Pour commencer, je souhaite remercier ma promotrice, Madame Willems, ainsi que mon co-promoteur, Monsieur Burnay, pour leur bienveillance, leur aide et leurs précieux conseils tout au long de la réalisation de ce mémoire qui n'aurait jamais vu le jour sans leur présence.

Je tiens également à remercier Madame Wansard et Monsieur Schmitz, pour le temps qu'ils auront accordé à la lecture de ce mémoire.

Merci aux responsables de la testothèque de la faculté de Psychologie, pour leur patience à chacune de mes venues malgré mes nombreuses requêtes.

Aux personnes qui m'ont aidé dans la relecture de mon travail.

À ma famille et mes proches, qui ont confiance en moi.

À mes amis, qui ont contribué à me changer les idées quand j'en avais plus que besoin.

Enfin, merci à Théo et Martin, pour m'avoir laissée plus ou moins tranquille quand je devais travailler.

Table des matières

I. Introduction

Actuellement, les domaines médical et paramédical cherchent de plus en plus à promouvoir une pratique basée sur les preuves, aussi appelée « Evidence-Based Practice » (EBP). Le champ de la neuropsychologie clinique s'inscrit dans ce contexte et vise également à tendre vers cette démarche. Pour les neuropsychologues, l'application d'une démarche EBP doit s'opérer dès l'étape de l'évaluation cognitive : elle se concrétise en choisissant les outils d'évaluation les plus adéquats pour les besoins spécifiques du patient rencontré. Le professionnel qui veut intégrer sa démarche dans une pratique EBP pourra sélectionner l'un ou l'autre outil d'évaluation, en s'appuyant sur les caractéristiques psychométriques de celui-ci, ainsi que sur son expertise professionnelle, tout en considérant la problématique spécifique au patient. On observe que les neuropsychologues ont tendance à employer certains outils de façon récurrente. Cependant, rien ne nous permet d'affirmer à ce jour que leur choix s'opère dans un souci de mener une démarche EBP. Dans ce cadre, notre travail visera à déterminer si les outils les plus employés en neuropsychologie clinique francophone le sont en raison de leurs bonnes qualités psychométriques.

A. Historique de l'Evidence-Based Practice

1. L'Evidence-Based Medicine (EBM)

À l'origine, le paradigme « Evidence-Based Medicine » (EBM) s'est développé en 1992 dans le domaine médical (Evidence-Based Medicine Working Group, 1992). L'objectif était de développer une méthode de prise de décision clinique, en intégrant les meilleurs résultats issus de la recherche directement dans la pratique de soin. L'EBM est définie comme « l'usage conscient, explicite et judicieux des meilleures preuves actuelles dans la prise de décision concernant le soin de patients individuels » (Sackett et al., 1996). Selon le « Three-Circle Model of Evidence-Based Clinical Decisions » (Haynes et al., 2002) représenté en figure 1, l'EBM se veut donc être une pratique de prise de décision basée sur la prise en compte de trois piliers : (1) l'état clinique et les circonstances du patient qui l'ont mené à consulter un médecin, (2) les préférences du patient concernant le traitement, et (3) les preuves issues de la recherche. À l'intersection de ces trois piliers se situe l'expertise clinique, qui a une place centrale dans la prise de décision *evidence-based* : elle regroupe les compétences générales de la pratique clinique ainsi que l'expérience personnelle du professionnel. Cette expertise clinique doit

inclure et équilibrer les trois piliers afin d'établir un diagnostic et un pronostic adéquats, et sélectionner le traitement à appliquer.

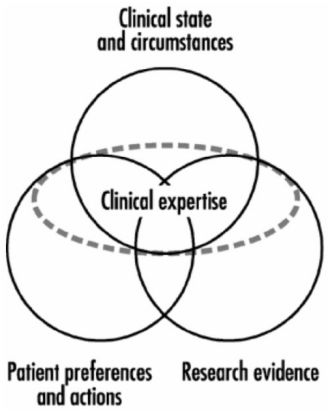

Figure 1. Updated Three-Circle Model of Evidence-Based Clinical Decisions (Haynes et al., 2002, figure 2)

Plus tard, à la suite du développement de l'EBM et de l'expansion de ses principes aux disciplines liées à la santé, certains auteurs (Satterfield et al., 2009) proposent d'aller vers un modèle EBP à visée transdisciplinaire (figure 2). Ce modèle révisé de l'EBP comprend : (1) les meilleures preuves issues de la recherche, (2) l'expertise du praticien, (3) l'état et les caractéristiques, valeurs et préférences du patient, (4) la prise de décision au centre, et enfin (5) l'environnement et le contexte organisationnel qui englobent l'ensemble.

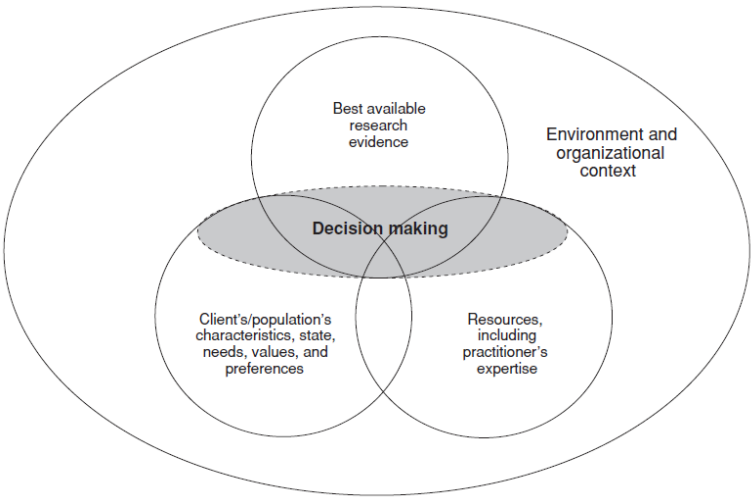

Figure 2. EBP Model révisé par Satterfield et al. (2009, figure 5)

2. De l'EBM à l'Evidence-Based Practice in Psychology (EBPP)

Au fil des années, les principes de l'EBM se sont donc répandus à plusieurs disciplines liées à la santé (Satterfield et al., 2009). En psychologie particulièrement, l'American Psychological Association (APA) est amenée, en 2005, à former « l'APA Presidential Task Force on EBP », un groupe de travail mis en place dans l'objectif de définir l'Evidence-Based Practice in Psychology (EBPP), en cohérence avec les précédents travaux en EBM. Après avoir mené une revue de la littérature sur le sujet et différentes délibérations, l'APA Presidential Task Force on EBP (2006) s'accorde pour reconnaître une valeur égale aux trois piliers de l'EBM qui, de plus, sont étroitement liés. En conséquence, le groupe définit l'EBPP de cette manière : « L'EBPP est l'intégration des (1) meilleures recherches disponibles avec (2) l'expertise clinique dans (3) le contexte des caractéristiques du patient, de sa culture et de ses préférences ». (1) Les meilleures preuves disponibles issues de la recherche dépendent directement de la question posée (par exemple, si nous nous interrogeons sur l'efficacité d'un traitement, le design expérimental le plus adéquat pour répondre à notre question sera une étude par essai clinique randomisé). De manière plus générale, il est admis que les meilleures preuves sont issues des revues systématiques (OCEBM Levels of Evidence Working Group, 2011). (2) L'expertise clinique est l'approche qui consiste à spécifier, opérationnaliser et s'entraîner aux compétences requises pour exécuter des pratiques spécifiques. (3) La prise en compte des préférences du patient vise à respecter et aider les patients à mettre au clair leurs valeurs et préférences concernant le traitement à appliquer, afin de parvenir à une prise de décision partagée (Spring, 2007).

En psychologie, un clinicien adoptera une démarche EBP pour répondre à une question qu'il se pose concernant un diagnostic, pour sélectionner une action de prévention à mener ou encore un choix de traitement à pratiquer (Durieux et al., 2017). La démarche EBP consiste alors en 5 étapes (Straus et al., 2011, cités par Durieux et al., 2017) : (1) poser une question clinique structurée et précise, (2) rechercher les meilleures données issues de la recherche, (3) évaluer ces données de manière critique, (4) appliquer les résultats dans la pratique, (5) évaluer la performance.

3. L'EBP en neuropsychologie clinique

La neuropsychologie est un domaine relativement récent de la psychologie qui s'est développé au fil des années et des avancées technologiques, notamment grâce au développement de nouvelles techniques d'imagerie cérébrale. C'est essentiellement le rôle du clinicien auprès du patient qui a changé. Actuellement, selon l'Unité de Neuropsychologie de l'Université de Liège (Faculté de Psychologie, Logopédie et Sciences de l'Éducation) (2018), bien que le neuropsychologue s'intéresse aux conséquences cognitives et comportementales de dysfonctionnements neurologiques, il reste avant tout un psychologue clinicien, dont l'objectif est d'aider le patient à retrouver une qualité de vie optimale à travers son évaluation, puis sa prise en charge.

Dans ce contexte, la neuropsychologie clinique fait donc partie intégrante de la psychologie clinique et développe, elle aussi, l'ambition de s'inscrire dans une pratique EBP. Celle-ci est définie comme suit : « La pratique *evidence-based* en neuropsychologie clinique est un pattern de pratique clinique centré sur des valeurs qui tentent d'intégrer la meilleure recherche issue des études de population afin d'informer les décisions cliniques concernant les individus au sein du contexte du prestataire d'expertise et des valeurs individuelles du patient, dans le but de maximiser les résultats cliniques et la qualité de vie pour le patient, d'une façon rentable, tout en répondant aux préoccupations et besoins du prestataire demandeur » (Chelune, 2017 [notre traduction]). Le schéma qui suit (figure 3) illustre la démarche EBP en neuropsychologie clinique comme un processus intégratif qui reconnecte la recherche scientifique à la pratique clinique.

Figure 3. EBP comme un processus intégratif (Chelune, 2017, figure 7.1)

B. L'évaluation en neuropsychologie clinique — Evidence-Based Assessment (EBA)

Après avoir identifié clairement la demande du patient lors de l'étape initiale de l'anamnèse, la première mission du neuropsychologue consiste en son évaluation. Seule une évaluation adéquate du fonctionnement cognitif du patient lui permettra de proposer ensuite une prise en charge individualisée optimale et spécifique à ses difficultés. Le neuropsychologue peut être amené à poursuivre différents objectifs selon la demande et l'émetteur de la demande : contribuer au diagnostic, planifier une prise en charge individualisée, évaluer l'efficacité de la prise en charge, fournir un avis dans le cadre d'une expertise, et enfin, mener des recherches (Lezak et al., 2004).

À la suite de cette étape d'identification de la demande, le clinicien pourra mettre en place une démarche évaluative *evidence-based* (Evidence-Based Assessment, EBA) qui permettra d'y répondre de façon appropriée en sélectionnant les outils d'évaluation adéquats. Ainsi, tout en considérant la demande du patient, le neuropsychologue se basera sur son expertise clinique ainsi que sur les données issues de la littérature pour récolter les informations nécessaires concernant les données psychométriques relatives aux outils qu'il envisage d'employer. Des auteurs proposent, entre autres, deux stratégies qui permettront de rendre l'entreprise réalisable (Youngstrom et al., 2017). Premièrement, se concentrer sur les caractéristiques du test les plus pertinentes selon la tâche à accomplir. En effet, l'un ou l'autre outil sera à privilégier selon ses caractéristiques psychométriques et l'objectif poursuivi lors de l'évaluation. Ensuite, ils proposent de se concentrer sur les caractères suffisants et satisfaisants (*satisficing*), c.-à-d. en utilisant une méthode suffisamment bonne, plutôt que de rechercher à tout prix la meilleure méthode, qui n'existe pas dans tous les cas. Dans ce souci d'une démarche EBA, nous sommes amenés à nous questionner sur les différentes caractéristiques psychométriques que peuvent présenter les outils d'évaluation régulièrement utilisés par les neuropsychologues, ainsi qu'à nous demander lesquelles de ces qualités psychométriques sont essentielles selon l'objectif visé par le clinicien lors de l'évaluation.

1. Les caractéristiques psychométriques des outils comme point de départ à l'EBA

Les informations psychométriques relatives aux épreuves utilisées sont essentielles à prendre en compte. En effet, les considérer et les comprendre permettra au clinicien de choisir les outils adéquats, et d'ensuite assurer une interprétation correcte des résultats obtenus par le patient. Cependant, il s'avère que la compréhension de ces indices psychométriques et l'importance qui leur est accordée sont souvent insuffisantes (Bowden, 2017). La considération de ces données est pourtant indispensable pour faire preuve d'une démarche EBP et pour sélectionner l'outil d'évaluation le plus adéquat selon l'objectif visé. D'ailleurs, la connaissance de la psychométrie fait partie des qualifications requises par l'APA pour les utilisateurs de tests (APA, 2000). En outre, employer à tort un outil qui fournit des mesures de pauvre qualité psychométrique pourrait entraîner le clinicien à formuler des jugements erronés, voire néfastes pour son patient (Haynes et al., 2019).

De plus, il faut noter que les qualités psychométriques des outils dépendent directement de l'échantillon de référence. En effet, le degré de corrélation d'un test avec le construit qu'il est effectivement censé mesurer, peut considérablement varier d'un contexte à l'autre. Il est idéal, dans la mesure du possible, d'utiliser des études de validation portant sur des échantillons au plus proche du patient évalué (Van Meter, 2020). Cela implique notamment que des études de validation devraient idéalement être menées de façon nationale, car la validité de l'évaluation et des résultats peut être atténuée lorsqu'un outil est utilisé de façon interculturelle (Maltzman, 2013). Actuellement, des auteurs cherchent à généraliser les qualités psychométriques des tests en menant des méta-analyses qui visent à synthétiser plusieurs études de validation réalisées pour un même outil (American Educational Research Association [AERA] et al., 2014). Cependant, cette tentative de généralisation est récente et nous estimons qu'elle est encore peu développée en neuropsychologie.

Dans ce contexte, nous estimons que, pour être applicables dans le cadre de la neuropsychologie francophone européenne, les données visant à valider un outil doivent idéalement avoir été recueillies sur base d'un échantillon franco-européen (Colombo et al., 2016 ; Maltzman, 2013). Plusieurs outils régulièrement utilisés ont fait l'objet d'une étude de validation[1] francophone.

[1] Nous entendons par « étude de validation », toute étude qui propose le calcul d'indices de fidélité (test-retest, inter-juges, consistance interne), de validité (de construit, convergente, concourante, discriminante), ou qui propose une standardisation ou des normes diagnostiques pour un outil spécifique.

C'est le cas par exemple de la batterie WAIS-IV (Wechsler, 2011). D'autres, en revanche, n'ont pas fait l'objet d'une validation francophone et sont pourtant largement employés, comme la batterie NEPSY-II (Korkman et al., 2012), qui n'est que partiellement normée en France. Cette batterie repose principalement sur un échantillon américain supposant que les données américaines seraient statistiquement proches des données issues de la culture francophone (Roy, 2015).

Les outils employés par les neuropsychologues doivent rencontrer différents critères de qualité psychométrique, à savoir : (a) des normes, (b) des preuves de validité, (c) un certain niveau de fidélité des mesures (Board of Directors, 2007). Les études de validation peuvent fournir ces différentes informations que nous allons passer en revue ci-dessous.

a. Données normatives

Une fois l'évaluation neuropsychologique effectuée, le praticien se réfère aux données normatives pour interpréter la performance du patient. Ces données correspondent aux scores obtenus par un échantillon de la population à une épreuve spécifique. Grâce à ces données, le clinicien pourra situer le score obtenu par son sujet à un test par rapport aux scores de sa population de référence (Colombo et al., 2016). Ainsi, disposer de normes est indispensable pour que le neuropsychologue puisse tirer des conclusions concernant la performance de son patient. Les données normatives peuvent s'accompagner de différentes informations pertinentes à prendre en considération lors de l'interprétation des résultats par le clinicien.

Premièrement, l'année de production des normes influence l'exactitude de l'interprétation de la performance cognitive du patient. En effet, une étude a démontré que l'effet Flynn, bien connu et démontré plus d'une fois (Flynn, 1984 ; Flynn et al., 2007), ne s'applique pas qu'aux tests évaluant le quotient intellectuel. Des chercheurs ont démontré que cet effet existe bel et bien à travers plusieurs outils d'évaluation neuropsychologique (Dickinson & Hiscock, 2011). Ainsi, l'utilisation de normes « trop vieilles » provoquerait une surestimation de la performance du sujet, car il serait comparé à une cohorte avec une moyenne de performance cognitive inférieure. En résumé, au plus les normes disponibles sont vieilles, au plus le professionnel devra faire preuve de prudence dans ses conclusions.

Deuxièmement, la représentativité de l'échantillon normatif par rapport au patient évalué influence également la qualité des interprétations futures. L'origine de la population de normalisation est dès lors un critère essentiel. On observe régulièrement que les neuropsychologues ont tendance à employer des normes issues d'autres pays lorsque des

normes ne sont pas disponibles dans leur pays d'origine. Comme le révèle une étude (Branco Lopes et al., 2019) sondant des neuropsychologues français, 99.1% des répondants affirment utiliser des données normatives issues de leur propre pays, mais 53% d'entre eux avouent également utiliser des données normatives provenant d'autres pays. La raison principale évoquée était simplement le manque de données pour leur pays. Or, des auteurs ont démontré que cet usage pouvait s'avérer extrêmement problématique, menant à considérer à tort des résultats normaux comme déficitaires (Alberto & Marcopulos, 2008 ; Raudeberg et al., 2018).

Un troisième point, qui découle du précédent, est la distinction des normes en plusieurs groupes d'âge et catégories socioprofessionnelles. En effet, cela permet au clinicien de disposer de normes issues d'une population qui se rapproche autant que possible de son patient. Chez les enfants, ces groupes doivent idéalement regrouper des tranches d'âge plus restreintes que pour les adultes, étant donné l'évolution plus rapide des fonctions cognitives en lien avec la maturation cérébrale toujours en cours, et plus particulièrement la maturation du cortex préfrontal, qui est directement lié au fonctionnement exécutif (Chevalier, 2010). Pour les adultes et personnes âgées, en revanche, les tranches d'âge peuvent être plus larges. À noter qu'il est également préférable de disposer du nombre de sujets constituant le groupe, ainsi que de leur moyenne d'âge. Enfin, pour les adultes et les personnes âgées, selon le processus cognitif évalué, il est important de disposer de normes regroupées par genre, niveau scolaire et/ou niveau professionnel, afin de pouvoir situer au mieux le sujet par rapport à sa population de référence.

Quatrièmement, le clinicien devra également avoir connaissance de la taille de l'échantillon utilisé et de la distribution des scores dans cet échantillon pour pouvoir nuancer ses interprétations. Plus l'échantillon est large, plus sa distribution se rapprochera de la distribution réelle des scores dans la population, qui est en moyenne généralement normale (Brooks et al., 2009). Si la distribution est normale, nous pourrons alors tirer des conclusions sur base des moyennes et écarts-types. En revanche, s'il s'avère qu'elle est anormale (ou que sa normalité n'a pas été démontrée), il sera préférable de disposer de mesures de type percentile afin d'éviter les erreurs d'interprétation (Brooks et al., 2009).

En conclusion, ce sera au clinicien, grâce à son expertise, de prendre en compte toutes ces variations possibles concernant les données normatives, de choisir les plus adéquates si plusieurs choix s'offrent à lui, et de faire preuve de nuance dans ses conclusions lorsqu'il le juge nécessaire.

b. Validité

Un autre élément essentiel à prendre en compte est la notion de validité. Une bonne validité permettra de réaliser une évaluation précise du fonctionnement neuropsychologique du patient, et d'assurer par la suite une prise en charge de qualité (Riley et al., 2017). Selon les « Standards for Educational and Psychological Testing » (AERA et al., 2014), « le concept de validité renvoie au degré auquel les preuves et la théorie soutiennent les interprétations des scores au test pour les emplois prévus au test ». En d'autres termes, la validité assure la capacité qu'une épreuve a effectivement mesurer ce qu'elle prétend mesurer.

La notion de validité comprend plusieurs spécificités importantes à mentionner. Premièrement, elle doit toujours être contextualisée : ce n'est pas le test qui est jugé valide ou non, mais ce sont les inférences qui seront réalisées sur base des résultats qui seront jugées valides ou non (par exemple, un test d'admission peut s'avérer valide pour prédire les résultats scolaires d'étudiants universitaires, mais non valide pour prédire la réussite professionnelle des étudiants universitaires une fois diplômés). Deuxièmement, la validité est un concept unitaire, qui varie selon différents degrés et s'évalue grâce à une accumulation de preuves fournies par différentes méthodes (AERA et al., 2014). Bien que les auteurs de la dernière version des Standards for Educational and Psychological Testing (AERA et al., 2014) ont fait le choix de ne pas distinguer différents types de validité, nous avons décidé de préserver les distinctions issues de la version précédente par souci de clarté (AERA et al., 1999). Nous distinguerons ainsi différentes méthodes qui permettent d'évaluer la validité d'un outil.

Tout d'abord, la *validité de construit* réfère au processus visant à valider simultanément les mesures des construits psychologiques ainsi que les théories dont ces construits font partie, afin de vérifier dans quelle mesure on évalue effectivement le construit désiré (Strauss & Smith, 2009). Les inférences réalisées à partir d'un test seront jugées valides en référence au cadre conceptuel initialement défini par les concepteurs du test. Dans ce contexte, le point de départ pour la conception et la validation d'un test consiste en la définition détaillée du cadre conceptuel, c.-à-d. du construit psychologique visé par l'évaluation dans lequel s'inscrit également la signification explicite de l'interprétation prévue des scores à un test (AERA et al., 2014). Les analyses factorielles (exploratoires ou confirmatoires) permettent d'analyser la structure interne d'un outil afin d'évaluer la concordance de la relation entre les constituants d'un test et le construit sur lequel se basent les interprétations.

Ensuite, d'autres méthodes fournissent des preuves sur base des relations avec d'autres variables externes. Parmi celles-ci, on peut distinguer : (1) la *validité convergente ou concomitante*, (2) la *validité divergente*, (3) la *validité prédictive*, (4) la *validité concourante*. (1) La première méthode implique que l'on s'attend à obtenir des résultats similaires chez un même sujet pour deux mesures évaluant un concept lié. (2) La seconde permet de vérifier que le test mesure spécifiquement le concept visé par l'évaluation et non un concept voisin. Par exemple, on s'attend à obtenir une corrélation faible entre les scores obtenus à un questionnaire de dépression et un questionnaire d'anxiété (Laveault & Grégoire, 2014). Les deux méthodes suivantes fournissent des preuves sur la relation entre les scores au test et un critère pertinent (par exemple, une autre épreuve supposée évaluer le même construit) : soit (3) de façon à prédire des résultats futurs à un autre test soit (4) de façon simultanée, (Laveault & Grégoire, 2014). Cette dernière méthode permet de s'assurer que les résultats obtenus ne sont pas impactés par le processus de mesure employé. Autrement dit, on s'attend à obtenir une corrélation élevée entre deux méthodes différentes évaluant le même construit issu du même cadre conceptuel.

Enfin, nous distinguerons la *validité discriminante*, utile lorsque l'outil a été constitué dans l'objectif de détecter l'appartenance du patient à une population clinique spécifique. En effet, elle permet de fournir des données diagnostiques qui aident le clinicien à évaluer la probabilité pour le patient de présenter une condition spécifique (par exemple, une démence) (Colombo et al., 2016). Ces données diagnostiques, qui constituent des preuves de la validité discriminante des mesures fournies par l'outil, peuvent présenter différents degrés de (1) sensibilité et (2) spécificité. (1) La sensibilité d'un test est définie par la proportion de personnes atteintes du diagnostic qui obtiendront effectivement un résultat positif au test diagnostic : on parle de « vrais positifs ». (2) La spécificité est la proportion de personnes non atteintes et qui auront effectivement un résultat négatif au test diagnostique, c.-à-d. les « vrais négatifs » (Akobeng, 2006). Autrement dit, un test avec une haute sensibilité permettra d'éviter de juger une personne atteinte du diagnostic comme « non porteuse » (soit, un faux négatif), tandis qu'une haute spécificité permettra d'éviter de poser à tort un faux diagnostic (soit, un faux positif). L'idéal recherché est de déterminer un « score seuil » à partir duquel le test disposerait de 100% de sensibilité et 100% de spécificité. C'est cependant rarement le cas, car il existe toujours un risque d'erreur. Il faudra alors choisir un score seuil qui permettra au mieux de distinguer les vrais positifs des vrais négatifs, en sachant qu'augmenter l'un en déplaçant le seuil aura pour effet de diminuer l'autre. Selon le diagnostic visé par l'évaluation, il sera plus important d'avoir

une haute sensibilité ou spécificité. Par exemple, dans le cadre du diagnostic de la maladie d'Alzheimer, bien qu'un diagnostic précoce (grâce à une haute sensibilité) permette une prise en charge plus efficace, utilisant un maximum des fonctions préservées, une haute spécificité peut être préférable (afin d'éviter les faux positifs), étant donné les répercussions et préjudices que peut provoquer ce type de diagnostic sur la vie du patient, notamment à cause de la stigmatisation liée à cette condition (De Mendonça Lima et al., 2003). Il en est de même dans le cadre de l'expertise, lorsqu'il s'agit de détecter des comportements « de manque d'effort » (autrefois appelés « comportements de simulation »), car il est essentiel d'éviter de considérer à tort un sujet comme simulateur (Meulemans & Seron, 2004). En revanche, lorsque l'évaluation a pour objectif de contribuer à un diagnostic afin d'obtenir le droit à des aménagements scolaires pour un enfant, il vaudra mieux privilégier un outil à haute sensibilité, au détriment d'une meilleure spécificité. En effet dans ce cas, identifier de faux positifs ne devrait pas entraîner de conséquences dommageables pour le patient.

c. Fidélité

En complément à la notion de validité arrive le concept de fidélité. On considère qu'un instrument fournit une mesure fidèle si celle-ci est reproductible (AERA et al., 2014). La fidélité peut être définie comme « la consistance de l'information récoltée à partir de tests » (Bowden & Finch, 2017). Dans ce contexte, la fidélité se rapporte aux scores fournis par l'outil et non à l'outil lui-même.

La fidélité des scores permet d'estimer l'erreur standard de mesure (ESM), qui représente la déviation des scores d'un sujet lors de la réplication de la procédure d'évaluation. L'ESM fournit une estimation du manque de consistance des scores donnés par un test. Ainsi, plus un test fournit une mesure fidèle, au plus l'ESM sera faible (relation inverse). Une ESM faible permettra au clinicien de disposer d'un score plus précis de la performance réelle du sujet, sans quoi l'intervalle de confiance (IC) sera trop large et donc peu précis (Brooks et al., 2009).

La fidélité peut être mesurée via différentes méthodes non équivalentes et non corrélées (Calamia et al., 2013) : (1) la *mesure de la consistance interne de l'outil*, (2) la *méthode inter-juges* et (3) la *méthode test-retest*.

Premièrement, la *consistance interne* permet de mesurer la concordance entre plusieurs parties ou plusieurs items d'un même outil (AERA et al., 2014). Celle-ci fournit une mesure de fidélité moins directe et peut être mesurée au moyen de différentes méthodes (par exemple : méthode split-half, coefficient alpha de Cronbach, etc.).

Deuxièmement, la *méthode de fidélité inter-juges* implique que l'administration indépendante d'un même test par deux évaluateurs qualifiés fournisse les mêmes résultats. Pour ce faire, il faut avant toute chose que les consignes d'administration soient communiquées de manière claire, précise et qu'elles soient applicables par l'évaluateur (Krippendorff, 2016).

Troisièmement, la *méthode de fidélité test-retest* permet d'évaluer la stabilité dans le temps des mesures fournies par l'outil. Elle consiste en l'administration d'un outil d'évaluation identique à un même sujet, en deux temps séparés par un intervalle variable. Ces réplications impliquent que le construit évalué soit stable dans le temps (AERA et al., 2014), ce qui est le cas des fonctions cognitives chez le sujet sain. Lorsqu'on mesure la fidélité test-retest d'un outil, il faut prêter attention à l'effet d'entraînement : un test peut fournir une mesure très fidèle, mais celle-ci peut différer en phase de retest à cause du biais d'entraînement (Calamia et al., 2012). Ce biais se traduit par une amélioration de la performance du sujet grâce à l'exposition répétée à une même épreuve (Heilbronner et al., 2010).

Différentes raisons peuvent expliquer ce biais d'entraînement. En effet, plusieurs variables modératrices du phénomène ont pu être mises en évidence dans la littérature scientifique. Nous en distinguerons trois types : (1) la méthodologie employée pour l'évaluation de la fidélité test-retest, (2) les caractéristiques liées à l'outil, et (3) les caractéristiques liées aux sujets.

1) Méthodologie test-retest

Tout d'abord, il faut considérer la taille de l'échantillon utilisé pour le calcul de l'indice de fidélité test-retest. Cette variable n'a pas d'effet modérateur, mais agit plutôt sur la précision de la corrélation fournie. Watson (2004) recommande un échantillon de minimum 300 sujets (400 au mieux) afin d'assurer une mesure suffisamment précise de la fidélité test-retest.

Ensuite, la durée de l'intervalle entre les deux passations peut modérer l'importance de l'effet d'entraînement. Il n'existe pas de recommandation absolue concernant l'étendue de l'intervalle. Celle-ci peut varier selon la nature du construit évalué et sa stabilité dans le temps. Seul un construit stable pourra entraîner une mesure stable, ce qui est généralement le cas des fonctions cognitives évaluées par les neuropsychologues (contrairement à des mesures plus variables, comme l'humeur). La longueur de l'intervalle permet de diminuer la familiarité au test du sujet, limitant ainsi l'effet d'entraînement (Calamia et al., 2012; Lievens et al., 2007; Scharfen et al., 2018). Un intervalle trop court peut permettre au sujet de se souvenir trop aisément de la première passation (Watson, 2004), tandis qu'un intervalle trop long peut augmenter la

probabilité que le sujet ait vécu un événement pouvant créer de réels changements significatifs chez celui-ci (Calamia et al., 2013; Watson, 2004). Une méta-analyse a mis en évidence que l'effet d'entraînement persiste même dans des intervalles allant jusqu'à cinq années pour certaines épreuves, bien que les gains observés soient moindres que lors d'intervalles plus courts (Calamia et al., 2012). Les auteurs ont suggéré que cette diminution de gains de performance pouvait être due à une diminution effective de l'effet d'entraînement, ou à un déclin cognitif chez le sujet adulte, ou encore, aux deux. En résumé, il n'existe pas de durée idéale pour la période séparant les deux séances d'administration.

Enfin, certains auteurs tentent de limiter l'effet d'entraînement en employant des formes parallèles d'une même épreuve (par exemple, une même épreuve évaluant le fonctionnement mnésique peut s'accompagner d'une forme parallèle constituée d'une autre liste d'items). Cependant, employer des formes parallèles afin d'évaluer la fidélité des scores fournis par un outil n'est pas forcément recommandé. En effet, une méta-analyse a démontré que la fidélité calculée au moyen de formes parallèles s'avère légèrement inférieure à la fidélité test-retest calculée à partir de l'administration d'une forme unique d'un même test (Calamia et al., 2013). Ainsi, dans le cadre du calcul de la fidélité test-retest d'un outil d'évaluation, nous estimons que l'emploi d'une forme parallèle n'est pas idéal, d'autant plus qu'il est essentiel d'assurer au préalable l'équivalence des deux formes du test à travers une étude de validation (Benedict & Zgaljardic, 1998).

2) Caractéristiques des outils

D'une part, parmi les caractéristiques de l'outil, des études ont démontré que l'effet d'entraînement peut être modéré par le type de construit évalué. Par exemple, les tests évaluant les fonctions mnésiques ou exécutives présentent une moindre fidélité test-retest (Benedict & Zgaljardic, 1998 ; Calamia et al., 2013) comparativement aux outils évaluant la vitesse de traitement (Scharfen et al., 2018). Ceci s'explique par la mémorisation des stratégies à employer pour réaliser l'épreuve, améliorant la seconde performance. Cet effet a été démontré pour des périodes allant jusqu'à 12 mois d'intervalle (Basso et al., 1999) et est d'autant plus important pour les outils qui impliquent de découvrir une stratégie unique de résolution de problème (par exemple, le Wisconsin Card Sorting Test). L'intérêt de ce type d'épreuve est justement la découverte et l'apprentissage de la stratégie à employer. Dès lors, une fois qu'elle est découverte lors de la première passation, celle-ci est mémorisée et directement appliquée lors du retest. Par conséquent, on peut supposer que ce type d'épreuve n'évalue pas tout à fait les

mêmes construits lors des deux passations, en raison d'un important effet d'entraînement, ce qui diminue l'importance de l'indice de fidélité test-retest (Lezak et al., 2004).

D'autre part, le contenu des items peut également avoir un effet modérateur. Cependant, les résultats issus de la littérature sont mitigés. En effet, des auteurs ont mis en évidence que les outils employant des items numériques entraînent un biais test-retest moindre, en comparaison des outils impliquant un contenu verbal ou multiple (Scharfen et al., 2018). Une explication proposée serait que ces items soient moins familiers, et donc plus difficiles à mémoriser, car ils requièrent davantage de ressources cognitives pour un encodage (Villado et al., 2016). Par exemple, les items verbaux activent également de l'information sémantique, ce qui augmente la probabilité de les mémoriser, contrairement aux chiffres, qui sont davantage dénués de sens (Scharfen et al., 2018). D'autres auteurs ont mis en évidence un effet d'entraînement pour les items visuo-spatiaux (Benedict & Zgaljardic, 1998 ; Salthouse & Tucker-Drob, 2008), justifiant également ces résultats par la notion de familiarité aux items : une première exposition augmenterait la familiarité, entraînant une amélioration de la performance du sujet lors de la seconde passation. En outre, d'autres auteurs encore ont mis en évidence que les épreuves employant plusieurs types d'items entraînent un gain de score plus large que les épreuves n'employant qu'un seul type d'item (Villado et al., 2016). En résumé, il n'existe pas de consensus sur l'effet modérateur des différents types d'items présentés.

3) Caractéristiques des sujets

L'âge des sujets de l'échantillon peut influencer l'importance de l'effet d'entraînement. En effet, chez les enfants, la maturation cérébrale est toujours en cours, et les capacités cognitives sont sujettes à évolution au fil des mois et années. Ainsi, une étude a par exemple démontré que la fidélité des mesures d'inhibition est faible chez les enfants d'âge préscolaire, mettant en évidence que la stabilité temporelle de ce type de mesure mérite d'être davantage investiguée à cette période (Müller et al., 2012). De même, il est possible d'observer plus de fluctuations dans la performance des enfants d'une séance à l'autre. En ce qui concerne les adultes, les résultats sont mitigés : deux méta-analyses (Calamia et al., 2012 ; Calamia et al., 2013) ont démontré un effet de l'âge sur la fidélité test-retest, mais cet effet variait selon les tests, tandis qu'une autre (Scharfen et al., 2018) n'a mis en évidence aucun effet.

De même, une fonction cognitive sera plus ou moins stable, et l'effet d'entraînement s'exprimera différemment selon que le sujet soit issu d'une population saine ou clinique. Cela varie également selon l'outil et le type de population clinique : on pourra observer une fidélité supérieure, inférieure ou égale au groupe contrôle (Calamia et al., 2013 ; Calamia et al., 2012). En effet, certaines populations cliniques présentent des troubles cognitifs acquis, évolutifs et parfois même fluctuants. Par exemple, un patient dément pourrait présenter une dégradation de performance à un test de mémoire en quelques mois. À l'inverse, une personne qui vient de subir un accident vasculaire cérébral pourrait présenter une amélioration de ses performances lors du retest à des épreuves attentionnelles en raison d'une récupération spontanée. Il est donc essentiel de prêter attention au type de population sur lequel la mesure de la fidélité test-retest s'est effectuée. Nous estimons que l'idéal est d'utiliser une population contrôle, afin d'éviter les fluctuations liées aux diverses conditions cliniques.

2. Données psychométriques selon l'objectif visé par l'évaluation

La prise en compte des qualités psychométriques des outils aidera le clinicien à sélectionner le meilleur outil selon son objectif, et lui permettra de nuancer les résultats de son patient. Ainsi, le neuropsychologue qui souhaite s'inscrire dans une démarche EBA et veut veiller aux qualités psychométriques de ses outils lors du choix de ceux-ci pourrait se retrouver perdu face à la quantité d'informations parfois disponibles. Pour parvenir à mener cette entreprise, des auteurs (Youngstrom et al., 2017) suggèrent de prêter attention à certaines données psychométriques spécifiques selon l'objectif visé par le clinicien lors de l'évaluation. Ils distinguent pour ce faire trois catégories qui regroupent la majorité des objectifs pouvant être recherchés par le psychologue (Youngstrom et al., 2017) : (1) *prediction*, (2) *prescription*, (3) *process*. La première réfère à la tentative de relier le profil du patient à un critère important (par exemple, un diagnostic). La deuxième correspond à la volonté de récolter des informations qui permettront d'influencer le choix du traitement. Enfin, la troisième correspond à l'obtention d'informations concernant les progrès du patient lors du traitement afin de quantifier des résultats.

Au départ de ces trois catégories d'objectif, Youngstrom et al. (2017) ont déterminé, pour chacune d'entre elles, quelles données psychométriques s'avèrent les plus pertinentes. Si la démarche évaluative du clinicien se situe dans la catégorie *prediction*, il devra s'assurer de la validité discriminante des mesures fournies par l'outil, qui assure sa capacité à classifier les

trajectoires. Ensuite, si sa démarche se situe plutôt dans la catégorie *prescription,* les informations centrales auxquelles le clinicien devra veiller sont *:* (1) la fidélité inter-juges, essentielle pour assigner un diagnostic et sélectionner un traitement adéquat pour le patient, (2) la validité de construit, étant donné que l'objectif est d'établir un diagnostic supposé être fondé empiriquement, ainsi que (3) la validité prescriptive, qui est la raison même pour laquelle le diagnostic est posé. Enfin, pour la catégorie *process*¸ seront essentielles une bonne validité de construit (qui assurera la sensibilité au traitement et la validité prédictive de l'outil) ainsi qu'une bonne sensibilité au traitement qu'il faudra quantifier de différentes façons tout au long de la prise en charge. Il est important d'ajouter que d'autres informations psychométriques peuvent être également importantes et pertinentes à prendre en compte, mais le sont cependant moins que celles précédemment citées (voir Youngstrom et al., 2017 pour plus de détails).

3. Qualités psychométriques des outils et pourcentage d'utilisation

Sur base de ses connaissances théoriques, le neuropsychologue clinicien est mieux armé dans le choix de ses outils et sera ensuite davantage capable de nuancer les résultats de son patient. Ainsi, pour mettre en place une pratique EBA, les outils privilégiés par les neuropsychologues cliniciens devraient être des outils à la fois valides, fidèles, adéquatement normés, et, dans l'idéal, validés sur base d'une population francophone.

Récemment, un sondage proposé à 804 neuropsychologues français s'est chargé de répertorier le pourcentage d'utilisation d'une liste de 59 outils d'évaluation (Branco Lopes et al., 2019). Les résultats issus de cette étude se trouvent en Tableau 1. En parallèle, une étude menée chez des logopèdes s'est penchée sur les différents facteurs influençant la sélection de tests dans le cadre du diagnostic de troubles spécifiques du langage (Betz et al., 2013). Cette étude visait à déterminer si la qualité des tests standardisés était liée à leur fréquence d'utilisation. Il ressort de cette étude que les propriétés psychométriques des tests, évaluées au travers de la validité concourante, fidélité et données diagnostiques, n'étaient pas corrélées avec leur fréquence d'utilisation par les logopèdes. En réalité, seul un échantillon de tests s'avérait être employé de façon régulière, sans lien significatif avec leur qualité. D'ailleurs, le facteur qui expliquait au mieux la fréquence d'utilisation était la composante historique de l'outil : il semblerait que les logopèdes avaient davantage tendance à employer des tests déjà bien connus, et ayant, dans certains cas, déjà fait l'objet de plusieurs révisions, sans se préoccuper de leurs propriétés psychométriques. En effet, la popularité d'un outil peut mener les professionnels à entretenir de fausses croyances selon lesquelles cet outil est valide et fidèle (Maltzman, 2013). À notre connaissance, aucune étude à ce jour n'a cherché à déterminer si les neuropsychologues francophones choisissent leurs outils d'évaluation en veillant à leurs qualités psychométriques.

Tableau 1. *Pourcentage d'utilisation d'outils d'évaluation parmi 804 neuropsychologues français sondés (Branco Lopes et al.. 2019).*

	Nom de l'outil	%		Nom de l'outil	%
1	Test de Stroop	76.5	31	Symbol Digit Modalities Test (SDMT)	11.1
2	Trail Making Test (TMT)	76.4	32	Rivermead Behavioral Memory Test (RBMT)	9.5
3	Rappel libre/Rappel indicé à 16 items (RL/RI16)	75	33	Batterie courte d'évaluation des fonctions cognitives destinées aux patients souffrant de sclérose en plaques (BCcogSEP)	9.1
4	Échelle d'intelligence de Wechsler pour adultes 4ème édition (WAIS-IV)	64.3	34	Batterie Rapide d'Évaluation des Praxies (BREP)	9
5	Test d'attention concentrée révisé (d2-r)	64.2	35	Hooper Visual Organization Test (HVOT)	8.8
6	Batterie Rapide d'Évaluation Frontale (BREF)	61.8	36	Judgment of Line Orientation (JLO)	8.8
7	Test de dénomination 80 images (DO80)	57	37	Boston Naming Test (BNT)	8.7
8	Visual Object and Space Perception battery (VOSP)	50	38	Buschke Selective Reminding Test (BSRT)	6.1
9	Wisconsin Card Sorting Test (WCST)	46	39	Boston Diagnostic Aphasia Examination (BDAE)	5.5
10	Échelle clinique de mémoire de Wechsler 4ème édition (MEM-IV)	46	40	Finger Tapping est (FTT)	4.5
11	California Verbal Learning Test (CVLT)	45.5	41	Continuous Performance Task (CPT)	3.9
12	Delayed Matching to Sample 48 items (DMS48)	45.5	42	Évaluation de la démence fronto-temporale (SEA)	3.2
13	Mini Mental State Examination (MMSE)	44.7	43	Hopkins Verbal Learning Test (HVLT)	3.1
14	Test des portes	38.2	44	Rey Auditoral Verbal Learning Test (RAVLT)	3
15	Test de l'horloge	34	45	Controlled Oral Word Association Test (COWAT)	2.7
16	Échelle d'intelligence de Wechsler pour enfants et adolescents 5ème édition (WISC-V)	31.8	46	Repeatable Battery for the Assessment of Neuropsychological Status (RBANS)	2.2
17	Test d'Évaluation de l'Attention (TEA / TAP)	31.5	47	Visual Motor Integration test (VMI)	1.2
18	Montreal Cognitive Assessment (MoCA)	30.3	48	Warrington Recognition Memory Test (WRMT)	1.2
19	Bilan neuropsychologique de l'enfant 2ème édition (NEPSY-II)	29.5	49	Test de Barcelone	1
20	Test de Hayling	29.1	50	Neurobehavioral Cognitive Status Examination (NCSE / Cognistat)	1
21	Batterie d'Évaluation des Connaissances Sémantiques du GRECO (BECS-GRECO)	28.1	51	Luria Nebraska Neuropsychological Battery (LNNB)	1
22	Échelle de démence de Mattis (MDRS)	25.9	52	Memory Assessment Scale (MAS)	1
23	Paced Auditory Serial Addition Test (PASAT)	24.8	53	Échelle d'intelligence de Wechsler pour adultes révisée (WAIS-R)	1
24	Échelle d'intelligence de Wechsler pour enfants 4ème édition (WPPSI-IV)	22.8	54	Test of Memory Malingering (TOMM)	0.7
25	Évaluation de la démence fronto-temporale (Mini-SEA)	19.9	55	Consortium to Establish a Registry for Alzheimer's Disease (CERAD)	0.6
26	Neuropsychiatric Inventory (NPI)	19.5	56	Halstead Reitan Neuropsychological Battery (HRNB)	0.5
27	Delis-Kaplan Executive Function System (D-KEFS)	18.9	57	Bender Visual-Motor Gestalt Test (BVMGT)	0.2
28	Behavioural Assessment of the Dysexecutive Syndrome (BADS)	14.2	58	Vigil Continuous Performance Test	0.1
29	Figure complexe de Rey	13.4	59	Wide Range Achievement Test (WRAT)	0.00
30	Token Test	11.4			

Note. L'abréviation du nom de l'outil est indiquée entre parenthèses. % = Pourcentage d'utilisation de l'outil.

II. Objectifs et hypothèses

L'adoption d'une pratique de type *evidence-based* est devenue une réelle nécessité dans le domaine de la neuropsychologie clinique, aussi bien dans le contexte de la prise en charge que celui de l'évaluation. Nous avons vu dans notre partie théorique que, pour mener une démarche EBA, le clinicien doit se baser : (1) sur le cas spécifique de son patient, (2) sur son expertise clinique ainsi que (3) sur les données issues de la littérature scientifique l'informant des qualités psychométriques de l'outil. Pour qu'un outil d'évaluation soit valide dans notre culture franco-européenne, il devrait, si possible, avoir fait l'objet d'une étude de validation sur base d'une population francophone. Ainsi, notre recherche a pour objectif principal de faire le point sur les qualités psychométriques des outils d'évaluation employés dans le cadre de la neuropsychologie francophone européenne. Pour ce faire, notre travail se composera de plusieurs étapes.

Premièrement, nous analyserons les outils recensés dans l'étude de Branco Lopes et al. (2019), qui répertorie 59 outils d'évaluation les plus fréquemment utilisés par les neuropsychologues français (voir Tableau 1). Nous tenterons de répertorier toute information psychométrique concernant ceux-ci. L'objectif de cette première étape sera de déterminer si les outils que nous employons ont fait ou non l'objet d'études de validation et, si oui, si cette validation a été réalisée sur base d'une population francophone.

À partir des données récoltées, notre deuxième étape visera à déterminer si les neuropsychologues choisissent leurs outils d'évaluation en fonction de leurs données psychométriques. Pour ce faire, nous observerons si le pourcentage d'utilisation des outils (rapporté par Branco Lopes et al., 2019) est positivement corrélée à leurs qualités psychométriques. Si c'est le cas, cela signifierait que les praticiens choisissent un outil d'évaluation en regard de ses informations psychométriques, ce qui serait conforme à une pratique EBA. Au vu des résultats d'une étude portant sur des logopèdes (Betz et al., 2013), notre hypothèse est plutôt la suivante : nous ne devrions pas observer de corrélation significative entre le pourcentage d'utilisation des outils et ses qualités psychométriques. En revanche, si cette hypothèse s'avère erronée, et que le pourcentage d'utilisation des outils est corrélé avec une ou plusieurs de ses qualités psychométriques, nous pourrons déterminer laquelle semble le plus influencer le choix des neuropsychologues.

Enfin, dans une dernière étape, nous évaluerons la fidélité de nos outils au moyen d'une analyse de type méta-analyse. Celle-ci se basera sur les indices de fidélité test-retest des outils et visera à évaluer la stabilité des mesures qu'ils fournissent. Dans cette analyse, nous tiendrons compte de variables relatives à (1) la méthodologie employée (par exemple, le délai test-retest), (2) aux caractéristiques propres à l'épreuve (par exemple, le construit évalué par l'outil), (3) ainsi qu'aux caractéristiques propres à l'échantillon constitué pour l'étude (par exemple, l'âge de l'échantillon).

III. Méthodologie

A. Sélection des études

Ce travail porte sur les études de validation d'outils d'évaluation employés par les neuropsychologues français. Les outils qui feront l'objet de notre étude ont été sélectionnés sur base de l'étude de Branco Lopes et al. (2019) qui répertorie 59 épreuves et batteries, ainsi que leur pourcentage d'utilisation pour 804 neuropsychologues français sondés (voir Tableau 1). Dans le cas où un outil mentionné aurait fait l'objet de plusieurs révisions, nous avons fait le choix de ne sélectionner que les études qui concernent leur dernière version (par exemple, pour l'Échelle d'Intelligence pour Adultes de Wechsler, nous nous sommes limités à la 4ème et dernière édition de la batterie).

B. Stratégie de recherche des études de validation

La recherche de données s'est opérée en plusieurs étapes. Dans un premier temps, nous avons eu la chance de disposer d'une base de données fournie par l'Organisation Française des Psychologues spécialisés en Neuropsychologie (OFPN)[2]. Cette base de données a été créée par un panel d'experts issus des différents champs de la neuropsychologie dans l'objectif de répertorier, le plus exhaustivement possible, une large gamme d'outils employés par les neuropsychologues francophones ainsi que les références des études de validation associées à ces outils.

Afin d'aller plus loin dans notre recherche, nous avons mis en place une étape supplémentaire en utilisant les moteurs de recherche Pubmed et PsycINFO. Sur PsycINFO par exemple, nous avons mis en place une recherche au moyen des mots-clés (MeSH terms) suivants : (*test reliability* [Mesh] OR *test validity* [Mesh] OR *test specificity* [Mesh] OR *test sensitivity* [Mesh] OR *test standardization* [Mesh]) AND *neuropsycholog**. Ces mots-clés ont été associés au nom et à l'abréviation (anglophone et francophone) de l'outil d'évaluation qui faisait l'objet de notre recherche, par exemple : AND (Rappel Libre Rappel Indice 16 items OR RL/RI16 OR Free and Cued Selective Reminding Test OR FCRST).

[2] L'OFPN vise à proposer une organisation rassemblant les psychologues spécialisés en neuropsychologie, afin de servir de porte-parole et promouvoir la pratique de la neuropsychologie clinique (https://ofpn.fr/).

C. Critères d'inclusion

Plusieurs critères d'inclusion ont été utilisés. Nous avons d'abord sélectionné toutes les études de validation fournissant des informations psychométriques sur un outil spécifique, en nous limitant aux études rédigées en français ou anglais. De plus, nous avons limité nos recherches aux travaux publiés à partir de janvier 1990. Nous avons fait ce choix en prenant en compte l'effet Flynn (Dickinson & Hiscock, 2011). Il n'existe pas de recommandation optimale concernant la « date limite » de validité d'une étude. De ce fait, l'année 1990 a été choisie de façon arbitraire car nous estimons que les résultats publiés auparavant risquent d'être aujourd'hui désuets.

Pour répondre à notre objectif visant à déterminer si les outils que les neuropsychologues emploient ont fait ou non l'objet d'études de validation, nous avons en priorité privilégié l'inclusion d'études de validation basées sur un échantillon issu d'une population franco-européenne (France, Belgique francophone et Suisse francophone). À défaut, nous avons répertorié des études issues d'autres populations. Dans la plupart des cas, plusieurs études remplissaient nos premiers critères d'inclusion. Nous avons alors établi cinq critères supplémentaires permettant de déterminer au mieux quelle étude serait de meilleure qualité pour les besoins spécifiques de notre travail. Les critères de sélection établis sont les suivants, énoncés par ordre d'importance : (1) échantillon issu d'une population franco-européenne, (2) échantillon se rapprochant le plus d'une population franco-européenne (par exemple, nous privilégierons d'abord les populations francophones, ensuite, si nous devons faire un choix entre une étude réalisée sur des participants américains, ou des participants espagnols, nous considérerons que les sujets espagnols partagent une culture plus proche de la culture franco-européenne), (3) l'emploi d'une population tout-venant (ce critère ne concerne pas les études évaluant la validité discriminante), (4) la plus grande taille d'échantillon (*N*), (5) l'année de publication la plus récente.

En ce qui concerne les études normatives et diagnostiques (validité discriminante), il arrive que plusieurs études de validation fournissent des informations complémentaires. En effet, deux cas de figure sont survenus. Dans un premier cas, il arrive qu'une étude normative propose des normes pour des sujets de 16 à 50 ans, alors qu'une autre concerne des sujets de 60 à 80 ans. Dans cette situation, les deux études ont été répertoriées à titre informatif. Dans un second cas, il arrive également qu'une étude de validité discriminante présente des données diagnostiques pour un type de population clinique (par exemple, des sujets Alzheimer) et qu'une autre en propose pour une autre population (par exemple, des patients ayant subi un trauma crânien). Ces données étant complémentaires, plusieurs études ont été répertoriées.

D. Extraction et encodage des données

Deux tableaux d'extraction ont été créés afin de faciliter l'analyse des données (voir Tableau 2 et 3). Une seule personne s'est chargée de l'encodage. En cas de doute concernant la façon d'encoder ou interpréter une donnée, nous avons consulté les promoteurs de ce mémoire afin d'avoir un avis complémentaire. Afin de déterminer quelles données psychométriques étaient disponibles pour chaque outil, et d'évaluer les variables qui influencent les neuropsychologues à utiliser ceux-ci, nous avons établi un premier tableau (voir résumé en Tableau 2). Celui-ci visait à répertorier, pour chaque étude sélectionnée, les informations suivantes :

Informations générales. Nous avons encodé le *nom de l'outil* ainsi que le *pourcentage d'utilisation* pour chaque outil comme rapporté dans l'étude de Branco Lopes et al. (2019). Nous avons également encodé le *construit « global »* évalué par la batterie ou l'épreuve (c.-à-d. fonctions mnésiques, fonctions exécutives, fonctionnement cognitif global, fonctions attentionnelles, niveau intellectuel, fonctions praxiques, fonctions visuo-spatiales, fonctions langagières). Si un outil comprend plusieurs *épreuves* ou *sous-tâches*, celles-ci ont également été encodées individuellement. Pour chaque épreuve et sous-tâche, nous avons encodé *le sous-construit évalué*, (c.-à-d. vitesse de traitement, inhibition, flexibilité, mise à jour, planification, raisonnement, mémoire sémantique, mémoire à long terme, mémoire à court terme et de travail – les deux construits ont été rassemblés car, dans la plupart des cas, les épreuves évaluent les deux construits, par exemple la Mémoire des chiffres de la WAIS-IV —, visuo-spatial, praxies, langage, attention sélective, alerte, attention divisée, attention soutenue, cognition sociale). Enfin, nous avons également indiqué l'*année de publication de la première version* de l'outil, pour évaluer si cette donnée historique exerce une influence sur le choix des outils.

Standardisation. Afin de décrire la qualité des études normatives sélectionnées pour chaque outil, nous avons précisé, pour chacune, la *taille de l'échantillon normatif* (*N*), l'*âge minimal et maximal* des sujets, ainsi que les différents *sous-groupes* proposés (par exemple, par niveau scolaire, genre, etc.). Nous avons également encodé le *pays* d'origine des sujets, afin d'évaluer si les neuropsychologues choisissent leurs outils selon l'origine de l'échantillon normatif. L'*année de publication* des normes a également été encodé afin de déterminer si les neuropsychologues choisissent leur outil en prenant en compte la récence des normes.

Validité. Afin de déterminer quelle méthode est la plus régulièrement utilisée pour démontrer la validité d'un outil, nous avons encodé pour chaque étude sélectionnée le *type de validité* (c.-à-d. validité de construit, convergente, concourante, et divergente – la validité discriminante sera analysée différemment), ainsi que la *méthode* (*c.-à-d.* corrélation ou analyse factorielle) et/ou l'*outil de comparaison* employé (à titre informatif). De plus, afin d'estimer si les professionnels prennent en compte la validité des mesures dans la sélection de leurs outils, nous avons encodé les valeurs de *validité concourante* sous forme de corrélation. Si l'étude de validité concourante préalablement sélectionnée fournit des corrélations avec plusieurs autres mesures, nous avons sélectionné celle qui mesure le construit se rapprochant le plus de celui mesuré par l'épreuve que nous voulons analyser. Par exemple, si un test de mémoire verbale est comparé avec une épreuve de mémoire visuelle ainsi qu'une autre épreuve de mémoire verbale, nous avons encodé cette dernière pour nos analyses futures.

Validité discriminante. Afin d'évaluer l'existence de données diagnostiques, nous avons encodé pour chaque étude incluse la valeur du *N* pour le groupe contrôle ainsi que pour la population clinique concernée. Ensuite, la précision diagnostique de l'outil a été déterminée en utilisant les valeurs de *sensibilité* et *spécificité* (encodées en valeurs décimales). Si l'étude fournissait des valeurs diagnostiques pour différents scores seuils, nous avons sélectionné celui recommandé par les auteurs, ou, à défaut, celui qui maximise la sensibilité et spécificité de détection. Le Rey Auditory Verbal Learning Test posait un cas particulier, car nous avons sélectionné une étude qui rapportait des valeurs diagnostiques pour trois groupes d'âge (Messinis et al., 2016). Pour cette étude, nous avons calculé la moyenne pondérée des données pour les trois groupes d'âge, afin d'obtenir une valeur unique. Enfin, l'*année* et le *pays* de publication ont également été encodés afin de déterminer si les neuropsychologues choisissent leur outil en prenant en compte la récence et/ou l'origine des normes diagnostiques.

Fidélité inter-juges. Afin d'évaluer si les mesures fournies par l'outil ont fait l'objet d'une étude de fidélité en utilisant une méthode inter-juges, nous avons encodé la *valeur* minimale et maximale rapportées pour les différents subtests ou sous-tâches, l'*indice* utilisé (par exemple, corrélation de Pearson, ou corrélation intra-classe), ainsi que la *référence* de l'étude.

Tableau 2. *Extraction des données en vue de la revue des qualités psychométriques.*

	Revue des études de validation
Informations générales	
Nom de l'outil	
Année de publication d'origine	
Pourcentage d'utilisation	Rapporté par Branco Lopes et al. (2019) (voir Tableau 1)
Construit global évalué	Fonctionnement cognitif global, fonctions attentionnelles, fonctions exécutives, fonctions mnésiques, fonctions langagières, fonctions visuo-spatiales, fonctions praxiques, cognition sociale
Subtests / Sous-tâches	
Construit évalué par le subtest / la sous-tâche	Vitesse de traitement, inhibition, flexibilité, mise à jour, planification, raisonnement, mémoire sémantique, mémoire à long terme, mémoire à court terme et de travail, visuo-spatial, praxies, langage, attention sélective, alerte, attention divisée, attention soutenue, cognition sociale
Standardisation	
N échantillon normatif	
Année de publication	
Nationalité	
Tranche d'âge des sujets	
Sous-groupe	Par exemple, niveau scolaire, genre, etc.
Présentation des normes	Par exemple, percentile, moyenne et écart-type, etc.
Validité	
De construit	Méthode employée (par exemple, analyse factorielle)
Convergente	Outil(s) de comparaison
Concourante	- Outil(s) de comparaison - Valeurs des corrélations pour chaque subtest ou sous-tâches
Divergente	Outils de comparaison
Discriminante	- Type de population clinique (par exemple, Alzheimer) - *N* contrôle - *N* clinique - Sensibilité - Spécificité
Fidélité inter-juges	
Indice	Par exemple, corrélation *r* de Pearson, ou corrélation intra-classe (ICC)
Valeur minimale et maximale	

Note. Nous avons également précisé, pour chacune des études de validation répertoriées, la référence ainsi que le pays de publication.

Le second tableau (voir résumé en Tableau 3) a été constitué pour mener notre méta-analyse sur les indices de fidélité test-retest. Pour ce faire, nous avons encodé les tailles d'effet ainsi que différentes variables potentiellement modératrices :

Informations générales. Nous avons encodé le *nom de l'outil*, les différents *subtests* et *sous-tâches* de celui-ci, l'*année de publication* (et référence) de l'étude sélectionnée, et le *pays* (qui correspond à la nationalité de l'échantillon).

Méthodologie test-retest. Nous avons encodé trois variables modératrices liées à la méthodologie de l'étude. Premièrement, nous avons encodé la *durée de l'intervalle test-retest*, exprimée en jours (les durées exprimées en intervalle ont été transformées en nombre de jours en utilisant la valeur centrale : par exemple, un intervalle de 14 à 84 jours a été codé *49*), car nous savons que cet intervalle a un impact important sur la mesure de la fidélité. Deuxièmement, nous avons également encodé la *taille de l'échantillon* (*N*) qui influence la précision de la mesure fournie. Troisièmement, nous avons précisé quelles *versions de l'outil* ont été employées lors des deux administrations (c.-à-d. identiques, parallèles ou mixtes).

Caractéristiques de l'outil. Nous avons encodé quatre variables modératrices liées aux caractéristiques propres à l'outil d'évaluation. Premièrement, le *type d'outil* (c.-à-d. *batterie*, si l'outil provient d'une batterie, ou *épreuve*) afin d'évaluer si cette variable exerce une influence sur la fidélité de la mesure. Deuxièmement, le *construit global évalué* (c.-à-d. fonctionnement cognitif global, fonctions attentionnelles, fonctions exécutives, fonctions mnésiques, fonctions langagières, fonctions visuo-spatiales, fonctions praxiques, cognition sociale) ainsi que le *sous-construit* évalué (dans le cas de substests ou sous-tâches) (c.-à-d. vitesse de traitement, inhibition, flexibilité, mise à jour, planification, raisonnement, mémoire sémantique, mémoire à long terme, mémoire à court terme et de travail, visuo-spatial, praxies, langage, attention sélective, alerte, attention divisée, attention soutenue, cognition sociale). En effet, nous savons que le construit évalué peut impacter l'effet d'entraînement et donc la stabilité de la mesure fournie par l'outil. Troisièmement, nous avons encodé le *contenu des items* (c.-à-d. *verbal* lorsque l'outil est présenté de façon orale uniquement, *visuel* lorsque les items sont constitués de formes ou d'images, *numérique* lorsque des chiffres sont présentés oralement ou visuellement, *multiple* lorsque l'outil propose plusieurs types de contenu, ou *autre* si aucune de ces catégories ne convient) qui peut également avoir un effet potentiel. Enfin, nous avons encodé la *consistance interne* de l'outil (valeur chiffrée), ainsi que la *méthode employée* pour évaluer celle-ci (par exemple : alpha de Cronbach α) afin d'évaluer si cette mesure a un impact sur la stabilité des mesures fournies par l'outil.

Caractéristiques de la population. Nous avons encodé trois variables liées aux sujets de l'étude. Premièrement, le *type de population* (c.-à-d. groupe *contrôle*, *clinique*, ou *multiple*) car cette variable peut impacter la stabilité de la mesure. Deuxièmement, nous avons encodé la *tranche d'âge de l'échantillon* (c.-à-d. *enfant* pour les moins de 18 ans, *adulte* à partir de 18 ans jusque 69 ans, et *personne âgée* dès 70 ans, ou *mixte* pour les échantillons regroupant plusieurs de ces catégories). Enfin, troisièmement, nous avons encodé la *moyenne d'âge,* exprimée en années, car l'âge des sujets peut également influencer la mesure de la fidélité test-retest en impactant l'importance de l'effet d'entraînement.

Fidélité test-retest. Étant donné que la fidélité test-retest peut être exprimée de différentes façons, nous avons encodé le *type d'indice* (*c.-à-d.* corrélation de Pearson *r*, *Z* de Fisher*)*, ainsi que sa *valeur* chiffrée en prenant en compte deux décimales.

Tableau 3. *Extraction des données en vue de la méta-analyse des indices de fidélité test-retest.*

Analyse de la fidélité test-retest	
Informations générales	
Nom de l'outil	
Subtests / Sous-tâches	
Pourcentage d'utilisation	Rapporté par Branco Lopes et al., 2019 (voir Tableau 1)
Année de publication	
Pays de publication	
Référence de publication	
Méthodologie test-retest	
Intervalle test-retest	Exprimé en jours (valeur centrale)
N	
Versions de l'outil	Identiques, parallèles, mixtes
Caractéristiques de l'outil	
Type d'outil	Batterie, épreuve
Construit global évalué	Fonctionnement cognitif global, fonctions attentionnelles, fonctions exécutives, fonctions mnésiques, fonctions langagières, fonctions visuo-spatiales, fonctions praxiques, cognition sociale
Sous-construit évalué	Vitesse de traitement, inhibition, flexibilité, mise à jour, planification, raisonnement, mémoire sémantique, mémoire à long terme, mémoire à court terme et de travail, visuo-spatial, praxies, langage, attention sélective, alerte, attention divisée, attention soutenue, cognition sociale
Contenu des items	Verbal, visuel, numérique, multiple, autre
Consistance interne et indice	Valeur chiffrée et, par exemple, alpha de Cronbach
Caractéristiques des sujets	
Type d'échantillon	Contrôle, clinique, mixte
Âge d'échantillon	Moyenne d'âge
Tranche d'âge de l'échantillon	Enfants, adultes, personnes âgées, mixte
Fidélité test-retest	
Type d'indice	*r* de Pearson ou *Z* de Fisher
Valeur de l'indice	Valeur chiffrée

E. Analyse des données

1. Analyses descriptives

D'abord, le Tableau 2 nous a permis de répondre à notre premier objectif visant à répertorier les études de validation pour différents outils d'évaluation employés par les neuropsychologues français. Les données répertoriées nous ont permis de déterminer si les outils les plus fréquemment utilisés sont aussi ceux ayant fait l'objet d'une étude de validation en langue française. Nous avons également pu déterminer, par exemple, parmi l'ensemble des études, quel type de validité est régulièrement proposé par les auteurs (de construit, convergente, concourante, discriminante, ou divergente), combien d'outils ont été validés ou normés en français ou encore, le nombre d'outils pour lequel se rapportent des données diagnostiques (validité discriminante), etc.

Ensuite, afin de déterminer si les neuropsychologues choisissent leurs outils d'évaluation en fonction de leurs qualités psychométriques, nous avons mené une analyse corrélationnelle. Une corrélation de Pearson a été menée sur les variables continues, tandis qu'une corrélation bisériale a été employée en cas de variable dichotomique. Pour rappel, nous disposions des pourcentages d'utilisation pour chaque épreuve et batterie considérées dans leur globalité, nous avons donc dû nous assurer que chaque pourcentage n'était corrélé qu'avec une seule donnée. Nous avons procédé de la sorte pour les différentes données psychométriques :

Standardisation. Lorsque plusieurs études ont été sélectionnées, nous avons choisi de corréler le pourcentage d'utilisation avec l'étude fournissant le plus grand *N*, son année de publication ainsi que la nationalité de l'échantillon (codé 1 pour un échantillon franco-européen, et 0 dans les autres cas). Enfin, nous avons corrélé le pourcentage d'utilisation avec la disponibilité de l'information (codée 1) ou non (codée 0).

Validité discriminante. Dans le cas où plusieurs études portant sur différentes populations cliniques ont été sélectionnées, nous avons utilisé, dans la mesure du possible, les données issues de la population prévue initialement par les créateurs de l'outil. Par exemple, pour le Mini Mental State Examination, nous disposions de données concernant des patients ayant subi un Accident Vasculaire Cérébral (AVC), des patients atteints cognitivement (sans davantage de détails), des patients Parkinson, ou encore des patients Alzheimer. Nous avons donc encodé les données diagnostiques correspondant aux patients Alzheimer. Ensuite, deux cas de figure se sont présentés : (1) dans le cas des épreuves, il arrive que les auteurs fournissent la sensibilité

et spécificité de plusieurs mesures. Nous avons alors choisi de corréler celle qui était recommandée par les auteurs ou, à défaut, celle qui maximisait la sensibilité et spécificité de l'outil (par exemple, pour le Trail Making Test, la partie B de l'épreuve est plus efficace que la partie A pour détecter les patients présentant une démence). (2) Dans le cas des batteries, nous avons moyenné les indices de sensibilité et spécificité des différentes épreuves afin de n'avoir plus qu'une donnée unique. Pour finir, nous avons également corrélé l'année de publication de l'étude utilisée pour la corrélation ainsi que la nationalité de l'échantillon (codé 1 pour un échantillon franco-européen, et 0 dans les autres cas). Enfin, nous avons corrélé le pourcentage d'utilisation avec la disponibilité de l'information (codée 1) ou non (codée 0).

Validité concourante et fidélité test-retest. La même procédure a été appliquée pour ces deux données : lorsque plusieurs corrélations exprimant une validité concourante ou une fidélité test-retest ont été répertoriées pour un même outil, nous avons moyenné ces corrélations (en veillant à éliminer les signes négatifs, étant donné que nous sommes intéressés par la force de la corrélation et non le sens de celle-ci). De plus, nous avons corrélé l'année de publication de l'étude ainsi que la nationalité de l'échantillon (codé 1 pour un échantillon franco-européen, et 0 dans les autres cas). Enfin, nous avons corrélé le pourcentage d'utilisation avec la disponibilité (codée 1) ou non (codée 0) de l'information.

Cette analyse corrélationnelle nous a permis de déterminer si les différentes données psychométriques répertoriées (c.à.d. année d'origine de l'outil, *N* de l'échantillon normatif, sensibilité, spécificité, validité concourante, fidélité test-retest, année et pays de publication de l'étude et disponibilité de l'information) étaient significativement corrélées au pourcentage d'utilisation des outils (rapportée par Branco Lopes et al., 2019).

2. Analyse de la fidélité

Pour mener notre méta-analyse, nous avons suivi les recommandations émises par Beretvas et Pastor (2003) concernant les méta-analyses portant sur des indices de fidélité. Nous utiliserons le programme *R Studio* pour nos analyses.

D'une part, certaines études de fidélité test-retest distinguent plusieurs sous-groupes (selon plusieurs groupes d'âge), et fournissent alors plusieurs tailles d'effet. Lorsque c'était le cas, nous ne pouvions pas assurer l'indépendance des différents sous-groupes car certaines similarités inter-groupes peuvent entraîner une dépendance entre les corrélations obtenues (Beretvas & Pastor, 2003). Afin d'éviter ce problème de dépendance des tailles d'effet, nous avons calculé la moyenne de ces corrélations, pondérée par la taille des différents échantillons, en passant par un *z* Fisher avant de revenir à un *r* de Pearson. Cela nous a permis de regrouper les différents sous-groupes en un seul échantillon auquel ne se rapporterait plus qu'une seule taille d'effet.

D'autre part, des études visant à évaluer la fidélité d'un outil proposent différentes tailles d'effet pour chacune des mesures fournies par l'outil. Lorsque ces mesures évaluent le même construit, nous avons moyenné les tailles d'effet afin de limiter l'indépendance des tailles d'effet (c'est le cas par exemple du California Verbal Learning Test, qui propose plusieurs scores évaluant la mémoire à long terme). D'autres outils en revanche, comme le test de Stroop, sont constitué de plusieurs sous-tâches évaluant des construits très différents (la vitesse de traitement et l'inhibition). Dans ce cas, nous avons considéré les tailles d'effet comme indépendantes, afin de pouvoir évaluer par la suite l'influence des différents modérateurs que peuvent représenter les construits. En ce qui concerne les batteries, nous avons considéré chaque épreuve de façon indépendante.

Enfin, nous avons utilisé une procédure méta-analytique à effets aléatoires. Ce modèle suppose que les tailles d'effet diffèrent de la moyenne de la population en raison de l'erreur d'échantillonnage et de la variabilité liée à l'étude (Borenstein et al., 2007). Nous avons employé une approche de type « *shifting unit of analysis* » (Cooper, 2017). Ainsi, les tailles d'effet ont été analysées comme indépendantes. De plus, afin d'assurer la puissance statistique de nos analyses, les distributions composées de moins de cinq tailles d'effet (c.-à-d. $k < 5$) n'ont pas été analysées. Par ailleurs, les valeurs numériques (c.-à-d. la durée de l'intervalle test-retest, la consistance interne de l'outil, et l'âge moyen des sujets) seront analysées au moyen d'une corrélation de Pearson avec les indices de fidélité test-retest.

IV. Résultats

A. Analyses descriptives

1. Revues des études de validation

Parmi les 59 outils issus de l'étude de Branco Lopes et al. (2019) (voir Tableau 1), plusieurs outils ont dû être exclus. Nous avons éliminé de cette sélection la batterie d'Intelligence de Wechsler pour Adultes révisée (WAIS-R), car nous avons privilégié la version plus récente de cette batterie (WAIS-IV). Nous avons également éliminé le Vigil Continuous Performance test, car nous n'avons trouvé aucune information à propos de cet outil. Au total, pour l'ensemble des 57 outils restants, nous avons pu répertorier 156 études de validation fournissant diverses informations psychométriques. L'ensemble des études de validation répertoriées est détaillé en annexe (voir Tableau A1 à A5), un résumé du nombre et du pourcentage d'outils validés est proposé en Tableau 4.

Données normatives. Nous avons répertorié des études de standardisation pour 52 outils (91.23%) (Tableau 4, voir Tableau A1 pour plus de détails). Pour certains outils, 5 études supplémentaires ont été sélectionnées à titre informatif car elles apportent des informations complémentaires, ce qui résulte en un total de 57 études de standardisation. Parmi ces 52 outils standardisés, 27 sont normés sur base d'une population francophone, ce qui représente un peu plus de la moitié des outils normés (51.92%). Concernant la présentation des normes, sur les 57 études, 34 proposent des mesures de type percentile, 25 des moyennes par groupe, 7 des équations de régression et enfin, 11 sont sous forme de notes étalonnées ou standards. La présentation en percentiles est donc la plus fréquemment proposée. De plus, 39 études distinguent les différents groupes d'âge par niveau d'éducation ou niveau socioculturel. Dix études proposent également une distinction de sous-groupes par genre.

Validité. Nous avons rapporté 43 études de validité discriminante pour 27 outils différents, soit 47.37% des 57 outils (Tableau 4, voir Tableau A2 pour plus de détails). Plusieurs d'entre eux ont été validés sur différentes populations cliniques, ce qui explique le fait que nous avons répertorié plus d'études que le nombre d'outils validés (par exemple, le MMSE a été validé sur des sujets Alzheimer, AVC, Parkinson, et d'autres présentant une atteinte cognitive). Parmi ces 27 outils, 33.33% l'ont été sur base d'une population française. En ce qui concerne les autres formes de validité, 26 outils ont été validés au moyen d'une analyse factorielle afin d'évaluer leur validité de construit (voir Tableau A3), soit 45.61%. La validité de construit a également

été évaluée au moyen d'autres méthodes : nous avons répertorié des preuves de validité convergente pour 29 outils (50.88%), de validité concourante pour 25 outils (43.86%) et enfin, de validité divergente pour 5 outils seulement (8.77%) (voir Tableau A3 et A4 pour plus de détails). Cela fait de la validité convergente la méthode la plus fréquemment employée pour évaluer la validité de construit des outils. Sur l'ensemble des 128 études de validation répertoriées (toutes méthodes confondues) seules 26 ont été réalisées sur base d'une population française, soit 20.30%. Enfin, il est intéressant de noter que nous n'avons trouvé aucune preuve de validité (toutes méthodes confondues) pour 3 outils (à savoir, le DO80, le BSRT, et la Test de Barcelone), soit 5.26% des outils.

Fidélité. La stabilité des mesures a été évaluée au moyen d'une méthode test-retest pour 33 outils (57.89%) (Tableau 4), soit pour l'ensemble des sous-tâches et subtests issus de l'outil, soit seulement de façon partielle (voir Tableau A5 pour plus de détails). Parmi ces études, seules 5 ont été réalisées sur une population française. En ce qui concerne les mesures de fidélité inter-juges, en revanche, seuls 18 outils (31.58%) ont été investigués, dont 6 en France (Tableau 4, voir Tableau A4 pour plus de détails).

Tableau 4. *Nombre et pourcentage d'outils validés.*

	N outils validés	% outils validés	*N* outils validés en France	% outils validés en France parmi l'ensemble des outils	% outils validés en France parmi les outils validés uniquement
Normes	52	91.23	27	47.37	51.92
Validité					
- Discriminante	27	47.37	9	15.79	33.33
- De construit	26	45.61	2	3.51	7.69
- Concourante	25	43.86	5	8.77	20.00
- Convergente	29	50.88	9	15.79	31.03
- Divergente	5	8.77	1	1.75	20.00
Fidélité					
- Test-retest	33	57.89	5	8.77	15.15
- Inter-juges	18	31.58	6	10.53	33.33

Note : N = nombre d'outils ; % = pourcentage.

2. Caractéristiques psychométriques des outils et fréquence d'utilisation

Des corrélations de Pearson ont été menées entre le pourcentage d'utilisation des 57 outils (Branco Lopes et al., 2019) et les caractéristiques psychométriques suivantes : date historique de l'outil, *N* de l'échantillon normatif, sensibilité, spécificité, validité concourante (sous forme de corrélation de Pearson), fidélité test-retest (exprimée en corrélation de Pearson) ainsi que l'année de publication de chacune de ces données psychométriques (Tableau 5). De plus, pour chacune des informations psychométriques répertoriées, nous avons mené des corrélations bisériales sur les variables dichotomiques suivantes : pays de publication (France ou autre) et disponibilité de l'information. Étant donné que nous n'avons pas trouvé l'ensemble des informations pour chacun des outils, le nombre de données inclus dans chaque corrélation varie. Nous nous sommes intéressés à la corrélation entre le pourcentage d'utilisation et les différentes variables encodées. Parmi toutes ces données, 4 corrélations se sont avérées significatives. Celles-ci concernent la taille de l'échantillon normatif *(N*), le pays de publication des normes, l'existence de données normatives (disponibilité de l'information), ainsi que le pays de publication des normes diagnostiques (ou validité discriminante). La valeur de ces corrélations significatives vaut respectivement $r = .42$, $r = .62$, $r = .32$, et $r = .55$. Aucune des autres données psychométriques analysées ne s'est avérée significativement corrélée avec le pourcentage d'utilisation. Les autres corrélations sont dans l'ensemble faibles, avec un $r \leq .30$ (Cohen, 1988). Mettons également en évidence que le simple fait qu'il existe (ou non) des études de validité discriminante, concourante ou de fidélité test-retest n'est pas non plus corrélé avec le pourcentage d'utilisation. Ensuite, il est intéressant de noter que trois corrélations non significatives se sont avérées négatives. Étonnamment, le pourcentage d'utilisation des outils et la fidélité test-retest, entretiennent une relation négative ($r = — .25$). Ceci peut signifier que les outils les plus fréquemment utilisés sont aussi ceux qui fournissent les mesures les moins fidèles. De façon moins étonnante, on observe une relation négative entre le pourcentage d'utilisation et la date historique de l'outil ($r = — .19$). Enfin, la spécificité entretient une relation négative également avec le pourcentage d'utilisation ($r = — .23$), à l'inverse de la sensibilité ($r = .18$).

Tableau 5. *Corrélation entre les caractéristiques psychométriques des outils et leur pourcentage d'utilisation par les neuropsychologues français, rapporté par Branco Lopes et al. (2019).*

Caractéristiques psychométriques	*r*
Année d'origine de l'outil	— .19 (55)
Normes	
- *N*	**.42*** (48)
- Année de publication	.16 (48)
- Pays	**.62*** (48)
- Disponibilité de l'information	**.32*** (56)
Validité discriminante	
- Sensibilité	.18 (25)
- Spécificité	— .23 (25)
- Année de publication	— .06 (25)
- Pays	**.55*** (25)
- Disponibilité de l'information	.009 (56)
Validité concourante	
- Corrélation	— .08 (24)
- Année de publication	.25 (24)
- Pays	.19 (24)
- Disponibilité de l'information	.23 (56)
Fidélité test-retest	
- Corrélation	— .25 (30)
- Année de publication	— .05 (30)
- Pays	.30 (30)
- Disponibilité de l'information	.23 (56)

Note. Les degrés de liberté sont indiqués entre parenthèses. *r* = corrélation de Pearson pour les variables continues et *r* = corrélation bisériale pour les variables dichotomiques.

*p < .05.

B. Méta-analyse des indices de fidélité test-retest

1. Résultats globaux

Au total, 32 études ont été répertoriées dans cette méta-analyse. Un outil (4 subtests de la D-KEFS) n'a pas pu être intégré à la méta-analyse car la taille d'échantillon (*N*) n'était pas rapportée dans la publication. L'échantillon de la distribution globale est de *N* = 17 042, pour un nombre total de tailles d'effet *k* = 135. Le tableau 6 présente les résultats pour l'ensemble des tailles d'effet (c.-à-d. sans prise en compte des modérateurs). Les trois premières colonnes indiquent la distribution analysée, le nombre de tailles d'effet (*k*) ainsi que le nombre de sujets individuels (*N*). Les colonnes 4 à 7 indiquent la moyenne des tailles d'effet observées ($\underline{r_o}$), l'intervalle de confiance à 95% (IC 95%), l'intervalle de prédiction à 90% (IP 90%) ainsi que la statistique d'hétérogénéité I^2.

Pour l'ensemble des échantillons (*k* = 135), la taille d'effet global est de .69, avec un IC 95% allant de .66 à .71. L'indice I^2 s'élève à 89.84, ce qui suggère que la distribution est hautement hétérogène (Higgins et al., 2003). Cette forte hétérogénéité dans la distribution globale peut évoquer la présence d'éventuels modérateurs. Nous analyserons d'abord les modérateurs liés à la méthodologie employée, ensuite les modérateurs relatifs aux caractéristiques des outils analysés et, enfin, les modérateurs correspondant aux caractéristiques des sujets de l'échantillon.

Tableau 6. *Résultats de la méta-analyse pour l'ensemble de la distribution.*

			Méta-analyse			
Distribution	*k*	*N*	$\bar{r}_o$	IC 95%	IP 90%	I^2
Globale	135	17 042	.69	.66, .71	.38, .86	89.84

Note. *k* = nombre de tailles d'effet ; *N* = échantillon total ; $\bar{r}_o$ = moyenne des corrélations observées ; IC 95% = intervalle de confiance à 95% ; 90% PI = intervalle de prédiction à 90% ; I^2 = rapport entre la vraie hétérogénéité et la variance totale.

2. Résultats par modérateur

a. Méthodologie test-retest

Formes de l'outil. Parmi les échantillons pour lesquels une version identique de l'outil a été employée lors des deux passations (k = 126), on observe une taille d'effet de .68, tandis qu'elle est de .72 lorsque les auteurs ont employé des versions parallèles d'un même outil, ce qui est le cas pour 8 tailles d'effet uniquement (k = 8) (Tableau 6). Les intervalles de confiance s'entrecoupent, ce qui suggère une différence non significative entre les moyennes des deux distributions. De plus, les indices I^2 sont respectivement de 90.21 et de 76.74, suggérant une hétérogénéité élevée pour les indices de fidélité évalués sur base de formes identiques de l'outil, mais modérée pour les formes parallèles (Higgins et al., 2003).

Intervalle test-retest. Étonnamment, on n'observe aucune corrélation significative entre la durée de l'intervalle test-retest et l'importance de la taille d'effet (Tableau 7). La corrélation est même pratiquement nulle avec r = .07.

Tableau 7. *Résultats de la méta-analyse selon la méthodologie employée (formes de l'outil utilisées lors des deux passations) et corrélation avec la durée de l'intervalle test-retest.*

			Méta-analyse			
Distribution	*k*	*N*	$\bar{r}_o$	IC 95%	IP 90%	I^2
Formes						
- Identiques	126	16 127	.68	.65, .71	.37, .86	90.21
- Parallèles	8	758	.72	.64, .79	.51, .85	76.74
			Corrélation			
Distribution		*r*		*p*	*N*	
Intervalle test-retest		.07		.41	135	

Note. k = nombre de tailles d'effet ; N = échantillon total ; $\bar{r}_o$ = moyenne des corrélations observées ; IC 95% = intervalle de confiance à 95% ; 90% PI = intervalle de prédiction à 90% ; I^2 = rapport entre la vraie hétérogénéité et la variance totale ; r = corrélation de Pearson.
*p < .05.

b. Caractéristiques de l'outil

Type d'outil. On observe un indice de fidélité moyen légèrement plus élevé pour les outils provenant d'une batterie ($\bar{r}_o$= .69) en comparaison aux mesures issues d'une épreuve unique ($\bar{r}_o$= .66) (Tableau 8). Les IC 95% sont respectivement de .66, .72 et .63, .69, indiquant un chevauchement entre les deux distributions. En outre, l'hétérogénéité de la distribution rapportée par l'I^2 est élevée pour les premiers (I^2 = 91.15) et modérée pour les seconds (I^2 = 60.17) (Higgins et al., 2003).

Contenu des items. La moyenne des indices test-retest ($\bar{r}_o$) varie selon le contenu des items (Tableau 8). En effet, elle est de .66 pour les items au contenu visuel (k = 77), de .69 pour les items verbaux (k = 32) et enfin, de .73 pour les items au contenu varié (k = 20). Les autres distributions n'ont pas pu être analysées en raison d'un échantillon trop petit *(k < 5)*. On observe une hétérogénéité plus élevée pour les outils constitués de contenu multiple (I^2 = 95.41 et IP 90% = .20, .93), et visuel (I^2 = 87.33 et IP 90% = .36, .84), tandis qu'elle est modérée pour les items de type verbal (I^2 = 66.36 et IP 90% = .57, .78) (Higgins et al., 2003).

Construits globaux. Les résultats prenant en compte les construits globaux comme modérateurs de la taille d'effet se trouvent en Tableau 8. Les distributions constituées des outils évaluant les fonctions praxiques et la cognition sociale n'ont pu être analysées en raison d'un nombre de tailles d'effet trop petit (k < 5). Parmi l'ensemble des construits globaux utilisés comme modérateurs, les outils évaluant le fonctionnement cognitif global présentent la moyenne test-retest la plus élevée ($\bar{r}_o$= .79) d'une part, mais avec un nombre de tailles d'effet le plus bas (k = 8) d'autre part, ce qui peut limiter la représentativité de la distribution. De plus, la variabilité est également la plus élevée avec I^2 = 97.59 (Higgins et al., 2003) et un large intervalle de prédiction (IP 90% = .19, .96), ce qui suggère une haute hétérogénéité parmi les tailles d'effet observées. Pour les autres construits utilisés comme modérateurs, les outils évaluant les fonctions attentionnelles (k = 23) présentent une moyenne test-retest de $\bar{r}_o$= .63. Le $\underline{r_o}$ est de .67 pour les fonctions visuo-spatiales (k = 19) et exécutives (k = 30), de $\bar{r}_o$= .68 pour les fonctions mnésiques (k = 28). Enfin, elle est de $\bar{r}_o$= .72 pour la distribution des fonctions langagières (k = 31). Seule la distribution des fonctions mnésiques présente une hétérogénéité faible (I^2 = 49.95 et IP 90% = .60, .75), tandis qu'elle est élevée pour l'ensemble des autres distributions (I^2 = de 74.94 à 90.80) (Higgins et al., 2003). Cette forte hétérogénéité suggère que la moyenne test-retest pour ces distributions est à nuancer, en raison de l'influence possible d'autres modérateurs, tandis qu'elle est davantage représentative pour les fonctions mnésiques.

De façon générale, l'ensemble des intervalles de confiance se chevauchent, suggérant que les 6 distributions analysées pour les construits globaux ne sont pas significativement différentes.

Sous-construits. Nous avons mené une étape supplémentaire dans nos analyses en évaluant plus spécifiquement les sous-construits compris au sein des différents construits globaux (Tableau 8). Parmi les fonctions attentionnelles, 3 distributions n'ont pu être analysées en raison d'un échantillon trop petit ($k < 5$). Les outils évaluant l'attention sélective ($k = 7$) présentent une moyenne test-retest $\bar{r}_o = .71$ tandis qu'elle est plus faible pour les outils évaluant la vitesse de traitement ($k = 12$) avec un $\bar{r}_o = .62$. Ensuite, parmi les fonctions exécutives, les distributions de mise à jour et planification n'ont pu être analysées en raison d'un échantillon trop petit ($k < 5$). Pour les distributions liées aux modérateurs inhibition ($k = 6$), raisonnement ($k = 11$) et flexibilité ($k = 7$), les moyennes test-retest sont respectivement de $\bar{r}_o = .63$, .66 et .74. L'ensemble des intervalles de confiance se recoupent, suggérant une différence non significative entre les moyennes des distributions. Les 3 distributions semblent hétérogènes (respectivement IP 90% = .30, .82 ; .33, .85 ; .35, .91) bien que le I^2 pour l'inhibition suggère une hétérogénéité plus modérée ($I^2 = 71.61$) que pour les distributions raisonnement et flexibilité (respectivement, $I^2 = 87.22$ et 92.43) (Higgins et al., 2003). Enfin, pour les sous-construits évaluant les fonctions mnésiques, les distributions de mémoire à long terme et mémoire à court terme / de travail ont une moyenne test-retest similaire ($\bar{r}_o = .68$ et $\bar{r}_o = .67$) avec des intervalles de confiance quasiment identiques (IC 95% = .65, .71 ; .63, .71). L'hétérogénéité est modérée pour la mémoire à long terme ($I^2 = 59.82$) et faible pour la distribution mémoire à court terme / de travail ($I^2 = 28.98$) (Higgins et al., 2003), ce qui peut être liée au nombre de tailles d'effet qui est inférieur pour la dernière, et donc moins sujet aux variations (respectivement $k = 17$ contre $k = 11$).

Consistance interne. Nous avons évalué le lien entre la consistance interne et la stabilité des mesures fournies par l'outil (Tableau 8). La majorité des valeurs de consistance interne rapportées était exprimé en *Z* de Fisher, tandis que seules 2 valeurs étaient exprimées en alpha de Cronbach (α). Nous avons alors mené une corrélation entre les indices de fidélité et de consistance interne exprimés en *Z* uniquement. Les résultats de la corrélation n'indiquent aucun lien significatif entre la consistance de l'outil et la stabilité des mesures qu'il fournit.

c. Caractéristiques des sujets

Type d'échantillon. Les moyennes des indices test-retest pour les distributions de sujets contrôles et cliniques sont relativement similaires avec, respectivement, $\bar{r}_o$= .66 et $\bar{r}_o$= .71 (Tableau 9). Cependant, étant donné que la taille de la distribution contrôle s'élève à k = 117, contre k = 6 pour la distribution clinique, ces deux résultats s'avèrent difficilement comparables. En effet, la seconde est sans doute moins proche de la distribution réelle étant donné le peu d'observations. En parallèle, la distribution mixte (composée de sujets contrôles et cliniques) s'avère significativement supérieure ($\bar{r}_o$= .84) avec un IC 95% = .79, .88 qui ne recouvre pas celui de la distribution contrôle (IC 95% = .64, .69) et très peu celui de la distribution clinique (IC 95% = .60, .80), suggérant une différence significative entre les moyennes. Cependant, cette distribution est composée de peu de tailles d'effet (k = 11). Enfin, l'ensemble des I^2 est élevé (de 80.49 à 95.79), indiquant des distributions très hétérogènes (Higgins et al., 2003).

Tranche d'âge de l'échantillon. Les distributions de sujets enfants (k = 59) et adultes (k = 29) obtiennent une moyenne des tailles d'effet de $\bar{r}_o$= .68 et $\bar{r}_o$= .60 (Tableau 9). La distribution mixte (composée de plusieurs tranches d'âge) (k = 29) présente un $\bar{r}_o$= .74 avec un IC 95% peu large (.69, .77), suggérant que la moyenne est représentative de la distribution. L'hétérogénéité des tailles d'effet est cependant très élevée (I^2 = 94.32) (Higgins et al., 2003).

Moyenne d'âge. Nous avons mené une corrélation entre la moyenne d'âge des participants et l'indice de fidélité test-retest qui s'avère significative (r = .28, p < .05) (Tableau 9). Ces résultats suggèrent qu'il existe un lien entre l'âge des participants et l'effet d'entraînement, et donc, indirectement, entre l'âge et la stabilité des mesures fournies par l'outil.

Tableau 8. *Résultats de la méta-analyse selon les caractéristiques de l'outil (contenu, construits et sous-construits) et corrélation avec la consistance interne.*

	Méta-analyse					
Distribution	*k*	*N*	$\bar{r}_o$	IC 95%	IP 90%	I^2
Type d'outil						
- Épreuve	29	3 548	.66	.63, .69	.54, .76	60.17
- Batterie	106	13 494	.69	.66, .72	.35, .87	91.15
Contenu des items						
- Verbal	32	4 525	.69	.66, .72	.57, .78	66.36
- Visuel	77	8 843	.66	.62, .69	.36, .84	87.33
- Numérique	*La distribution n'a pas été analysée (k < 5).*					
- Autre	*La distribution n'a pas été analysée (k < 5).*					
- Multiple	20	2 501	.73	.62, .81	.20, .93	95.41
Construits globaux (en gras) et sous-construits						
Fonctionnement cognitif global	8	1 427	.79	.61, .89	.19, .96	97.59
Fonctions attentionnelles	23	2 553	.63	.57, .69	.39, .79	79.79
- Attention sélective	7	1 185	.71	.65, .77	.57, .82	74.03
- Vitesse de traitement	12	1 251	.62	.53, .69	.36, .79	79.49
- Alerte	*La distribution n'a pas été analysée (k < 5).*					
- Attention soutenue	*La distribution n'a pas été analysée (k < 5).*					
- Attention divisée	*La distribution n'a pas été analysée (k < 5).*					
Fonctions exécutives	30	2 838	.67	.60, .73	.30, .86	88.71
- Flexibilité	7	912	.74	.60, .84	.35, .91	92.43
- Inhibition	6	321	.63	.47, .74	.30, .82	71.61
- Raisonnement	11	1 228	.66	.56, .74	.33, .85	87.22
- Mise à jour	*La distribution n'a pas été analysée (k < 5).*					
- Planification	*La distribution n'a pas été analysée (k < 5).*					
Fonctions mnésiques	28	4 311	.68	.65, .70	.60, .75	49.95
- Mémoire à long terme	17	3 179	.68	.65, .71	.59, .76	59.82
- Mémoire à court terme / de travail	11	1 132	.67	.63, .71	.60, .73	28.98
Fonctions langagières	31	3 015	.72	.66, .78	.42, .88	90.80
Fonctions visuo-spatiales	19	1 720	.67	.61, .72	.46, .81	74.94
Fonctions praxiques	*La distribution n'a pas été analysée (k < 5).*					
Cognition sociale	*La distribution n'a pas été analysée (k < 5).*					

	Corrélation		
Distribution	*r*	*p*	*N*
Consistance interne	.12	.31	79

Note. k = nombre de tailles d'effet ; *N* = échantillon total ; $\bar{r}_o$ = moyenne des corrélations observées ; IC 95% = intervalle de confiance à 95% ; 90% PI = intervalle de prédiction à 90% ; I^2 = rapport entre la vraie hétérogénéité et la variance totale ; *r* = corrélation de Pearson.
*p < .05.

Tableau 9. *Résultats de la méta-analyse selon les caractéristiques de l'échantillon (type d'échantillon, tranche d'âge de l'échantillon) et corrélation avec la moyenne d'âge des sujets.*

			Méta-analyse			
Distribution	*k*	*N*	$\bar{r}_o$	IC 95%	IP 90%	I^2
Type d'échantillon						
- Contrôle	117	11 853	.66	.64, .69	.41, .82	81.89
- Clinique	6	931	.71	.60, .80	.47, .85	80.49
- Mixte	11	4 218	.84	.79, .88	.66, .93	95.79
Tranche d'âge de l'échantillon						
- Enfants (< 18 ans)	59	5 712	.68	.35, .70	.55, .77	58.90
- Adultes (18 – 69 ans)	29	2 007	.60	.52, .66	.25, .81	80.88
- Personnes âgées (> 70 ans)	*La distribution n'a pas été analysée (k < 5).*					
- Mixte	46	9 283	.74	.69, .77	.43, .89	94.32

	Corrélation		
Distribution	*r*	*p*	*N*
Moyenne d'âge	.28	.02	66

Note. k = nombre de tailles d'effet ; *N* = échantillon total ; $\bar{r}_o$ = moyenne des corrélations observées ; IC 95% = intervalle de confiance à 95% ; 90% PI = intervalle de prédiction à 90% ; I^2 = rapport entre la vraie hétérogénéité et la variance totale ; *r* = corrélation de Pearson.
*p < .05.

V. Discussion

A. Interprétation des résultats

Notre travail contribue à faire le point sur l'état actuel de la pratique EBA en neuropsychologie clinique francophone. Pour ce faire, l'objectif premier de ce travail était de recenser un maximum d'informations psychométriques sur la liste d'outils prédéfinie afin de déterminer si les outils employés par les neuropsychologues francophones ont fait, ou non, l'objet d'études de validation françaises. Les résultats étaient divers selon les données psychométriques que nous avons considérées. Le second objectif, complémentaire, visait à déterminer, au moyen de corrélations, les variables qui influenceraient les neuropsychologues lors du choix de leurs outils d'évaluation. Enfin, notre dernier objectif visait à évaluer la fidélité test-retest des mesures fournies par nos outils au moyen d'une méta-analyse.

Pour commencer, nous avions émis l'hypothèse que le pourcentage d'utilisation d'un outil ne devrait pas être corrélée avec ses caractéristiques psychométriques. Or, les corrélations que nous avons menées ont révélé des relations significatives entre le pourcentage d'utilisation des outils et la taille de l'échantillon normatif (*N*), le pays de publication des normes, l'existence de normes, ainsi que le pays de publication de l'étude de validité discriminante. Ces résultats contredisent en partie notre hypothèse de départ, que nous avions émise sur base de l'étude de Betz al. (2013). En effet, nous ne nous attendions pas à observer une corrélation significative entre le pourcentage d'utilisation des outils et leurs qualités psychométriques. Cette divergence de résultats s'explique par une raison très simple : les variables qui entretiennent une relation significative avec le pourcentage d'utilisation n'ont, en réalité, pas été analysées par Betz et al. (2013).

Ensuite, nous n'avons observé aucune relation significative entre la date de publication originelle des outils et leur pourcentage d'utilisation, ce qui contraste avec les résultats issus de l'étude de Betz et al. (2013). En effet, les auteurs avaient mis en évidence que les outils les plus fréquemment utilisés par les logopèdes étaient caractérisés par une forte composante historique et avaient déjà fait l'objet de plusieurs révisions. Nous pouvons expliquer cette absence de significativité par le fait que la liste d'outils que nous avons utilisée est composée de peu d'outils récents. En effet, lorsqu'on compare la première date de publication des outils avec la date des dernières normes répertoriées, on remarque que la majorité de ces outils ont, eux aussi, déjà bénéficié de plusieurs mises à jour. Concrètement, la quasi-totalité des outils a initialement

été publiée au 20ième siècle et bénéficie malgré tout de normes datant des années 2000. Ainsi, bien que nos résultats ne rejoignent pas ceux de Betz et al. (2013), ils ne nous permettent pas d'affirmer que les neuropsychologues n'ont pas tendance à employer des outils déjà bien établis historiquement.

Par ailleurs, nos résultats mettent clairement en évidence la qualité des normes utilisées par les neuropsychologues. En effet, parmi les études répertoriées, nous avons vu que les plus fréquentes sont les études de standardisation. Ce résultat est rassurant étant donné que les normes représentent un outil indispensable à l'interprétation des résultats du patient. Plusieurs éléments contribuent à assurer la représentativité de l'échantillon normatif par rapport au patient, et donc à assurer une interprétation adéquate de sa performance. Premièrement, nos résultats montrent que les études de standardisation sont le type d'étude de validation le plus souvent mené sur des populations franco-européennes. De plus, nous avons observé une corrélation significative entre le pourcentage d'utilisation des outils et le pays de publication des études normatives, signifiant que les outils les plus utilisés sont aussi ceux qui sont normés en France. Deuxièmement, la majorité des études de standardisation proposent une présentation des scores sous forme de percentiles, qui permettent une interprétation des scores du sujet plus exacte, car non influencée par la distribution des scores de la population normative (Brooks et al., 2009). Troisièmement, une grande majorité des normes disponibles distinguent plusieurs sous-groupes par âge, et par niveau socio-culturel. Cette distinction est importante car nous connaissons l'influence réciproque existant entre le niveau d'éducation et les capacités cognitives, qui a pu être démontrée à plusieurs reprises (Guerra-Carrillo et al., 2017 ; Lövdén et al., 2020). Dernièrement, la corrélation positive et significative observée entre le pourcentage d'utilisation et la taille de l'échantillon révèle que les neuropsychologues emploient des normes constituées sur base d'un large échantillon, ce qui est plus représentatif de la population. Tous ces éléments mettent en évidence que les outils employés par les neuropsychologues français sont adéquatement normés. Cependant, ces résultats attestant de la bonne qualité des normes ne nous permettent pas d'affirmer que les neuropsychologues choisissent leurs outils en raison de la qualité de leurs normes. Il se peut également que ces outils soient correctement normés précisément parce qu'ils sont les plus fréquemment utilisés.

En ce qui concerne la validité, nous avons vu qu'à peu près la moitié des outils ont fait l'objet d'une étude de validité. Nos résultats sont similaires à ceux d'une étude menée aux États-Unis qui avait mis en évidence que seuls 55% des études répertoriées à propos de différents outils d'évaluation rapportaient une preuve de validité (Hogan & Agnello, 2004). Nous avons vu que la validité de construit est le plus souvent rapportée (au moyen d'analyses factorielles, ou via des preuves de validité concourante ou convergente), suivie de la validité discriminante. Bien qu'elle ne soit pas rapportée pour la majorité des outils, ce résultat n'est pas étonnant étant donné que tous n'ont pas été conçus dans un objectif de diagnostic ou de détection d'une population clinique spécifique. Nous avons pu également observer que peu d'outils à visée diagnostique ont été validés sur base d'une population française. Cela peut s'avérer problématique si les professionnels utilisent ces instruments, car cela signifierait qu'ils utilisent des normes issues de populations cliniques étrangères, dans le but de contribuer à un diagnostic. En effet, nous savons que l'emploi de normes issues d'un autre pays peut entraîner des erreurs diagnostiques (Alberto & Marcopulos, 2008 ; Raudeberg et al., 2018). Cependant, nous pouvons mettre ce résultat en lien avec la corrélation positive et significative obtenue entre le pourcentage d'utilisation et le pays de publication des normes diagnostiques. Cette relation indique que les neuropsychologues veillent à utiliser des normes diagnostiques issues d'une population francophone. Ainsi, bien que peu d'outils diagnostiques ont fait l'objet d'une étude de validité discriminante en France, nos résultats montrent que les cliniciens veillent à utiliser ceux pour lesquels ils disposent de normes diagnostiques françaises. De plus, de façon intéressante, nous avons observé une relation positive entre le pourcentage d'utilisation des outils et la sensibilité de leurs mesures, tandis qu'elle est négative pour leur spécificité. Bien qu'elles ne soient pas significatives et mériteraient d'être approfondies, ces corrélations semblent indiquer que les neuropsychologues prêtent davantage attention à utiliser des mesures sensibles, plutôt que spécifiques. Enfin, la validité divergente des outils est rarement évaluée. Il peut en effet paraître plus logique pour les développeurs d'outils de chercher à démontrer que leurs instruments fournissent des mesures valides en les comparant à des outils qui évaluent des construits similaires ou identiques. De façon générale, quel que soit le type de validité que nous avons étudié, peu d'outils ont fait l'objet de l'une des différentes méthodes de validation directement en France.

Enfin, la fidélité des mesures, quant à elle, est plus fréquemment évaluée au moyen d'une méthode test-retest plutôt qu'une méthode inter-juges. Nous avons observé une relation non significative entre la fidélité test-retest et le pourcentage d'utilisation. Cette absence de relation peut s'expliquer par le fait que les tests intégrés à nos analyses présentent globalement une fidélité test-retest similaire. Tout de même, il est intéressant de noter que la relation observée tend plutôt à être négative, ce qui peut révéler que les outils les plus fidèles ne sont pas les plus utilisés par les neuropsychologues francophones. Cette relation mériterait d'être davantage investiguée. En parallèle, la fidélité inter-juges, bien qu'elle soit moins régulièrement étudiée, semble plus essentielle pour certains outils que d'autre. Par exemple, elle nous semble plus utile lorsque l'outil a une visée diagnostique (Youngstrom et al., 2017), ou lorsque la cotation des résultats implique une part de subjectivité, et requiert donc la mise en place de règles de cotation très détaillées (c'est le cas par exemple du test de Hayling), ce qui n'est pas le cas de la majorité des outils. De façon similaire aux études de validité, la fidélité des mesures est rarement évaluée sur base d'une population francophone.

Conjointement à ces résultats, la méta-analyse que nous avons menée a révélé que la fidélité test-retest des mesures fournies par nos outils est globalement bonne : en effet, la taille d'effet de la distribution globale est élevée et similaire à celle rapportée par la méta-analyse menée auparavant par Calamia et al. (2013). Plusieurs résultats ressortent de l'analyse des différents modérateurs et méritent d'être discutés.

En premier lieu, nous n'avons, de façon étonnante, observé aucun lien significatif entre la durée de l'intervalle et la taille d'effet. La valeur de la corrélation est même très faible, voire nulle. Ces résultats entrent en contradiction avec les études précédentes qui mettaient en évidence une relation entre la durée de l'intervalle et l'effet d'entraînement, et donc avec l'indice de fidélité test-retest (Calamia et al., 2012 ; Calamia et al., 2013 ; Scharfen et al., 2018). Ce résultat peut s'expliquer par la large diversité d'outils et de sujets intégrés à notre analyse : aucune limite n'a été établie à ce sujet. Or, il se pourrait que l'effet produit par la durée de l'intervalle varie selon le construit évalué, ou encore selon l'âge des participants. Nous pensons que la durée d'intervalle gagnerait à être analysée de façon plus scindée, c.-à.-d. en distinguant les différentes construits évalués ainsi que les différentes tranches d'âge des sujets.

Ensuite, nous avons observé un lien entre l'âge des sujets et les indices test-retest, ce qui rejoint des résultats mis en évidence auparavant (Calamia et al., 2012 ; Calamia et al., 2013 ; Scharfen et al., 2018). Nous avons analysé l'effet de l'âge à travers les distributions de différentes tranches d'âge, ainsi qu'au moyen d'une corrélation entre la moyenne d'âge et la fidélité test-retest. Lors de la première analyse, nous avons défini une tranche d'âge large pour la population « adultes », qui s'étendait jusque 69 ans. Par conséquent, la tranche « personnes âgées » n'était pas constituée de suffisamment d'observations pour pouvoir être analysée. En ce qui concerne les enfants, il aurait pu être judicieux de distinguer une tranche d'âge pour les enfants d'âge préscolaire/scolaire et les adolescents, afin d'obtenir des analyses plus détaillées. Les résultats issus de l'analyse distincte des différentes distributions ne nous ont pas permis de tirer de conclusions. En revanche, l'analyse corrélationnelle révèle une relation faible mais significative et positive entre la moyenne d'âge et l'indice test-retest. Cette relation peut s'expliquer de deux façons : (1) la stabilité des mesures est meilleure pour les outils d'évaluation destinés aux adultes, (2) les fonctions cognitives se stabilisent avec l'âge. Bien que les deux propositions ne soient pas nécessairement mutuellement exclusives, nous penchons plutôt pour la seconde explication. En effet, une étude regroupant une tranche d'âge plus restreinte (des sujets âgés de 52 à 80 ans) n'a mis en évidence aucun lien entre l'âge et l'effet d'entraînement pour des tests exécutifs et attentionnels (Lemay et al., 2004). À l'inverse, une autre étude (Spencer et al., 2003) a démontré que, dans un groupe restreint d'enfants âgés de 5 à 8 ans, les fonctions cognitives étaient moins stables chez les plus âgés d'entre eux. En effet, ceux-ci se développeraient plus rapidement au niveau cognitif, entraînant un biais test-retest dans le calcul de l'indice de fidélité. La littérature met en évidence que les fonctions cognitives se stabilisent avec l'âge, ce qui explique que nos résultats mettent en évidence une corrélation positive significative entre l'âge et la fidélité test-retest des outils.

Pour poursuivre, en ce qui concerne l'effet modérateur produit par le domaine cognitif évalué par l'outil, bien qu'aucune distribution ne soit significativement différente des autres, nous avons remarqué quelques légères divergences. Nous avons observé une fidélité test-retest supérieure pour les outils évaluant le fonctionnement global (qui évaluent donc plusieurs fonctions cognitives). La distribution est fortement hétérogène, ce qui remet en question l'interprétation d'une moyenne unique et révèle la présence d'autres modérateurs. Cette forte hétérogénéité est sans doute liée au fait que ces outils évaluent des construits différents. Ensuite, nous avons vu que la fidélité des outils évaluant les fonctions langagières est supérieure à celle des outils évaluant les fonctions attentionnelles ou les fonctions mnésiques. Ces résultats

trouvent sens en rejoignant partiellement ceux issus d'une autre étude (Ivnick et al., 1999), qui suggérait que les fonctions cognitives les plus stables dans le temps seraient les fonctions langagières, suivies des fonctions attentionnelles et enfin, en dernier lieu, les fonctions mnésiques. La faible hétérogénéité observée au sein de la distribution des fonctions mnésiques est confirmée par les moyennes test-retest similaires des deux sous-construits qui la composent (mémoire à long terme et mémoire à court terme / de travail). Par ailleurs, nous pensons que la forte hétérogénéité observée dans les autres distributions s'explique par les différents sous-construits qui les composent. Cette suggestion est cependant difficile à confirmer, étant donné que plusieurs distributions de sous-construits n'ont pas pu être analysées individuellement en raison d'un nombre de tailles d'effet trop faible. En résumé, nous avons pu observer de légères variations dans les moyennes test-retest selon le construit évalué bien que celles-ci ne soient pas significatives. Il est difficile de déterminer sur base de nos résultats si ces variations sont liées au fonctionnement des outils ou à la stabilité des fonctions cognitives. En effet, des auteurs suggèrent que la stabilité des mesures fournies par nos outils d'évaluation varie plutôt selon le construit évalué, reflétant en réalité davantage différents niveaux de stabilité selon la fonction cognitive (Ivnick et al., 1995 ; Ivnick et al., 1999).

En ce qui concerne les versions de l'outil employées lors de chaque passation, nos résultats ne montrent aucune différence significative. Ces résultats contrastent avec ceux obtenus par Calamia et al. (2013), qui ont observé une fidélité légèrement inférieure dans le cas de l'emploi de formes parallèles. Cela peut s'expliquer par le faible nombre de tailles d'effet observées pour la distribution des formes parallèles. De plus, la distribution des formes identiques est associée à une forte hétérogénéité, suggérant la présence d'autres modérateurs qui auraient une influence sur l'importance de l'indice de fidélité test-retest. Nos résultats ne nous permettent donc pas de confirmer l'existence d'une différence entre les deux méthodes, et mériteraient d'être davantage investigués.

Enfin, on retrouve une influence du contenu des items. Des auteurs avaient par exemple mis en évidence un biais test-retest moindre pour les outils composés d'items numériques, suggérant par conséquent une meilleure fidélité test-retest (Scharfen et al., 2018). Malheureusement, nous disposions d'un nombre insuffisant de tailles d'effet pour pouvoir analyser ces données. Nos résultats montrent en revanche une fidélité test-retest inférieure pour les items au contenu visuel, en comparaison aux items au contenu verbal, ce qui rejoint les observations issues d'autres études (Benedict & Zgaljardic, 1998 ; Salthouse & Tucker-Drob, 2008). Enfin, on, observe une plus grande fidélité pour les outils aux items multiples, ce qui contraste avec l'étude

de Villado et al. (2016), qui avaient mis en évidence un gain de score plus important pour ces outils lors de la seconde passation. Globalement, bien que nos résultats permettent de confirmer une influence du type d'item sur l'importance de l'effet d'entraînement, ceux-ci contribuent également à renforcer les divergences au sein de la littérature.

En résumé, en recoupant les résultats issus de nos deux premiers objectifs, ceux-ci vont plutôt dans le sens de notre hypothèse de départ selon laquelle les qualités psychométriques des outils ne sont pas un critère primordial dans le choix d'un outil. Les neuropsychologues français utilisent des outils adéquatement normés, mais qui ne présentent pas nécessairement une bonne fidélité ou validité de mesure. En effet, bien que notre méta-analyse ait mis en évidence une bonne fidélité test-retest globale, celle-ci semble varier selon différents modérateurs et est rarement évaluée sur des populations francophones. En somme, les paramètres de validité et fidélité des mesures semblent moins primordiaux aux neuropsychologues français. Néanmoins, nous pouvons émettre plusieurs pistes susceptibles d'expliquer ce manque de considération. Celles-ci ne se veulent ni exhaustives, ni mutuellement exclusives.

Premièrement, il est possible que les neuropsychologues soient conscients que les outils qu'ils utilisent ne présentent pas les meilleures caractéristiques psychométriques. En effet, nos résultats suggèrent que la validité et la fidélité des mesures ne représentent pas un critère de choix, mais rien ne nous permet d'affirmer que ces données n'entrent pas du tout en compte dans leur sélection. Au contraire, 50.2% des neuropsychologues français affirment que certains des outils qu'ils emploient ne présentent pas de bonnes qualités psychométriques (Branco Lopes et al., 2019). Cela indique que la moitié d'entre eux prennent ces données en considération et qu'ils sont conscients des qualités et défauts de leurs outils. Ainsi, on peut supposer que, pour une partie des professionnels, leur choix s'opère par dépit, à défaut de disposer d'outils de meilleure qualité.

Deuxièmement, l'utilisation fréquente d'outils indépendamment de leurs caractéristiques psychométriques peut s'expliquer par un manque de compréhension de leur intérêt. Des auteurs pointent d'ailleurs le manque de connaissance de la psychométrie des psychologues (Borsboom, 2006 ; Bowden, 2017). On peut dès lors supposer qu'il est difficile pour les professionnels concernés de considérer des données qu'ils comprennent mal, ou dont l'intérêt leur est méconnu.

Troisièmement, la non-considération des caractéristiques psychométriques des outils peut s'expliquer par un manque d'accès à l'information. D'une part, tous les professionnels n'ont pas accès aux bases de données de littérature scientifique. D'autre part, dans le cas des épreuves uniques, il arrive fréquemment que les neuropsychologues utilisent des épreuves photocopiées : 87% des neuropsychologues sondés dans l'étude de Branco Lopes et al. (2019) révèlent avoir recours à la reproduction de documents. Or, utiliser des versions photocopiées plutôt que des publications officielles peut mener à de la perte d'informations.

Quatrièmement, comme nous l'avons vu, certains outils sont utilisés par une grande majorité de neuropsychologues (par exemple, le test de Stroop, le TMT ou encore, le RL/RI16). La « popularité » de ces outils peut entraîner un biais cognitif chez les professionnels. Ce biais peut les amener à penser à tort, et sans vérification, que leurs outils sont valides et fidèles (Maltzman, 2013). À l'inverse, il se peut que les cliniciens se montrent davantage réticents face à des outils plus récents et moins largement utilisés.

Dernièrement, nous pouvons noter un manque de recherche dans le domaine, qui se marque principalement en Europe francophone. En effet, nous avons pu observer qu'il était rare qu'un outil ait fait l'objet d'une étude de validité ou fidélité en France. Une alternative pour les neuropsychologues pourrait être de se renseigner dans les études issues d'autres pays, mais celles-ci peuvent s'avérer plus difficiles d'accès. En outre, il n'existe aucune obligation pour les développeurs d'outils d'évaluer les caractéristiques psychométriques de ceux-ci (Borsboom, 2006). Des auteurs ont d'ailleurs mis en évidence que les preuves de validité et fidélité des outils étaient encore trop peu rapportées par les chercheurs (Hogan & Agnello, 2004 ; Slaney et al., 2009 ; Vacha-Haase et al., 2002). Par conséquent, face à ce manque de données, il peut s'avérer difficile pour les neuropsychologues soucieux de mettre en place une évaluation *evidence-based* de considérer les caractéristiques psychométriques comme un critère de choix essentiel.

B. Perspectives et limites

Nos résultats montrent globalement un manque d'intégration des preuves issues de la littérature lors de la mise en place de l'évaluation de patients. D'une part, cela peut être dû à un manque de considération de ce pilier par les neuropsychologues. En réalité, nos résultats mettent en évidence qu'ils prêtent davantage d'importance au pilier correspondant aux besoins et préférences du patient. En effet, ils utilisent des normes de bonne qualité, c.-à-d. adaptées à la majorité des patients. De plus, nous pouvons supposer que le choix des outils s'opère sur base des plaintes du patient. En parallèle, il semblerait que l'expertise clinique pousse les professionnels à utiliser fréquemment certains outils qu'ils maîtrisent et en lesquels ils ont confiance. Cependant, ce pilier ne se limite pas à la maîtrise d'outils : il requiert également pour le neuropsychologue de continuer à développer ses compétences en utilisant les preuves issues de la littérature. En réalité, selon l'APA (2021), le pilier « expertise clinique » est également censé faire le lien entre les preuves issues de la littérature et les données cliniques du patient. Or, le manque de considération des caractéristiques psychométriques des outils suggère justement un manque d'intégration des preuves issues de la littérature dans la planification de l'évaluation.

D'autre part, nos résultats permettent de mettre en évidence un manque de données au sein même de la littérature, particulièrement en Europe francophone. Or, tant que le domaine n'est pas plus développé au sein de la littérature scientifique, il sera difficile pour les professionnels de faire des données psychométriques un critère de choix essentiel.

Cependant, le manque de données issues de la littérature n'empêche pas la prise en compte des données psychométriques des outils d'évaluation. En effet, les neuropsychologues peuvent tout de même se servir des données actuellement disponibles. Plusieurs pistes sont envisageables pour intégrer au mieux les trois piliers et tendre davantage vers une évaluation *evidence-based*.

Des auteurs proposent différentes procédures pour aider les professionnels à mener une évaluation *evidence-based.* Par exemple, Youngstrom et al. (2017) suggèrent trois stratégies pour faciliter l'EBA : (1) accorder davantage d'attention aux caractéristiques psychométriques pertinentes selon l'objectif visé par l'évaluation (*prediction, prescription* ou *process*), (2) se concentrer sur les cas de patients le plus souvent rencontrés, afin de préparer un *set* d'outils à adapter selon les situations, (3) utiliser une méthode suffisamment bonne, plutôt que de rechercher à tout prix la méthode parfaite, qui n'existe pas dans tous les cas.

En parallèle, Bornstein (2017) propose 9 étapes à mettre en place pour développer une démarche évaluative qui intègre les trois piliers nécessaires à l'EBP. Les étapes qu'il suggère intègrent plusieurs compétences relatives à l'EBP, dont la capacité à sélectionner, administrer et interpréter les outils d'évaluation.

Enfin, une autre piste serait de développer, chez les étudiants futurs psychologues, un intérêt pour la psychométrie à travers sa compréhension. L'objectif serait de les amener à intégrer les informations dont ils disposent directement à leur pratique (pour des pistes concrètes à appliquer chez les étudiants, nous vous invitons à consulter la publication de Haverkamp, 2013). De plus, développer un intérêt pour les données psychométriques des outils d'évaluation pourrait, *in fine*, contribuer à développer la recherche dans le domaine. De manière plus générale, les universités ont un rôle à jouer pour former les étudiants à développer une pratique *evidence-based* qui intègre efficacement les différents piliers (Durieux et al., 2017).

Plusieurs limites méthodologiques sont à considérer dans notre travail. D'abord, la liste d'outils issue de l'étude de Branco Lopes et al. (2019) que nous avons utilisée n'a pas été initialement conçue pour les analyses effectuées. Il en résulte plusieurs défauts. Premièrement, beaucoup d'outils issus de la liste ne sont pas traduits en français. Il est donc normal que ceux-ci fassent partie des moins utilisés, et ce, que leurs caractéristiques psychométriques soient bonnes ou mauvaises. Il aurait été intéressant d'utiliser une liste exclusivement composée d'outils publiés et/ou traduits en français. Ainsi, nous aurions pu réaliser nos analyses sur des outils réellement accessibles aux neuropsychologues français (autrement dit, sans barrière linguistique). Cela nous aurait permis de mieux déterminer ce qui influence la sélection de leurs outils, sans que le critère de la langue intervienne. Deuxièmement, les batteries ont été considérées dans leur entièreté dans le sondage. Or, il est fréquent que les cliniciens n'utilisent que certaines épreuves issues de batteries (par exemple, il est certain que tous les subtests issus de la batterie NEPSY-II ne sont pas utilisés à la même fréquence par les neuropsychologues). Il aurait donc été préférable de mettre en lien les différentes caractéristiques psychométriques et fréquences d'utilisation des épreuves considérées de façon individuelle. Enfin, le sondage mené par Branco Lopes et al. (2019) que nous avons utilisé pour nos analyses n'était destiné qu'à des neuropsychologues français. Or, il est probable que les habitudes et pratiques évaluatives françaises et belges soient différentes. Ainsi, il pourrait être intéressant de reproduire ce travail sur des cliniciens belges afin de mieux connaître leurs habitudes.

Ensuite, concernant la méta-analyse menée, nous n'avons pas pu assurer l'indépendance de l'ensemble des tailles d'effet intégrées à notre méta-analyse. Certaines tailles d'effet issues d'un même échantillon ont été considérées comme indépendantes. Or, l'indépendance des tailles d'effet fait partie des postulats statistiques de base pour mener une méta-analyse (Beretvas & Pastor, 2003). De plus, plusieurs distributions n'ont pu être analysées en raison d'un nombre trop faible d'observations. Certaines distributions associées aux sous-construits n'ont pu être analysée. Par exemple, trois distributions issues de la distribution des fonctions attentionnelles n'ont pu être investiguées. Or, il aurait été utile d'analyser ces distributions étant donné l'hétérogénéité observée dans la distribution correspondant au construit global.

Une autre limite concerne les analyses corrélationnelles que nous avons menées. En effet, afin de pouvoir corréler chaque pourcentage d'utilisation à une donnée unique, nous avons été contraints de moyenner certaines informations entre elles. Par exemple, pour la NEPSY-II, nous avons dû calculer les moyennes pondérées des indices de fidélité test-retest relatifs aux différentes épreuves issues de la batterie. Or, cela a pu entraîner un manque d'exactitude dans nos résultats. D'une part, comme nous l'avons énoncé précédemment, nous ne pouvons pas affirmer que toutes les épreuves issues de cette batterie sont utilisées à la même fréquence. D'autre part, le calcul d'une moyenne unique n'est pas forcément représentatif de la batterie prise dans son ensemble.

Enfin, une dernière limite concerne la qualité de l'encodage des données. En effet, celui-ci a été réalisé par un encodeur unique. Or, dans l'idéal, plusieurs juges doivent se charger de l'extraction de données. Cela permet ensuite de comparer les données récoltées et discuter des éventuelles divergences observées.

C. Conclusion

Ce travail permet dans un premier temps de faire le point sur l'état actuel des données issues de la littérature à propos des caractéristiques psychométriques des outils employés par les neuropsychologues. Nous avons d'abord pu mettre en évidence un manque de recherche dans ce domaine en Europe francophone.

Dans un second temps, nous avons cherché à déterminer si les données psychométriques des outils étaient liées à leur fréquence d'utilisation. Nous avons mis en évidence que les outils les plus utilisés par les neuropsychologues français se démarquent par la qualité de leurs données normatives. En revanche, la validité et fidélité des outils n'apparaissent pas comme des critères primordiaux. Le manque de connaissance de la psychométrie des neuropsychologues ainsi que le manque de données issues de la littérature font partie des pistes explicatives que nous avons suggérées.

Par ailleurs, la méta-analyse menée révèle globalement une bonne fidélité test-retest des mesures fournies par nos outils. Nous n'avons cependant pas pu confirmer l'ensemble des effets modérateurs auparavant mis en évidence dans la littérature.

Bien que nous pensons que les neuropsychologues ont encore des progrès à faire dans la mise en place de l'EBA, les différents obstacles discutés (notamment, le manque de données issues de la littérature) nous amènent également à suggérer qu'il est nécessaire de rester indulgent envers les professionnels. Nous pointons d'ailleurs plusieurs pistes susceptibles de faciliter la mise en place de l'EBA à l'avenir.

Enfin, nous avons émis plusieurs limites liées à la méthodologie de ce travail. Celui-ci mériterait en effet d'être perfectionné afin de mieux cerner les pratiques des neuropsychologues francophones.

VI. Références

A. Références théoriques

Akobeng, A. K. (2006). Understanding diagnostic tests 1: Sensitivity, specificity, and predictive values. *Acta Paediatrica, 96*(30), 338-341. https://doi.org/10.1111/j.1651-2227.2006.00180.x

Alberto, L. F., & Marcopulos, B. A. (2008). A comparison of normative data for the Trail Making Test from several countries: Equivalence of norms and considerations for interpretation. *Scandinavian Journal of Psychology, 49*(3), 239-246. https://doi.org/10.1111/j.1467-9450.2008.00637.x

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* American Educational Research Association.

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

American Psychological Association. (2000). Report of the Task Force on Test User Qualifications. *Practice and Science Directorates.*

American Psychological Association. (2021, April 21). *Policy statement on evidence-based practice in psychology.* http://www.apa.org/practice/guidelines/evidence-based-statement

Basso, M. R., Bornstein, R. A., & Lang, J. M. (1999). Practice effects on commonly used measures of executive function across twelve months. *The Clinical Neuropsychologist, 13*(3), 283-292. https://doi.org/10.1076/clin.13.3.283.1743

Benedict, R. H. B., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology, 20*(3), 339-352. https://doi.org/10.1076/jcen.20.3.339.822

Beretvas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement, 63*(1), 75–95. https ://doi.org/10.1177/ 0013164402239318

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools, 44*(2), 133-146. https://doi.org/10.1044/0161-1461(2012/12-0093)

Board of Directors. (2007). American Academy of Clinical Neuropsychology (AACN) practice guidelines for neuropsychological assessment and consultation. *The Clinical Neuropsychologist, 21*(2), 209-231. https://doi.org/10.1080/13825580601025932

Borenstein, M., Hedges, L., & Rothstein, H. (2007). *Meta-analysis: Fixed effect vs. random effects*. Meta-Analysis. https://www.meta-analysis.com/downloads/Meta-analysis%20fixed%20effect%20vs%20random%20effects%20072607.pdf

Bornstein, R. F. (2017). Evidence-based psychological assessment. *Journal of Personnality Assessment, 99*(4), 435-445. https://doi.org/10.1080/00223891.2016.1236343

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425-440. https://doi.org/10.1007/s11336-006-1447-6

Bowden, S. C. (2017). Why do we need evidence-base neuropsychological practice? In S. C. Bowden (Ed.), *Neuropsychological assessment in the age of evidence-based practice* (pp. 22-37). Oxford University Press.

Bowden, S. C., & Finch, S. (2017). When is a test reliable enough and why does it matter? In S. C. Bowden (Ed.), *Neuropsychological assessment in the age of evidence-based practice* (pp. 129-156). Oxford University Press.

Branco Lopes, A., Leal, G., Malvy, L., Wauquiez, G., Ponchel, A., Rivera, D., & Arango-Lasprilla, J. C. (2019). Neuropsychology in France. *Applied Neuropsychology: Adult.* 1-12. https ://doi.org/https://doi.org/10.1080/23279095.2019.1633329

Brooks, B. L., Strauss, E., Sherman, E. M. S., Iverson, G. L., & Slick, D. J. (2009). Developments in neuropsychological assessment: Refining psychometric in clinical interpretive methods. *Canadian Psychology, 50*(3), 196-209. https://doi.org/10.1037/a0016066

Calamia, M., Markon, C., & Traniel D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test-retest correlations. *The Clinical Neuropsychologist, 27*(7), 1077-1105. https://doi.org/10.1080/13854046.2013.809795

Calamia, M., Markon, K, & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist, 26*(4), 543-570. https://doi.org/10.1080/13854046.2012.680913

Chelune, G. J. (2017). Evidence-based practice in neuropsychology. In S. C. Bowden (Ed.), *Neuropsychological assessment in the age of evidence-based practice* (pp. 200-231). Oxford University Press.

Chevalier, N. (2010). Les fonctions exécutives chez l'enfant : concepts et développement. *Canadian Psychology, 51*(30), 149-163. https ://doi.org/10.1097/a0020031

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale.

Colombo, F., Amieva, H., Lecerf, T., & Verdon, V. (2016) La norme en neuropsychologie, un concept à facettes multiples. *Revue de Neuropsychologie, 8*(1), 61-69. https ://doi.org/10.1684/nrp.2016.0365

Cooper, H. (2017). *Research synthesis and meta-analysis: A step-by-step approach*. Los Angeles, CA : Sage.

De Mendonça Lima, C. A., Levav, I., Jacobsson, L., & Rutz, W. (2003). Stigma and discrimination against older people with mental disorders in Europe. *International Journal of Geriatric Psychiatry, 18*(8), 679-682. https://doi.org/10.1002/gps.877

Dickinson, M. D., & Hiscock, M. (2011). The Flynn effect in neuropsychological assessment. *Applied Neuropsychology, 18*(2), 136-142. https://doi.org/10.10/09084282.2010.547785

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

Evidence-Based Medicine Working Group. (1992). Evidence-based medicine: a new approach to teaching the practice of medicine. *Journal of the American Medical Association 268* (17), 2420–2425. https://doi.org/10.1001/jama.1992.03490170092032

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*(1), 29–51. https://doi.org/10.1037/0033-2909.95.1.29

Flynn, J. R., & Weiss, L. G. (2007). American IQ gains from 1932 to 2002: The significance of the WISC subtests. *Journal of International Testing, 7*(2), 209–224. https://doi.org/10.1080/15305050701193587

Guerra-Carrillo, B., Katovich, K., & Bunge, S. A. (2017) Does higher education hone cognitive functioning and learning efficacy? Findings from a large and diverse sample. *PloS one, 12*(8). https://doi.org/10.1371/journal.pone.0182276

Haverkamp, B. F. (2013). Education and training in assessment for professional psychology: Engaging the "reluctant student". In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds), *APA handbook of testing and assessment in psychology, volume 2: Testing and assessment in clinical and counseling psychology (pp. 63-82).* American Psychological Association. https://doi.org/10.1037/14048-000

Haynes, R. B., Devereaux, P., & Guyatt, G. H. (2002). Clinical expertise in the era of evidence-based medicine and patient choice. *ACP Journal Club, 136*(2), 11–14. https ://doi.org/10.1136-ebm.7.2.36

Haynes, S. N., O'Brien, W. H., & Kaholokula, J. K. A. (2019). Behavioral assessment of adults in clinical settings. In G. Gorth-Marnat, & J. Wright (Eds.), *Handbook of psychological assessment* (4th ed., pp. 461-501). Academic Press.

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry G. K., & Hart, R. P. (2010). Official position of the American academy of clinical neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts. *The Clinical Neuropsychologist, 24*(8), 1267-1278. https ://doi.org/ 10.1080/13854046.2010.526785

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal, 327(*7414), 557-560. https://doi.org/10.1136/bmj.327.7414.557

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*(4), 802-812. https://doi.org/10.1177/0013164404264120

Ivnick, R. J., Smith, G. E., Lucas, J. A., Petersen, R. C., Boeve, B. F., Kokmen, E., & Tangalos, E. G. (1999). Testing normal older people three or four times at 1- to 2- year intervals: Defining normal variance. *Neuropsychology, 13*(1), 121-127.

Ivnick, R. J., Smith, G. E., Malec, J. F., Petersen; R. C., & Tangalos, E. G. (1995). Long-term stability and intercorrelations of cognitive abilities in older persones. *Psychological Assessment, 7*(2), 155-161.

Korkman, M., Kirk, U., & Kemp, S. (2012). *Bilan neuropsychologique de l'enfant*, seconde édition : NEPSY II (2e éd., version française). Éditions du Centre de Psychologie Appliquée.

Krippendorff, K. (2016). Misunderstanding reliability. *Methodology, 12*(4), 139-144. https://doi.org/10.1027/1614-2241/a000119

Laveault, D., & Grégoire, J. (2014). *Introduction aux théories de tests en psychologie et sciences de l'éducation* (3rd ed.). De Boeck Supérieur.

Lemay, S., Bédard, M.-A., Rouleau, I., & Tremblay, P.-L. (2004). Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *The Clinical Neuropsychologist, 18*(2), 284-302. https://doi.org/10.1080/13854040490501718

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment* (4th ed.). Oxford University Press.

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology*, *92*(6), 1672-1682. https://doi.org/10.1037/0021-9010.92.6.1672

Lövdén, M., Fratiglioni, L., Glymour, M. M., Lindenberger, U., & Tucker-Drob, E. M. (2020). Education and cognitive functioning across the life span. *Psychological Science in the Public Interest*, *21*(1), 6-41. https://doi.org/10.1177/1529100620920576

Maltzman, S. (2013). The assessment process. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds), *APA handbook of*

testing and assessment in psychology, volume 2: Testing and assessment in clinical and counseling psychology (pp. 19-34). American Psychological Association. https ://doi.org/10.1037/14048-000

Meulemans, T., & Seron, X. (2004). *L'exagération et la simulation des troubles* (chap. 8). Mardaga.

Müller, U., Kerns, K. A., & Konkin, K. (2012). Test-retest reliability and practice effects of executive function tasks in preschool children. *The Clinical Neuropsychologist, 26*(2), 271-287. https://doi.org/10.1080/13854046.2011.645558

OCEBM Levels of Evidence Working Group. (2011). *The Oxford levels of evidence 2.* https ://www.cebm.net/index.aspx?o=5653

Raudeberg, R., Iverson, G. L., & Hammar, A. (2018). Norms matter: U.S. normative data underestimate cognitive deficits in Norwegians with schizophrenia spectrum disorders. *The Clinical Neuropsychologist, 33*(1), 58-74. https://doi.org/10.1080/13854046.2019.1590641

Riley, E. N., Combs, H. L., Davis, H. A., & Smith, G. T. (2017). Theory as evidence: criterion validity in neuropsychological testing. In S. C. Bowden (Ed.), *Neuropsychological assessment in the age of evidence-based practice* (pp. 38-55). Oxford University Press.

Roy, A. (2015). Les fonctions exécutives chez l'enfant : des considérations développementales et cliniques à la réalité scolaire. *Développements, 7*, 13-40.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence-based medicine: What it is and what it isn't. *British Medical Journal, 312* (7023), 71-72. https://doi.org/10.1136/bmj.312.7023.71

Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology*, *22*(6), 800-811. https://doi.org/10.1037/a0013091

Satterfield, J. M., Spring, B., Brownson, R. C., Mullen, E. J., Newhouse, R. P., Walker, B. B., & Whitlock, E. P. (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

Scharfen, J., Peter, J. M., Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analyses. *Intelligence, 67,* 44-66. https://doi.org/10.1016/j.intell.2018.01.003 R

Slaney, K. L., Tkatchouk, M., Gabriel, S. M., Maraun, M. D. (2009). Psychometric assessment and reporting practices: Incongruence between theory and practice. *Journal of Psychoeducational Assessment, 27*(6), 465-476. https://doi.org/10.1077/0734282909335781

Spencer, F. H., Bornholt, L. J., & Ouvrier, R. A. (2003). Test reliability and stability of children's cognitive functioning. *Journal of Child Neurology, 18*(1), 5-11. https://doi.org/10.1177/08830738030180010901

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

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*(1), 1-25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Unité de Neuropsychologie de la Faculté de Psychologie, Logopédie et Sciences de l'Éducation de l'Université de Liège. (2018). *Qu'est-ce que la neuropsychologie ?* http ://www.neuropsycho.ulg.ac.be/news/1/15/Qu-est-ce-que-la-neuropsychologie/d,AccueilInformationsGenerales.html

Vacha-Haase, T., Henson, R., K., & Caruso, J., C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Pyschological Measurement, 62*(4), 562-569.

Van Meter, A. (2020). The prediction phase of evidence-based assessment. In E. A. Youngstrom, M. J. Prinstein, E. J. Mash, & R. A., Barkley (Eds.), *Assessment of disorders in childhood and adolescence* (pp. 30-48). Guilford.

Villado, A. J., Randall, J. G., & Zimmer, C. U. (2016). The effect of method characteristics on retest score gains and criterion-related validity. *Journal of Business and Psychology, 31*(2), 233-248. https://doi.org/10.1007/s10869-015-9408-7

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personnality over time. *Journal of Research in Personality, 38*(4)*,* 319-335. https://doi.org/10.1016/j.jrp.2004.03.001

Wechsler, D. (2011). *WAIS-IV – Échelle d'intelligence de Wechsler pour adultes – 4ème édition.* ECPA Pearson.

Youngstrom, E. A., Van Meter, A., Frazier, T. W., Hunsley, J., Prinstein M. J., Ong, M.-L., & Youngstorm, J. K. (2017). Evidence-based assessment as an integrative model for applying psychological science to guide the voyage of treatment. *Clinical Psychology : Science and Practice, 24*(40). 331-363. https ://doi.org/10.1111/cpsp.12207

B. Références des études de validation

Alioto, A. G., Kramer, J. H., Borish, S., Neuhaus, J., Saloner, R., Wynn, M., & Foley, J. M. (2017). Long-term test-retest reliability of the California Verbal Learning Test–second edition. *The Clinical Neuropsychologist*, *31*(8), 1449-1458. https ://doi.org/10.1080/13854046.2017.1310300

Appollonio, I., Leone, M., Isella, V., Piamarta, F., Consoli, T., Villa, M. L., Forapani, E., Russo, A., & Nichelli, P. (2005). The Frontal Assessment Battery (FAB): Normative values in an Italian population sample. *Neurological Sciences*, *26*(2), 108–116 https://doi.org/10.1007/s10072-005-0443-4

Ashendorf, L., Jefferson, A. L., Green, R. C., & Stern, R. A. (2009). Test-retest stability on the WRAT-3 reading subtest in geriatric cognitive evaluations. *Journal of Clinical and Experimental Neuropsychology, 31*(5), 605-610. https://doi.org/10.1080/13803390802375557

Auriacombe, S., Helmer, C., Amieva, H., Berr, C., Dubois, B., & Dartigues, J. F. (2010). Validity of the Free and Cued Selective Reminding test in predicting dementia: The 3C study. *Neurology, 74*(22), 1760–1767. https ://doi.org/10.1212/WNL.0b013e3181df0959

Axelrod, B. N., Meyers, J. E. & Davis, J. J. (2014). Finger Tapping Test Performance as a measure of performance validity. *The Clinical Neuropsychologist, 28*(5) 876-888. https://doi.org/10.1080/13854046.2014.907583

Basso, M. R., Bornstein, R. A., & Lang, J. M. (1999). Practice effects on commonly used measures of executive function across twelve months. *The Clinical Neuropsychologist*, *13*(3), 283-292. https://doi.org/10.1076/clin.13.3.283.1743

Bayard, S., Gély-Nargeot, M.-C., Raffard, S., Guerdoux-Ninot, E., Kamara, E., Gro-Balthazard, F., Jacus, J.-P., Moroni, C., & CPCN-Languedoc-Roussillon. (2017). French version of the Hayling Sentence Completion, part 1: Normative data and guidelines for error scoring. *Archives of Clinical Psychology, 32*(5)*,* 1-7. https ://doi.org/10.1093/arclin/acx010

Beeri, M. S., Schmeidler, J., Sano, M., Wang, J., Lally, R., Grossman, H., & Silverman, J. M. (2006). Age, gender, and education norms on the CERAD neuropsychological battery in the oldest old. *Neurology*, *67*(6), 1006–1010. https ://doi.org/10.1212/01.wnl.0000237548.15734.cd

Beery, K. E., Beery, N. A., & Beery, N. A. (2010). *The Beery-Buktenica Developmental Test of Visual-Motor Integration – 6th Edition*. Pearson.

Bertoux, M., Delavest, M., de Souza, L. C., Funkiewiez, A., Lépine, J. P., Fossati, P. Dubois, B., & Sarazin, M. (2012). Social cognition and emotional assessment differentiate frontotemporal dementia from depression. *Journal of Neurology, Neurosurgery & Psychiatry*, *83*(4), 411-416. https://doi.org/10.1136/jnnp-2011-301849

Bertoux, M., Michalon, S., & Blanc, F. (2020). Validation de la mini-SEA dans une population française variée : Données de référence pour la pratique clinique. *Revue de Neuropsychologie*, *12*(4), 367-375. https://doi.org/10.1684/nrp.2020.0609

Bezdicek, O., Majerova, V., Novak, M., Nikolai, T., Ruzicka, E., & Roth, J. (2013). Validity of the Montreal Cognitive Assessment in the detection of cognitive dysfunction in Huntington's disease. *Applied Neuropsychology: Adult*, *20*(1), 33–40. https://doi.org/10.1080/09084282.2012.670158

Binetruy, M., Mauny, F., Lavaux, M., Meyer, A., Sylvestre, G., Puyraveau, M., Berger, E., Magnin, E., Vandel, P., Galmiche, J., Chopard, G., & RAPID-II study group. (2018). The RAPID-II neuropsychological test battery for subjects aged 20 to 49 years: Norms and cognitive profile. *Revue neurologique*, *174*(1-2), 44–55. https ://doi.org/10.1016/j.neurol.2017.05.010

Bird, C. M., Papadopoulou, K., Ricciardelli, P., Rossor, M. N., & Cipolotti, L. (2003). Test-retest reliability, practice effects and reliable indice for the Recognition Memory Test. *British Journal of Clinical Psychology, 42*(2), 407-425. https ://doi.org/10.1348/014466503322528946

Bird, C. M., Papadopoulou, K., Ricciardelli, P., Rossor, M. N., & Cipolotti, L. (2004). Monitoring cognitive changes: Psychometric properties of six cognitive tests. *British Journal of Clinical Psychology*, *43*(2), 197-210. https://doi.org/10.1348/014466504323088051

Bolen, L. M. (2003). Constructing local age norms based on ability for the Bender-Gestalt Test. *Perceptual and Motor Skills*, *97*(2), 467-476. https://doi.org/10.2466/pms.2003.97.2.467

Brickenkamp, R., Schmidt-Atzert, L., & Liepmann, D. (2015). *D2-R : Test d'attention concentrée révisé.* Hogrefe.

Campo, P., & Morales, M. (2004). Normative data and reliability for a Spanish version of the verbal Selective Reminding Test. *Archives of Clinical Neuropsychology*, *19*(3), 421-435. https://doi.org/10.1016/S0887-6177(03)00075-1

Cangoz, B., Karakoc, E., & Selekler, K. (2009). Trail Making Test: Normative data from 50+ Turkish and elderly. *Journal of the Neurological Sciences*, *283*(1), 317. https://doi.org/10.1016/j.jns.2009.02.289

Carcaillon, L., Amieva, H., Auriacombe, S., Helmer, C., & Dartigues, J. F. (2009). A subtest of the MMSE as a valid test of episodic memory? Comparison with the Free and Cued Reminding Test. *Dementia and Geriatric Cognitive Disorders*, *27*(5), 429-438. https://doi.org/10.1159/000214632

Clerici, F., Ghiretti, R., Di Pucchio, A., Pomati, S., Cucumo, V., Marcone, A., Vanacore, N., Mariani, C., & Cappa, S. F. (2017). Construct validity of the Free and Cued Selective Reminding Test in older adults with memory complaints. *Journal of Neuropsychology*, *11*(2), 238-251. https://doi.org/10.111/JNP.12087

Crawford, J. R., Obonsawin, M. C., & Allan, K. M. (1998). PASAT and Components of WAIS-R Performance: Convergent and discriminant validity. *Neuropsychological Rehabilitation, 8*(3), 255-272. https://doi.org/10.1080/713755575

Cumming, T. B., Churilov, L., Linden, T., & Bernhardt, J. (2013). Montreal Cognitive Assessment and Mini-Mental State Examination are both valid cognitive tools in stroke. *Acata Neurologica Scandinavica, 128*(2), 122-123. https ://doi.org/10.1111/ane.12084

Cummings, J. L. (1997). The Neuropsychiatric Inventory: Assessing psychopathology in dementia patients. *Neurology, 48*(5), 10S-16S.

Dagenais, E., Rouleau, I., Demers, M., Jobin, C., Roger, E., Chamelian, L., & Duquette, P. (2013). Value of the MoCA test as a screening instrument in multiple sclerosis. *The Canadian Journal of Neurological Sciences, 40*(3), 410–415. https://doi.org/10.1017/s0317167100014384

Damian, A. M., Jacobson, S. A., Hentz, J. G., Belden, C. M., Shill, H. A., Sabbagh, M. N., Caviness, J. N., & Adler, C. H. (2011). The Montreal Cognitive Assessment and the Mini-Mental State Examination as screening instruments for cognitive impairment: Item analyses and threshold scores. *Dementia and Geriatric Cognitive Disorders*, *31*(2), 126-131. https ://doi.org/10.1159/000323867

De la Torra, G. G., Suárez-Lorens, A., Caballero, F. J., Ramallo, M. A., Randolph, C., Lleó, A., Sala, I., & Sánchez, B. (2014). Norms and reliability for the Spanish version of the Repeatable Battery for the Assessment of Neuropsychological Statut (RBANS) form A. *Journal of Clinical and Experimental Neuropsychology, 36*(10), 1023-1030. https://dx.doi.org/10.1080/13803395.2014.965664

De la Torre, G. G., Perez, M. J., Ramallo, M. A., Randolph, C., & González-Villegas, M. B. (2016). Screening of cognitive impairment in schizophrenia: Reliability, sensitivity, and specificity of the Repeatable Battery for the Assessment of Neuropsychological Status in a Spanish sample. *Assessment*, *23*(2), 221-231. https://doi.org/10.1177/1073191115583715

Decker, S. L., Allen, R., & Choca, J. P. (2006). Construct validity of the Bender-Gestalt II: Comparison with Wechsler Intelligence Scale for Children-III. *Perceptual and Motor Skills, 102*(1),133-141. https://doi.org/10.2466/pms.102.1.133-141

Deloche, G., & Hannequin, D. (1997). *DO80 : Test de dénomination orale d'images*. ECPA.

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test-retest reliability and practice effect of expanded Halstead-Reitan Neuropsychological test battery. *Journal of the International Neuropsychological Society*, *5*(4), 346-356. https://doi.org/10.1017/S1355617799544056

Drane, D. L., & Osato, S. S. (1997). Using the Neurobehavioral Cognitive Status Examination as a screening measure for older adults. *Archives of Clinical Neuropsychology, 12*(2), 137-143. https://doi.org/10.1016/S0887-6177(96)00057-1

Dubois, B., Slachevsky, A., Litvan, I., & Pillon, B. (2000). The FAB: a Frontal Assessment Battery at bedside. *Neurology, 55*(11), 1621–1626. https://doi.org/10.1212/wnl.55.11.1620

Duff, K., Clark, H. J. D., O'Bryant, S. E., Mold, J. W., Schiffer, R. B., & Sutker, P. B. (2008). Utility of the RBANS in detecting cognitive impairment associated with Alzheimer's disease: Sensitivity, specificity, and positive and negative predictive powers. *Archives of Clinical Neuropsychology*, *23*(5), 603-612. https://doi.org/10.1016/j.acn.2008.06.004

Duff, K., Hobson, V. L., Beglinger, L. J., & O'Bryant, S. E. (2010). Diagnostic accuracy of the RBANS in mild cognitive impairment: Limitations on assessing milder impairments. *Archives of Clinical Neuropsychology*, *25*(5), 429-441. https://doi.org/10.1093/arclin/acq045

Duff, K., Patton, D., Schoenberg, M. R., Mold, J., Scott, J. G., & Adams, R. L. (2003). Age- and education-corrected independent normative data for the RBANS in a community dwelling elderly sample. *The Clinical Neuropsychologist*, *17*(3), 351-366. https://doi.org/10.1076/clin.17.3.351.18082

Dujardin, K., Sockeel, P., Cabaret, M., De Sève, J., & Vermersch, P. (2003). La BCcogSEP : Une batterie courte d'évaluation des fonctions cognitives destinées aux patients souffrant de sclérose en plaques. *Revue Neurologique, 160*(1), 51-62.

Egeland, J., & Kovalik-Gran, I. (2010). Validity of the factor structure of Conner's CPT. *Journal of Attention Disorders, 13*(4), 347-357. https://doi.org/10.1177/1087054709332477

Eisenstein, N., Engelhart, C. I., Johnson, V., Wolf, J., Williamson, J., & Losonczy, M. B. (2002). Normative data for healthy elderly persons with the Neurobehavioral Cognitive Status Exam (Cognistat). *Applied Neuropsychology*, *9*(2), 110-113. https://doi.org/10.1207/S15324826AN0902_6

Engelhart, C., Eisenstein, N., Johnson, V., Wolf, J., Williamson, J., Steitz, D., Girard, V., Paramatmuni, K., Ouzounian, N., & Losonczy, M. (1999). Brief report factor Structure of the Neurobehavioral Cognitive Status Exam (COGNISTAT) in healthy, and psychiatrically and neurologically impaired, elderly adults. *The Clinical Neuropsychologist*, *13*(1), 109-111. https://doi.org/10.1076/clin.13.1.109.1975

Fernandez, A. L., & Scheffel, D. L., (2003). A study of the criterion validity of the Mattis Dementia Rating Scale. *International Journal of Testing, 3*(1), 49-58. https://doi.org/10.1207/S15327574IJT0301_3

Fong, M. W. M., Van Patten, R., & Fucetola, R. P. (2019). The factor structure of the Boston Diagnostic Aphasia Examination, third edition. *Journal of the International Neuropsychological Society, 25*, 772-776. https://doi.org/10.1017/S1355617719000237

Friedman, M. A., Schinka, J. A., Mortimer, J. A., & Borenstein Graves, A. (2002). Hopkins Verbal Learning Test-Revised: Norms for elderly African Americans. *The Clinical Neuropsychologist, 16*(3), 356-372. https://doi.org/10.1076/clin.16.3.356.13857

Funkiewiez, A., Bertoux, M., de Souza, L. C., Lévy, R., & Dubois, B. (2011). The

Giannakou, M., & Kosmidis, M. H. (2006). Cultural appropriateness of the Hooper Visual Organization Test? Greek normative data. *Journal of Clinical and Experimental Neuropsychology, 28*(6), 1023-1029, https://doi.org/10.1080/138033905910043

Giulioli, C., Meillon, C., Gonzalez-Colaço Harmand, M., Dartigues, J.-F., & Amieva, H. (2016). Normative scores for standard neuropsychological tests in the oldest old from the French population-based PAQUID study. *Archives of Clinical Neuropsychology*, *31*(1), 58-68. https://doi.org/10.1093/arclin/acv055

Godefroy, O., Moroni, C., Quaglino, V., Theunssens, É., Beaunieux, H., & Roussel, M. (2016). Données normatives. In M. Roussel, & O. Godefroy (Eds.), *La batterie GRECOGVASC : Évaluation et diagnostic des troubles neurocognitifs vasculaires avec ou sans contexte d'accident vasculaire cérébral* (pp. 221-246). DeBoeck Supérieur.

González-Palau, F., Franco, M., Jiménez, F., Parra, E., Bernate, M., & Solis, A. (2013). Clinical utility of the Hopkins Verbal Test-revised for detecting Alzheimer's disease and mild cognitive impairment in Spanish population. *Archives of Clinical Neuropsychology*, *28*(3), 245-253. https://doi.org/10.1093/arclin/act004

Greve, K. W., Etherton, J. L., Ord, J., Bianchini, K. J., & Curtis, K. L. (2009). Detecting malingered pain-related disability: Classification accuracy of the Test of Memory Malingering. *The Clinical Neuropsychologist*, *23*(7), 1250-1271. https://doi.org/10.1080/13854040902828272

Grober, E., Ocepek-Welikson, K., & Teresi, J. A. (2009). The Free and Cued Selective Reminding Test: Evidence of psychometric adequacy. *Psychology Science Quarterly*, *51*(3), 266-282.

Harvey, E. M., Leonard-Green, T. K., Mohan, K. M., Kulp, M. T., Davis, A. L., Miller, J. M., Twelker, J. D., Campus, I., & Dennis, L. K. (2017). Inter-rater and test-retest reliability of the Beery VMI in schoolchildren. *Optometry and Vision Science*, *94*(5), 598-605. https://doi.org/10.1097/OPX.0000000000001068

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. J., & Curtiss, G. (2002). *Wisconsin Card Sorting Test*. ECPA Pearson.

Herrera-Guzmán, I., Peña-Casanova, J., Lara, J. P., Gudayol-Ferré, E., & Böhm, P. (2004). Influence of age, sex and education on the Visual Object and Space Perception Battery (VOSP) in a healthy normal verbal population. *The Clinical Neuropsychologist, 18*(3), 385-394. https://doi.org/10.1080/1385404049052421

Heyanka, D. J., Thaler, N. S., Linck, J. F., Pastorek, N. J., Miller, B., Romesser, J., & Sim, A. H. (2015). A factor analytic approach to the validation of the Word Memory Test and Test of Memory Malingering as measures of effort and not memory. *Archives of Clinical Neuropsychology*, *30*(5), 369-376. https://doi.org/10.1093/arclin/acv025

Jacobs, M. L., & Donders, J. (2007). Criterion validity of the California Verbal Learning Test-(CVLT-II) after traumatic brain injury. *Archives of Clinical Neuropsychology*, *22*(2), 143-149. https://doi.org/10.1016/j.acn.2006.12.002

Jardim de Paula, J., Bertola, L., Nicolato, R., Nunes de Moraes, E., & Fernandes Malloy-Diniz, L. (2011). Evaluating language comprehension in Alzheimer's disease: The use of the Token test. *Arquivos de Neurro-Psiquiatria, 70*(60), 435-440. https://doi.org/10.1590/S0004-282X2012000600010

Kiely, K. M., Butterworth, P., Watson, N., & Wooden, M. (2014). The Symbol Digit Modalities Test: Normative data from a large nationally representative sample of Australians. *Archives of Clinical Neuropsychology*, *29*(8), 767–775. https://doi.org/10.1093/arclin/acu055

Kopecek, M., Bezdicek, O., Sulc, Z., Lukavsky, J., & Stepankova, H. (2017). Montreal Cognitive Assessment and Mini-Mental State Examination reliable change indices in healthy older adults. *International Journal of Geriatric Psychiatry*, *32*(8), 868-875. https://doi.org/10.1002/gps.4539

Korman, M., Kirk, U., & Kemp, S. (2012). *NEPSY-II – Bilan neuropsychologique de l'enfant – 2nde édition*. ECPA Pearson.

Kwok Chu Wong, G., Ngai, K., Wai Lam, S., Wong, A., Mok, V., & Sang Poon, W. (2013). Validity of the Montreal Cognitive Assessment for traumatic brain injury patients with intracranial hemorrhage. *Brain Injury*, *27*(4), 394-398. https://doi.org/10.3109/02699052.2012.750746

Lam, B., Middleton, L. E., Masellis, M., Stuss, D. T., Harry, R. D., Kiss, A., & Black, S. E. (2013). Criterion and convergent validity of the Montreal Cognitive Assessment with screening and standardized neuropsychological testing. *Journal of the American Geriatrics Society*, *61*(12), 2181-2185. https://doi.org/10.1111/jgs.12541

Lamarre, C. J., & Battern, S. B. (1994). A clinical examination of the Neurobehavioral Cognitive Status Examination in a general psychiatric inpatient population. *Journal of Psychiatric Neurosciences, 19*(2), 103-108.

Larson, E. B., Kirschner, K., Bode, R., Heinemann, A., & Goodman, R. (2005). Construct and predictive validity of the repeatable battery for the assessment of neuropsychological status in the evaluation of stroke patients. *Journal of Clinical and Experimental Neuropsychology*, *27*(1), 16-32. https://doi.org/10.1080/138033990513564

Lavoie, M., Bherer, L., Joubert, S., Gagnon, J. F., Blanchet, S., Rouleau, I., Macoir, J., & Hudon, C. (2018). Normative data for the Rey Auditory Verbal Learning Test in the older French-Quebec population. *The Clinical Neuropsychologist*, *32*(1), 15-28. https://doi.org/10.1080/13854046.2018.1429670

Lavoie, M., Callahan, B., Belleville, S., Simard, M., Bier, N., Gagnon, L., Gagnon, J.-F., Blanchet, S., Potvin, O., Hudon, C., & Macoir, J. (2013). Normative data for the Dementia Rating Scale-2 in the French-Quebec Population. The Clinical Neuropsychologist, *27*(7), 1150-1166. https://doi.org/10.1080/13854046.2013.825010

Llebaria, G., Pagonabarraga, J., Kulisevsky, J., García-Sánchez, C., Pascual-Sedano, B., Gironell, A., & Martínez-Corral, M. (2008). Cut-off score of the Mattis Dementia Rating Scale for screening dementia in Parkinson's disease. *Movement Disorders*, *23*(11), 1546-1550. https://doi.org/10.1002/mds.22173

López-Góngora, M., Querol, L., & Escartin, A. (2015). A one-year follow-up study of the Symbol Digit Modalities Test (SDMT) and the Paced Auditory Verbal Serial Addition

Test in relapsing-remitting multiple sclerosis: An appraisal of comparative longitudinal sensitivity. *BMC Neurology, 15*(1), 1-8. https://doi.org/10.1186/s12883-015-0296-2

Lopez, M. N., Lazar, M. D., & Oh, S. (2003). Psychometric properties of the Hooper Visual Organization Test. *Assessment*, *10*(1), 66-70. https://doi.org/10.1177/1073191102250183

Lucas, J. A., Ivnik, R. J., Smith, G. E. Ferman, T. J., Willis, F. B., Petersen, R. C., & Graff-Radford, N. R. (2005). Mayo's older African Americans normative studies: norms for Boston Naming Test, Controlled Oral Word Association, Category Fluency, Animal Naming, Token Test, Wrat-3 Reading, Trail Making Test, Stroop test, and Judgment of Line Orientation. *The Clinical Neuropsychologist, 19*(2), 243-269. https://doi.org/10.1080/13854040590945337

Mahieux-Laurent, F., Fabre, C., Galbrun, E., Dubrulle, A., & Moroni, C. (2009). Validation d'une batterie brève d'évaluation des praxies gestuelles pour consultation mémoire : Évaluation chez 419 témoins, 127 patients atteints de troubles cognitifs légers et 320 patients atteints d'une démence. *Revue Neurologique*, *165*(6-7), 560-567. https://doi.org/10.1016/j.neurol.2008.11.016

Malloy, P., Belanger, H., Hall, S., Aloia, M., & Salloway, S. (2003). Assessing visuoconstructional performance in AD, MCI, and normal elderly using the Beery Visual-Motor Integration test. *The Clinical Neuropsychologist, 17*(4), 544-550. https://doi.org/10.1076/clin.17.4.544.27947

Mattioli, F., Stampatori, C., Bellomi, F., Scarpazza, C., Galli, P., Guarneri, C., Corso, B., Montomoli, C., Niccolai, C., Goretti, B., Amato, M. P., Riboni, E., Dalla Tomasina, C., Falautano, M., & Capra, R. (2014). Assessing executive function with the D-KEFS sorting test: Normative data for a sample of the Italian adult population. *Neurological Sciences, 35*(12)*,* 1895–1902. https://doi.org/10.1007/s10072-014-1857-7

Mazancova, A. F., Růžička, E., Jech, R., & Bezdicek, O. (2020). Test the best: Classification accuracies of four cognitive rating scales for Parkinson's disease mild cognitive impairment. *Archives of Clinical Neuropsychology*, *35*(7), 1069-1077. https://doi.org/10.1093/arclin/acaa039

McKay, C., Wertheimer, J. C., Fichtenberg, N. L., & Casey, J. E. (2008). The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): Clinical utility in a traumatic brain injury sample. *The Clinical Neuropsychologist*, *22*(2), 228-241. https://doi.org/10.1080/13854040701260370

Merck, C., Charnallet, A., Auriacombe, S., Belliard, S., Hahn-Barma, V., Kremin, H., Lemesle, B., Mahieux, F., Moreaud, O., Perrier Palisson, D., Roussel, M., Sellal, F., & Siegwart, H. (2011). La batterie d'évaluation des connaissances sémantiques du GRECO (BECS-GRECO) : Validation et données normatives. *Revue de Neuropsychologie,3*(4), 235-255. https://doi.org/10.1684/nrp.2011.0194

Merten, T. (2005). Factor structure of the Hooper Visual Organization Test: A cross-cultural replication and extension. *Archives of Clinical Neuropsychology, 20*, 123-128. https://doi.org/10.1016/j.acn.2004.03.001

Merten, T. (2006). An analysis of the VOSP Silhouettes test with neurological patients. *Psychology Science, 48*(4), 451-462.

Messinis, L., Nasios, G., Mougias, A., Politis, A., Zampakis, P., Tsiamaki, E., Malefaki, S., Gourzis, P., & Papathanasopoulos, P. (2016). Age and education adjusted normative data and discriminative validity for Rey's Auditory Verbal Learning Test in the elderly Greek population. *Journal of Clinical and Experimental Neuropsychology*, *38*(1), 23-39. https://doi.org/10.1080/13803395.2015.1085496

Miranda, A. R., Franchetto Sierra, J., Martínez Roulet, A., Rivadero, L., Serra, S. V., & Soria, E. A. (2020). Age, education and gender effects on Wisconsin Card Sorting Test: Standardization, reliability and validity in healthy Argentinian adults. *Aging, Neuropsychology and Cognition*, *27*(6), 807–825. https://doi.org/10.1080/13825585.2019.1693491

Montani, C., Bouati, N., Pellisier, C., Couturier, P., Jasso-Mosqueda, G., Hugonot, R., & Franco, A. (1994). Cotation et validation du test du cadran de l'horloge en psychométrie chez le sujet âgé. *L'Encéphale,23*(3), 194-199.

Mougias, A. A., Christidi, F., Kiosterakis, G., Messinis, L., & Politis, A. (2018). Dealing with severe dementia in clinical practice: A validity and reliability study of Severe Mini-Mental State Examination in Greek population. *International Journal of Geriatric Psychiatry*, *33*(9), 1236-1242. https://doi.org/10.1002/gps.4915

Mullen, C. M., & Fouty, H. E. (2014). Comparison of the WRAT4 Reading subtest and the WTAR for estimating premorbid ability level. *Applied Neuropsychology: Adult, 21*(1), 69-7. https://doi.org/10.1080/09084282.2012.727111

Nakhutina, L., Pramataris, P., Morrison, C., Devinsky, O., & Barr, W. B. (2010). Reliable change indices and regression-based measures for the Rey-Osterreith Complex Figure test in partial epilepsy patients. *The Clinical Neuropsychologist*, *24*(1), 38-44. https://doi.org/10.1080/13854040902960091

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society, 53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Nolin, P. (1999). Analyses psychométriques de l'adaptation française du California Verbal Learning Test. *Revue Québécoise de Psychologie, 20*(1), 39-55.

Norris, G., & Tate, R. L. (2000). The Behavioural Assessment of the Dysexecutive Syndrome (BADS): Ecological, concurrent and construct validity. *Neuropsychological Rehabilitation, 10*(1), 33-45. https://doi.org/10.1080/096020100389282

O'Neil-Pirozzi, T. M., Goldstein, R., Strangman, G. E., & Glenn, M. B. (2012). Test–re-test reliability of the Hopkins Verbal Learning Test-Revised in individuals with traumatic brain injury. *Brain Injury*, *26*(12), 1425-1430. https://doi.org/10.3109/02699052.2012.694561

Ouvrard, C., Berr, C., Meillon, C., Ribet, C., Goldberg, M., Zins, M., & Amieva, H. (2018). Norms for standard neuropsychological tests from the French CONSTANCES cohort. *European Journal of Neurology, 26*(5), 786-793. https://doi.org/10.1111/ene.13890

Paci, M., Lorenzini, C., Fioravanti, E., Poli, C., & Lombardi, B. (2015). Reliability of the 36-item version of the Token test in patients with poststroke aphasia. *Topics in Stroke Rehabilitation*, *22*(5), 374-376. https://doi.org/10.1179/1074935714Z.0000000049

Paolo, A. M., Tröster, A. I., Axelrod, B. N., & Koller, W. C. (1995). Construct validity of the WCST in normal elderly and persons with Parkinson's disease. *Archives of Clinical Neuropsychology*, *10*(5), 463-473. https://doi.org/10.1016/0887-6177(95)00052-6

Peña-Casanova, J., Quiñones-Úbeda, S., Quintana-Aparicio, M., Aguilar, M., Badenes, D., Molinuevo, J. L., Torner, L., Robles, A., Sagrario Barquero, M., Villanueva, C., Antúnez, C., Martínez-Parra, C., Frank-García, A., Sanz, A., Fernández, M., Alfonso, V., Sol, J. M., & Blesa, R. (2009). Spanish multicenter normative studies (NEURONORMA Project): Norms for verbal span, visuospatial span, letter and number sequencing, Trail Making Test, and Symbol Sigit Modalities Test. *Archives of Clinical Neuropsychology*, *24*(4), 321-341. https://doi.org/10.1093/arclin/acp038

Pereira, D. R., Costa, P., & Cerqueira, J. J. (2015). Repeated assessment and practice effects of the written symbol digit modalities test using a short inter-test interval. *Archives of Clinical Neuropsychology*, *30*(5), 424-434. https://doi.org/10.1093/arclin/acv028

Periáñez, J. A., Lubrini, G., García-Gutiérrez, A., & Ríos-Lago, M. (2021). Construct Validity of the Stroop Color-Word Test: Influence of speed of visual search, verbal fluency, working memory, cognitive flexibility, and conflict monitoring. *Archives of Clinical Neuropsychology, 36*(1), 99–111. https://doi.org/10.1093/arclin/acaa034

Pettersson, R., Söderström, S., & Nilsson, K. W. (2018). Diagnosing ADHD in adults: An examination of the discriminative validity of neuropsychological tests and diagnostic assessment instruments. *Journal of Attention Disorders, 22*(11), 1019-1031. https://doi.org/10.1177/1087054715618788

Poitrenaud, J., Deweer, B., Kalafat, M., & Van der Linden, M. (2007). *CVLT (California Verbal Learning Test) : Test d'apprentissage et de mémoire verbale.* ECPA Pearson.

Quental, N. B. M., Brucki, S. M. D., & Bueno, O. F. A. (2013). Visuospatial function in early Alzheimer's disease: The use of the Visual Object and Space Perception (VOSP) battery. *PloS One*, *8*(7), e68398. https://doi.org/10.1371/journal.pone.0068398

Quintana, M. Peña-Casanova, J., Sánchez-Benavides, G, Langohr, K., Manero, R. M., Aguilar, M., Badenes, D., Molinuevo, J. L., Robles, A., Sagrario Barquero, M., Antúnez, C., Martínez-Parra, C., Frank-García, A., Fernández, M., & Blesa, R. (2011). Spanish multicenter normative studies (neuronorma project): Norms for the abbreviated Barcelona test. *Archives of Clinical Neuropsychology, 26*, 144-157. https://doi.org/10.1093/arclin/acq098

Radanovic, M., Mansur, L. L., & Scaff, M. (2004). Normative data for the Brazilian population in the Boston Diagnostic Aphasia Examination: Influence of schooling. *Brazilian Journal of Medical and Biological Research, 37*(11), 1371-1378. https://doi.org/10.1590/20100-879X2004001100019

Rapport, L. J., Millis, R., & Bonello, P. J. (1998). Validation of the Warrington theory of visual processing and the Visual Object and Space Perception battery. *Journal of Clinical and Experimental Neuropsychology, 20*(2), 211-220. https://doi.org/10.1076/jcen.20.2.211.1169

Reuter, F., Baumstarck-Barrau, K., Loundou, A., Pelletier, J., & Auquier, P. (2010). Paced Auditory Serial Addition Test : données normatives dans une population française. *Revue Neurologique*, *166*(11), 944-947. https://doi.org/10.1016/j.neurol.2010.01.018

Rivera, D., Perrin, P. B., Weiler, G., Ocampo-Barba, N., Aliaga, A., Rodríguez, W., Rodríguez-Agudelo, Y., Aguayo, A., Longoni, M., Trapp, S., Esenarro, L., & Arango-Lasprilla, J. C. (2015). Test of Memory Malingering (TOMM): Normative data for the Latin American Spanish speaking adult population. *NeuroRehabilitation*, *37*(4), 719-735. https://doi.org/10.3233/NRE-151287

Ross, S. A., Allen, D. N., & Goldstein, G. (2013). Factor structure of the Halstead-Reitan neuropsychological battery: A review and integration. *Applied Neuropsychology : Adult*, *20*(2), 120-135. https://doi.org/10.1080/09084282.2012.690798

Ross, T. P. (2003). The reliability of cluster and switch scores for the Controlled Oral Word Association Test. *Archives of Clinical Neuropsychology*, *18*(2), 153-164. https://doi.org/10.1093/arclin/18.2.153

Roussel, M., & Godefroy, O. (2008). La batterie GREFEX : Données normatives. In O. Godefroy, & GREFEX (Eds.), *Fonctions exécutives et pathologies neurologiques et psychiatriques* (pp. 231-252). Solal.

Ruff, R. M., Light, R. H., Parker, S. B., & Levin, H. S. (1996). Benton Controlled Oral Word Association Test: Reliability and updated norms. *Archives of Clinical Neuropsychology*, *11*(4), 329-338. https://doi.org/10.1016/0887-6177(95)00033-X

Rullier, L., Matharan, F., Barbeau, E. J., Mokri, H., Dartigues, J. F., Pérès, K., & Amieva, H. (2014). Test du DMS 48 : Normes chez les sujets âgés et propriétés de détection de la maladie d'Alzheimer dans la cohorte AMI. *Gériatrie et Psychologie Neuropsychiatrie du Vieillissement, 12*(3), 321-330. https://doi.org/10.1684/pnv.2014.0486

Sachs, B. C., Lucas, J. A., Smith, G. E., Ivnik, R. J., Petersen, R. C., Graff-Radford, N. R., & Pedraza, O. (2012). Reliable change on the Boston Naming Test. *Journal of the International Neuropsychological Society*, *18*(2), 375-378. https://doi.org/10.1017/S1355617711001810

Sánchez-Cubillo, I., Periáñez, J. A., Adrover-Roig, D., Rodríguez-Sánchez, J. M., Ríos-Lago, M., Tirapu, J., & Barceló, F. (2009). Construct validity of the Trail Making Test: Role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *Journal of the International Neuropsychological Society*, *15*(3), 438–450. https://doi.org/10.1017/S1355617709090626

Schmidt, K. S., Mattis, P. J., Adams, J., & Nestor, P. (2005). Alternate-form reliability of the Dementia Rating Scale-2. *Archives of Clinical Neuropsychology*, *20*(4), 435-441. https://doi.org/10.1016/j.acn.2004.09.011

SEA (Social Cognition and Emotional Assessment): A clinical neuropsychological tool for early diagnosis of frontal variant of frontotemporal lobar degeneration. *Neuropsychology, 26*(1), 81. https://doi.org/10.1037/a0025318

Shapiro, A. M., Benedict, R. H. B., Schretlen, D., & Brandt, J. (1999). Construct and concurrent validity of the Hopkins Verbal Learning Test – Revised. *The Clinical Neuropsychologist, 13*(3) 348-358. https://doi.org/10.1076/clin.13.3.348.1749

Sherman, E. M., Strauss, E., & Spellacy, F. (1997). Validity of the Paced Auditory Serial Addition Test (PASAT) in adults referred for neuropsychological assessment after head injury. *The Clinical Neuropsychologist*, *11*(1), 34-45. https://doi.org/10.1080/13854049708407027

Shunk, A. W., Davis, A. S., & Dean, R. S. (2006). Test review: Delis-Kaplan Executive Function System (D-KEFS). *Applied Neuropsychology, 13*(4), 275-379. https://doi.org/10.1207/s15324826an1304_9

Slachevsky, A., Villalpando, J. M., Sarazin, M., Hahn-Barma, V., Pillon, B., & Dubois, B. (2004). Frontal Assessment Battery and differential diagnosis of frontotemporal dementia and Alzheimer disease. *Archives of Neurology, 61*(7), 1104-1107. https://doi.org/10.1001/archneur.61.7.1104

Smith, S. R., Chang, J., Schnoebelen, K. J., Edwards, J. W., Servesko, A. M., & Walker, S. J. (2007). Psychometrics of a simple method for scoring organizational approach on the Rey-Osterrieth complex figure. *Journal of Neuropsychology*, *1*(1), 39-51. https://doi.org/10.1348/174866407X180800

Soble, J. R., Rhoads, T., Carter, D. A., Bernstein, M. T., Ovsiew, G. P., & Resch, Z. J. (2020). Out of sight, out of mind: The impact of material-specific memory impairment on Rey 15-item test performance. *American Psychological Association, 32*(11), 1087-1093. http://dx.doi.org/10.1037/pas0000854

Sonder, J. M., Burggraaff, J., Knol, D. L., Polman, C. H., & Uitdehaag, B. M. J. (2014). Comparing long-term results of PASAT and SDMT scores in relation to neuropsychological testing in multiple sclerosis. *Multiple Sclerosis Journal, 20*(4), 481-488. https://doi.org/10.1177/1352458513501570

Soreni, N., Crosbie, J., Ickowicz, A., & Schachar, R. (2009). Stop signal and Conners' Continuous Performance Tasks: Test—retest reliability of two inhibition measures in ADHD children. *Journal of Attention Disorders*, *13*(2), 137-143. https://doi.org/10.1177/1087054708326110

Sotaniemi, M., Pulliainen, V., Hokkanen, L., Pirttilä, T., Hallikainen, I., Soininen, H., & Hänninen, T. (2012). CERAD-neuropsychological battery in screening mild Alzheimer's disease. *Acta Neurologica Scandinavica*, *125*(1), 16-23. https://doi.org/10.1111/j.1600-0404.2010.01459.x

Soukup, V. M., Bimbela, A., & Schiess, M. C. (1999). Recognition memory for faces: Reliability and validity of the Warrington Recognition Memory Test (WRMT) in a neurological sample. *Journal of Clinical Psychology in Medical Settings, 6*(3), 287-293. https://doi.org/10.1023/A:1026243822356

Springate, B. A., Tremont, G., Papandonatos, G., & Ott, B. R. (2014). Screening for Mild Cognitive Impairment using the Dementia Rating Scale-2. *Journal of Geriatric Psychiatry and Neurology, 27*(2), 139-144. https://doi.org/10.1177/0891988714522700

Steinberg, B. A., Bieliauskas, L. A., Smith, G. E., & Ivnik, R. J. (2005) Mayo's older Americans normative studies: Age- and IQ-adjusted norms for the Trail-Making Test, the Stroop Test, and MAE Controlled Oral Word Association Test. *The Clinical Neuropsychologist, 19*(3-4), 329-377. https://doi.org/10.1080/13854040590945210

Steinberg, B. A., Bieliauskas, L. A., Smith, G. E., Langellotti, C., & Ivnik, R. J. (2005). Mayo's older Americans normative studies: Age- and IQ-adjusted norms for the Boston Naming Test, the MAE Token Test, and the Judgment of Line Orientation Test. The Clinical Neuropsychologist, *19*(3-4), 280-328. https://doi.org/10.1080/13854040590945229

Strauss, M. E., & Fritsch, T. (2004). Factor structure of the CERAD neuropsychological battery. *Journal of the International Neuropsychological Society*, *10*(4), 559-565. https://doi.org/10.10170S1355617704104098

Strong, C., Tiesma, D., & Donders, J. (2010). Criterion validity of the Delis-Kaplan Executive Function System (D-KEFS) fluency subtests after traumatic brain injury. *Journal of the International Neuropsychological Society, 17*(2), 230-237. https://doi.org/10.1017/S1355617710001451

Sugarman, M. A., & Axelrod, B. N. (2015). Embedded measures of performance validity using verbal fluency tests in a clinical sample. *Applied Neuropsychology: Adult, 22*(2), 141-146. https://doi.org/10.1080/23279095.2013.873439

Sutton, G. P., Barchard, K. A., Bello, D. T., Thaler, N. S., & Ringdahl, E. (2011). Beery-Buktenica development test of visual-motor integration performance in children with traumatic brain injury and attention-deficit/hyperactivity disorder. *American Psychological Association, 23*(3), 805-809. https://doi.org/10.1037/a0023370

Sylvestre, G., Chopard, G., Tio, G., Magnin, E., Rumbach, L., Vandel, P., & Galmiche, J. (2011). Normes diagnostiques de la batterie de tests neuropsychologiques RAPID pour les sujets âgés de 60 à 89 ans présentant une maladie d'Alzheimer. *Revue Neurologique, 167*(6-7), 495–504. https://doi.org/10.1016/j.neurol.2010.12.004

Tallberg, I. M. (2005). The Boston naming test in Swedish: Normative data. *Brain and Language*, *94*(1), 19-31. https://doi.org/10.1016/j.bandl.2004.11.004

Thielen, H., Verleysen, G., Huybrechts, S., Lafosse, C., & Gillebert, C. R. (2019). Flemish normative data for the Buschke Selective Reminding Test. *Psychologica Belgica*, *59*(1), 58-77. https://doi.org/10.5334/pb.486

Torrence, N. D., John, S. E., Gavett, B. E., & O'Bryant, S. E. (2016). An empirical comparison of competing factor structures for the repeatable battery for the assessment of neuropsychological status: A project FRONTIER study. *Archives of Clinical Neuropsychology*, *31*(1), 88-96. https://doi.org/10.1093/arclin/acv057

Trakoshis, S., Ioannou, M., & Fanti, K. (2020). The factorial structure of the tower test from the Delis–Kaplan Executive Function System: A confirmatory factor analysis study. *Assessment*, 1-15. https://doi.org/10.1177/1073191120960812

Turcotte, V., Gagnon, M.-E., Joubert, S., Rouleau, I., Gagnon, J.-F., Escudier, F., Koski, L., Potvin, O., Macoir, J., & Hudon, C. (2018). Normative data for the Clock Drawing test for French-Quebec mid- and older aged healthy adults. *The Clinical Neuropsychologist, 32*(1), 91-101. https://doi.org/10.1080/13854046.2018.1473495

Van der Elst, W., Van Boxtel, M. P., Van Breukelen, G. J., & Jolles, J. (2008). Detecting the significance of changes in performance on the Stroop Color-Word Test, Rey's Verbal Learning Test, and the Letter Digit Substitution Test: The regression-based change approach. *Journal of the International Neuropsychological Society*, *14*(1), 71-80. https://doi.org/10.10170S1355617708080028

Van der Linden, M., Coyette, F., Poitrenaud, J., Kalafat, M., Calacis, F., Wyns, C., Adam, S., & GREMEM. (2004). L'épreuve de rappel libre/rappel indicé à 16 items (RL/RI-16). In M. Van der Liden, S. Adam, A. Agniel, & C. Baisset Mouly (Eds.), *L'évaluation des troubles de la mémoire : Présentation de quatre tests de mémoire épisodique (avec leur étalonnage)* (pp. 25-47). Solal.

Van Schependom, J., D'hooghe, M. B., Cleynhens, K., D'hooge, M., Haelewyck, M. C., De Keyser, J., & Nagels, G. (2014). The Symbol Digit Modalities Test as sentinel test for cognitive impairment in multiple sclerosis. *European Journal of Neurology*, *21*(9), 1219-1225. https://doi.org/10.1111/ene.12463

Volker, M. A., Lopata, C., Vujnovic, R. K., Smerbeck, A. M., Toomey, J. A., Rodgers, J. D., Schiavo, A., & Thomeer, M. L. (2010). Comparison of the Bender Gestalt-II and VMI-V in samples of typical children and children with high-functioning autism spectrum disorders. *Journal of Psychoeducational Assessment*, *28*(3), 187-200. https://doi.org/10.1177/0734282909348216

Wallon, P., & Mesmin, C. (2009). *Test de la figure complexe de Rey*. ECPA Pearson.

Webber, T. A., Bailey, K. C., Alverson, W. A., Critchfield, E. A., Bain, K. M., Messerly, J. M., O'Rourke, J. J. F., Kirton, J. W., Fullen, C., Marceaux, J. C., & Soble, J. R. (2018). Further validation of the Test of Memory Malingering (TOMM) trial 1 performance validity index: Examination of false positives and convergent validity. *Psychological Injury and Law*, *11*(4), 325-335. https://doi.org/10.1007/s12207-018-9335-9

Wechsler, D. (2011). *WAIS-IV – Échelle d'intelligence de Wechsler pour adultes – 4ème édition.* ECPA Pearson.

Wechsler, D. (2012). *MEM-IV – Échelle clinique de mémoire de Wechsler – 4ème édition.* ECPA Pearson.

Wechsler, D. (2014). *WPPSI-IV – Échelle d'intelligence de Wechsler pour enfants – 4ème édition.* ECPA Pearson.

Wechsler, D. (2016). *WISC-V – Échelle d'intelligence de Wechsler pour enfants et adolescents – 5ème édition.* ECPA Pearson.

Williams, J. M. (1999). *Memory assessment scales*. Pyschological Assessment Resources.

Wilson, B. A., Alderman, N., Burgess, P. W., Emslie, H., & Evans, J. J. (1996). *Behavioural Assessment of the Dysexecutive Syndrome.* Pearson.

Wilson, B. A., Evans, J. J., Emslie, H., Alderman, N., & Burgess, P. (1998). The development of an ecologically valid test for assessing patients with a dysexecutive syndrome. *Neuropsychological Rehabilitation, 8*(3), 213-228. https://doi.org/10.1080/713755570

Wilson, B. A., Greenfield, E., Baddeley, A., Cockburn, J., Watson, P., & Tate, R. (2010). *The Rivermead Behavioral Memory Test – 3ème edition.* ECPA Pearson.

Yang, Z., Rashid, N. A. A., Quek, Y. F., Lam, M., See, Y. M., Maniam, Y., Dauwels, J., Tan, B. L., & Lee, J. (2018). Montreal Cognitive Assessment as a screening instrument for cognitive impairments in schizophrenia. *Schizophrenia Research*, *199*, 58-63. https://doi.org/10.1016/j.schres.2018.03.008

VII. Annexes

Tableau A1. *Données normatives par outil (ordre décroissant de pourcentage d'utilisation).*

	Outil	*N*	Âge	Sous-groupes	Présentation des normes	Pays	Références
1	**Test de Stroop**	718	< 40 à ≥ 60 ans	3 niveaux scolaires	percentile, moyenne et écart-type	France	Roussel et Godefroy, 2008
2	**TMT**	51 879	45 à 70 ans	3 niveaux scolaires	percentile	France	Ouvrard et al., 2018
		335	20 à 49 ans	2 niveaux scolaires	percentile	France	Binetruy et al., 2018
3	**RL/RI16**	51 879	45 à 70 ans	3 niveaux scolaires	percentile	France	Ouvrard et al., 2018
		483	16 à 100 ans	3 niveaux socio-culturels	équation de régression	France	Van der Linden et al., 2004
4	**WAIS-IV**	876	16 à 79 ans		notes standards et percentile	France	Wechsler, 2011
5	**d2-r**	1365	9 à > 55 ans	7 niveaux scolaires	moyenne et écart-type	France	Brickenkamp et al., 2015
6	**BREF**	42			moyenne et écart-type	France	Dubois et al., 2000
7	**DO80**	108		2 niveaux scolaires	moyenne et écart-type	France	Deloche et Hannequin, 1997
8	**VOSP**	90	50 à 80 ans	3 niveaux scolaires	moyenne et écart-type	Espagne	Herrera-Guzmán et al., 2004
9	**WCST**	718	< 40 ans à ≥ 60 ans	3 niveaux scolaires	percentile, moyenne et écart-type	France	Roussel et Godefroy, 2008
10	**MEM-IV**	408	16 à 90 ans		notes standards et percentile	France	Wechsler, 2011
11	**CVLT**	337	20 à 89 ans	niveau scolaire, genre	équation de régression	France	Poitreneaud et al., 2007
12	**DMS48**	750	65 à ≥ 75 ans	2 niveaux scolaires, genre	percentile	France	Rullier et al., 2014
13	**MMSE**	51 879	45 à 70 ans	3 niveaux scolaires	percentile	France	Ouvrard et al., 2018
		335	20 à 49 ans	2 niveaux scolaires	percentile	France	Binetruy et al., 2018
		285	≥ 85 ans	2 niveaux scolaires	moyenne	France	Giulioli et al., 2016
14	**Test des portes**	1 003	40 à 85 ans	3 niveaux scolaires	moyenne et écart-type	France	Godefroy et al., 2016
15	**Test de l'horloge**	593	43 à 93 ans	2 niveaux scolaires	percentile	Québec	Turcotte et al., 2018
16	**WISC-V**	1012	6 à 16 ans		notes standards et percentile	France	Wechsler, 2016
17	**TEA / TAP**	17 à 811 selon les subtests	6 à 90 selon les subtests	niveau scolaire, genre	moyenne, écart-type et percentile	Europe	Zimmermann et Fimm, 2010

18	**MoCA**	1 003	40 à 85 ans	3 niveaux scolaires	moyenne et écart-type	France	Godefroy et al., 2016
19	**NEPSY-II** *Subtests Catégorisation, Copie de figures, Mémoire de figures, Interférence de liste de mots, mémoire narrative, Processus phonologiques, Reconnaissances d'affects, Théorie de l'esprit ** Autres subtests	185*	5 à 16 ans		notes standards et percentile	France	Korkman et al., 2012
		1 200**	5 à 16 ans		notes standards et percentile	États-Unis	Korkman et al., 2012
20	**Test de Hayling**	426	20 à 87 ans	niveau scolaire	équation de régression	France	Bayard et al., 2017
21	**BECS-GRECO**	164	20 à > 75 ans	2 niveaux scolaires	moyenne et écart-type	France	Merck et al., 2011
22	**MDRS**	470	55 à 85 ans	2 niveaux scolaires	percentile	Québec	Lavoie et al., 2013
23	**PASAT**	360	18 à 60 ans	3 niveaux scolaires	moyenne et écart-type	France	Reuter et al., 2010
24	**WPPSI-IV**	1 005	2,6 mois à 7,7 ans		notes standards et percentile	France	Wechsler, 2014
25	**Mini-SEA**	150	< 46 à > 65 ans	3 niveaux scolaires	moyenne et écart-type	France	Bertoux et al., 2020
26	**NPI**						
27	**D-KEFS** (*Subtests TMT et Fluence verbale normés en France)	181	20 à 69 ans		moyenne, écart-type et percentile	Italie	Mattioli et al., 2014 (*Binetruy et al., 2018 ; Godefroy et al., 2016 ; Ouvrard et al., 2018)
28	**BADS**	216	16 à 87 ans		note étalonnée et percentile cumulé	Royaume-Uni	Wilson et al., 1996
29	**Figure complexe de Rey**	1 003	40 à 85 ans	3 niveaux scolaires	moyenne et écart-type	France	Godefroy et al., 2016
		1 700			moyenne, écart-type et percentile	France	Wallon et Mesmin, 2009
30	**Token test**	348	50 à > 84 ans		percentile	Espagne	Penã-Casanova, et al., 2009
31	**SDMT**	14 456	15 à > 85 ans	4 niveaux d'éducation, genre	moyenne, écart-type et percentile	Australie	Kiely et al., 2014
32	**RBMT**	274	16 à 101 ans		note étalonnée et percentile	France	Wilson et al., 2010

33	**BCCOGSEP** (*Subtests Fluence verbale et PASAT normés en France)	93 contrôles et 51 patients SEP		années de scolarité, genre	équation de régression	France	Dujardin et al., 2003 (*Godefroy et al., 2016 ; Reuter et al., 2010)
34	**BREP**	419		3 niveaux scolaires	médiane et 5ème percentile	France	Mahieux-Laurent et al., 2009
35	**HVOT**	206	18 à 79 ans	3 niveaux d'éducation	moyenne, écart-type et percentile	Grèce	Giannakou et Kosmidis, 2006
36	**JLO**	216	56 à 81 ans		percentile	États-Unis	Steinberg, Bielauskas, Smith, Langelloti et al., 2005
37	**BNT**	1 003	40 à 85 ans	3 niveaux scolaires	moyenne et écart-type	France	Godefroy et al., 2016
38	**BSRT**	3 257	18 à > 70 ans	4 niveaux d'éducation, genre	percentile	Belgique néerlandophone	Thielen et al., 2019
39	**BDAE**	107	15 à 84 ans	2 niveaux scolaires	moyenne, écart-type et percentile	Brésil	Radanovic et al., 2004
40	**FTT**	783	40 à 85 ans	3 niveaux scolaires	moyenne et écart-type	France	Godefroy et al., 2016
41	**CPT**						
42	**SEA**						
43	**HVLT**	237	60 à 84 ans	3 niveaux d'éducation, genre	percentile	États-Unis	Friedman et al., 2002
44	**RAVLT**	432	55 à 93 ans	niveau d'éducation, genre	moyenne, écart-type et équation de régression	Québec	Lavoie et al., 2018
45	**COWAT**	777	55 à > 95 ans-	niveau d'éducation	note étalonnée et percentile	États-Unis	Steinberg, Bielauskas, Smith, et Ivnick, 2005
46	**RBANS**	718	65 à 94 ans	3 niveaux d'éducation	note étalonnée et percentile	États-Unis	Duff et al., 2003
47	**VMI**	2 758	2 à 100 ans			États-Unis	Beery et Beery, 2010
48	**WRMT**						
49	**Test de Barcelone** (*Subtest Fluence verbale normé en France)	346		années d'éducation, genre	équation de régression et percentile	Espagne	Quintana et al., 2011 (*Godefroy et al., 2016)
50	**NCSE / Cognistat**	134	60 à 85 ans		moyenne et écart-type	États-Unis	Eisenstein et al., 2002
51	**LNNB**						
52	**MAS**	843	18 à 90 ans		moyenne et écart-type	États-Unis	Williams, 1999

53	**TOMM**	2 266	18 à 95 ans	2 niveaux d'éducation	percentile	Amérique latine	Rivera et al., 2015
54	**CERAD** (*Subtests Fluence verbale normé en France et Horloge normé au Québec)	196		niveau d'éducation, genre	équation de régression	États-Unis	Beeri et al., 2006 (*Godefroy et al., 2016 ; Turcotte et al., 2018)
55	**HRNB** (*Subtests FTT et TMT normés en France)						(*Godefroy et al., 2016 ; Ouvrard et al., 2018)
56	**BVMGT**	4 014	6 à 18 ans		moyenne et écart-type	États-Unis	Bolen, 2003
57	**WRAT**	309	56 à 94 ans		note étalonnée et percentile	États-Unis	Lucas et al., 2005

Note. Si aucune information n'est mentionnée, cela signifie que nous n'avons pas trouvé d'étude correspondante pour l'outil concerné.

Tableau A2. *Validité discriminante pour chaque outil (ordre décroissant de pourcentage d'utilisation).*

	Outil	*N* contrôle	*N* clinique	Population clinique	Sensibilité	Spécificité	Subtest / Sous-tâche	Pays	Référence
1	**Test de Stroop**								
2	**TMT**	352	676	Alzheimer	.73	.76	Partie B	France	Sylvestre et al., 2011
3	**RL/RI16**	1 464		Démence	.73	.89	Rappel libre immédiat	France	Auriacombe et al., 2010
4	**WAIS-IV**								
5	**d2-r**								
6	**BREF**		64 26	Alzheimer Démence fronto-temporale (DFT)	.81	.72	Score total	France	Slachevsky et al., 2004
7	**DO80**								
8	**VOSP**	44	31	Alzheimer	.73	.69	Subtest Silhouette	Brésil	Quental et al., 2013
9	**WCST**								
10	**MEM-IV**								
11	**CVLT**	100	100	Trauma crânien	.74	.63	Rappel total	États-Unis	Jacobs et Donders, 2007
12	**DMS48**	716	34	Alzheimer	.71	.79	Reconnaissance immédiate	France	Rullier et al., 2014
13	**MMSE**	352	676	Alzheimer	.89	.82	Rappel	France	Sylvestre et al., 2011
			60	Accident cardio-vasculaire (AVC)				Australie	Cumming et al., 2013
		89	46	Atteinte cognitive				États-Unis	Damian et al., 2011
		66	141	Parkinson				République Tchèque	Mazancova, 2020
14	**Test des portes**								
15	**Test de l'horloge**								
16	**WISC-V**								
17	**TEA / TAP**								

18	**MoCA**		94	Mild Cognitive Impairment (MCI)	.90	.87	Score total		
		90	93	Alzheimer	1.00	.87	Score total	Québec	Nasreddine et al., 2005
		66	141	Parkinson	.84	.66	Score total	République Tchèque	Mazancova, 2020
		40	48	Trauma crânien	.75	.67	Score total	Chine	Kwok Chu Wong et al., 2013
			64	Schizophrène	.78	.77	Score total	Singapour	Yang et al., 2018
		23	20	Huntington	.94	.84	Score total	République Tchèque	Bezdicek et al., 2013
			60	Accident cardio-vasculaire (AVC)	.92	.67	Score total	Australie	Cumming et al., 2013
		89	46	Atteinte cognitive	.87	.75	Score total	États-Unis	Damian et al., 2011
19	**NEPSY-II**								
20	**Test de Hayling**								
21	**BECS-GRECO**			Démence sémantique					
		164	25 11	Alzheimer	.88	.96	Ensemble des subtests (moyenne)	France	Merck et al., 2011
22	**MDRS**		98	MCI					
		50	46	Alzheimer léger	.81	.86	Score total	États-Unis	Springate et al., 2014
			57	Parkinson non dément					
			35	Parkinson dément				Espagne	Llebaria et al., 2008
23	**PASAT**			Sclérose en plaque					López-Góngora et al.,
		57	237	(SEP)	.78	.64	Score total	Espagne	2015
		48	60	Trouble Déficitaire de l'Attention avec Hyperactivité (TDAH)	.33	.77		Suède	Pettersson et al., 2018
24	**WPPSI-IV**								
25	**Mini-SEA**		37	DFT	.89	1.00			
		30	19	Dépression majeure	.94	1.00	Score total	France	Bertoux et al., 2012
26	**NPI**								
27	**D-KEFS** (*TMT validé en France)	352	676	Trauma crânien	.66	.65	Fluence verbale et Color-word interference test	États-Unis	Strong et al., 2010 (*Auriacombe et al., 2010)

28	**BADS**								
29	**Figure complexe de Rey**								
30	**Token test**	80	80	Alzheimer	.72	.64	Score total	Brésil	Jardim de Paula et al., 2011
31	**SDMT**		359	SEP	.91	.60	Score total	Belgique néerlandophone	Van Schependom et al., 2014
32	**RBMT**								
33	**BCCOGSEP** (*Subtest PASAT validé individuellement)								(*López-Góngora et al., 2015 ; Pettersson et al., 2018)
34	**BREP**	419	447	Démence	.50	.98	Ensemble des subtests (moyenne)	France	Mahieux-Laurent et al., 2009
35	**HVOT**								
36	**JLO**								
37	**BNT**								
38	**BSRT**								
39	**BDAE**								
40	**FTT**	264		Validité de performance	.40	.90	Score total	États-Unis	Axelrod et al., 2014
41	**CPT**	48	60	TDA/H	.33	.92	Score total	Suède	Pettersson et al., 2018
42	**SEA**	30	37 19	DFT Dépression majeure	1.00	1.00	Score total	France	Bertoux et al., 2012
		30	22 22	DFT MCI	.86	.88	Score total	France	Funkiewiez et al., 2011
43	**HVLT**	109	54 132	Alzheimer MCI	.85	.88	Score total	Espagne	González-Palau et al., 2013
44	**RAVLT**	258	65 192	Alzheimer MCI	.95	.81	Rappel différé	Grèce	Messinis et al., 2016
45	**COWAT**	969		Validité de performance	.36	.89	Score total	États-Unis	Sugarman et Axelrod, 2015
46	**RBANS**	69	69	Alzheimer	.56	.85	Ensemble des subtests (moyenne)	États-Unis	Duff et al., 2008

		88	88	Schizophrène	.87	.86	Score total	Espagne	De la Torre et al., 2016
		34	51	Trauma crânien	.82	.94	Score total	États-Unis	McKay et al., 2008
		71	72	MCI	.36	.80	Ensemble des subtests (moyenne)	États-Unis	Duff et al., 2010
47	**VMI**	43	40 43	Alzheimer MCI	.83	.80	Score total	États-Unis	Malloy et al., 2003
48	**WRMT**								
49	**Test de Barcelone**								
50	**NCSE / Cognistat**		72	Troubles psychiatriques	.83	.47	Score total	États-Unis	Lamarre et Battern, 1994
		20	21	Démence	1.00	.70	Score total	États-Unis	Drane & Osate, 1997
51	**LNNB**								
52	**MAS**								
53	**TOMM**		604	Douleur chronique	.35	1.00	Score total	États-Unis	Greve et al., 2009
54	**CERAD**	315	171	Alzheimer	.76	.76	Ensemble des subtests (moyenne)	Finlande	Sotaniemi et al., 2012
55	**HRNB** (*Subtests FTT validé aux États-Unis)								(*Axelrod et al., 2014)
56	**BVMGT**								
57	**WRAT**								

Note. Si aucune information n'est mentionnée, cela signifie que nous n'avons pas trouvé d'étude correspondante pour l'outil concerné.

Tableau A3. *Autres formes de validité et fidélité inter-juges pour chaque outil (ordre décroissant de pourcentage d'utilisation).*

	Outil	Validité de construit	Validité convergente	Validité divergente	Fidélité inter-juges
1	**Test de Stroop**		Mémoire des chiffres WAIS-IV, Fluence phonémique, TMT (Periáñez et al., 2021) (Espagne)		
2	**TMT**		FTT, Empan de chiffres WAIS-III, Test de Stroop (Sánchez-Cubillo et al., 2009) (Espagne)		*r* = .93 à .99 (Cangoz et al., 2009) (Turquie)
3	**RL/RI16**	Grober et al., 2009 (États-Unis)	Fluence sémantique, Fluence phonémique, Figure complexe de Rey, Test de l'horloge, Matrices de Raven, TMT, BREF, Test de Stroop (Clerici et al., 2017) (Italie)		
4	**WAIS-IV**	Wechsler, 2011 (États-Unis)	Wechsler Memory Scale-III, Échelle de mémoire pour enfants CMS, D-KEFS, CVLT (Wechsler, 2011) (États-Unis)		Subtests Similitudes, Vocabulaire, Information et Compréhension : *r* = .91 à .99 (Wechsler, 2011) (France)
5	**d2-r**		Test des dominos D2000, Test des labyrinthes Laby 5-12 (Brickenkamp et al., 2015) (France)	Brief Big 5 (Brickenkamp et al., 2015) (France)	
6	**BREF**		WCST (Dubois et al., 2000) (France)		Kappa de Cohen (κ) = .87 (Dubois et al., 2000) (France)
7	**DO80**				
8	**VOSP**	Rapport et al., 1998 (États-Unis)	Subtest Silhouettes : Cubes WAIS-R, SKT test battery (Merten, 2006) (Allemagne)		
9	**WCST**	Paolo et al., 1995 (États-Unis)	MMSE (Miranda et al., 2020) (Argentine)		Corrélation intraclasse (ICC) = .88 à .96 (Heaton et al., 2002) (États-Unis)
10	**MEM-IV**	Wechsler, 2012 (France)	TMT, Fluences verbales, WISC-IV (Wechsler, 2012) (États-Unis) WAIS-IV (Wechsler, 2011) (France)		Subtests Mémoire logique et Reproduction visuelle : Pourcentage d'accord = 97% (Wechsler, 2009) (France)

11	**CVLT**	Nolin, 1999 (Québec)		
12	**DMS48**			
13	**MMSE**		RL/RI16 (Carcaillon et al., 2009) (France)	Corrélation intraclasse (ICC) = .99 (Mougias et al., 2018) (Grèce)
14	**Test des portes**			
15	**Test de l'horloge**	Montani et al., 1994 (France)		Corrélation intraclasse (ICC) = .84 (Turcotte et al., 2018) (Québec)
16	**WISC-V**	Wechsler, 2016 (États-Unis)	Vineland-II (Wechsler, 2016) (États-Unis) KABC-II (Wechsler, 2016) (France)	*r* = .96 à .99 (Wechsler, 2016) (France)
17	**TEA / TAP**	Zimmermann & Fimm, 2010 (Europe)		
18	**MoCA**		RL/RI16, Figure complexe de Rey, Test de Stroop, COWAT, SDMT (Bezdicek et al., 2013) (République tchèque)	
19	**NEPSY-II**		NEPSY, WISC-IV, WNV, CMS (Korkman et al., 2012) (États-Unis)	Pourcentage d'accord = 93 à 99% (Korkman et al., 2012) (États-Unis)
20	**Test de Hayling**			Corrélation intraclasse (ICC) = .82 (Bayard et al., 2017) (France)
21	**BECS-GRECO**			
22	**MDRS**			

23	**PASAT**	Crawford et al., 1998 (Royaume-Uni)	Arithmétique WAIS-R, Mémoire des chiffres WAIS-R, Répétition de phrases, Trigrammes, Blocs de Corsi, Test de barrage, Temps de réaction, Symboles WAIS-R, TMT, WCST, Stroop (Sherman et al., 1997) (Canada)	QI WAIS-R, GATB General learning, Matrices de Raven, Wonderlic personnel test, Compétences mathématiques, Vocabulaire WAIS-R, BNT, GATB Aptitude verbale, Fluences verbales, Achèvement académique, BSRT, Mémoire logique WMS-R, Rappel figure de Rey, Apprentissage visuel de Rey, Cubes WAIS-R, Complètement d'image WAIS-R, Assemblage d'objets WAIS-R, HVOT, copie figure de Rey, Habiletés motrices, Dynamomètre, FTT (Sherman et al., 1997) (Canada)	
24	**WPPSI-IV**	Wechsler, 2012 (États-Unis)	KABC II (Wechsler, 2012) (France) Bayley-III, NNAT2, NEPSY-II (Wechsler, 2012) (États-Unis)		*r* = .96 à .99 (Wechsler, 2012) (France)
25	**Mini-SEA**		BREF, MMSE, MDRS, WCST, Fluence morphologique (Bertoux et al., 2012) (France)		
26	**NPI**		Hamilton Rating Scale for Depression (Cummings et al., 1997) (États-Unis)		
27	**D-KEFS**	Subtest de la tour : Trakoshis et al., 2020 (Chypre)	Subtest TMT validé (Sánchez-Cubillo et al., 2009) (Espagne)		
28	**BADS**	Wilson et al., 1996 (Royaume-Uni)	COWAT (Norris & Tate, 2000) (Australie)		*r* = .88 à 1.00 (Wilson et al., 1998) (États-Unis)

29	**Figure complexe de Rey**		Test de Stroop, Developmental Test of Visual Perception, Matrices WAIS-III, Cubes WAIS-III (Smith et al., 2007) (États-Unis)		*r* = .45 à 1.00 (Smith et al., 2007) (États-Unis)
30	**Token test**	Jardim de Paula et al., 2011 (Brésil)			Corrélation intra-classe (ICC) = .12 à 1.00 (Paci et al., 2015) (Italie)
31	**SDMT**		Brief Repeatable Battery of Neuropsychological Tests (Sonder et al., 2014) (Pays-Bas)		
32	**RBMT**		Empan de chiffres, Blocs de Corsi (Wilson et al., 2010) (France)		
33	**BCCOGSEP**	Subtest PASAT : Crawford et al., 1998 (États-Unis)	Subtest PASAT validé (Sherman et al., 1997) (Canada)	Subtest PASAT validé (Sherman et al., 1997) (Canada)	
34	**BREP**				
35	**HVOT**	Merten, 2005 (Allemagne)	MMSE, Syndrom-Kurztest, Line orientation test de Benton, Three-dimensional block constuction de Benton, TMT, Cubes WAIS-R, Vocabulaire WAIS-R, Empan de chiffres WAIS-R, Matrices de Raven, Blocs de Corsi, Alerte TAP, Memo test, WST Vocabulary test, Naming test Aachener Aphasie Test (Merten, 2005) (Allemagne)		*r* = .99 (Lopez et al., 2003) (États-Unis)
36	**JLO**				
37	**BNT**				Accord inter-juges : 83% (Tallberg, 2005) (Suède)
38	**BSRT**				
39	**BDAE**	Fong et al., 2019 (États-Unis)			

40	**FTT**				
41	**CPT**	Egeland et al., 2010 (Norvège)	Knox Cube test, Empan envers et endroit WAIS-R, PASAT, TMT, Test de Stroop (Egeland et al., 2010) (Norvège)		
42	**SEA**		BREF, MMSE, MDRS, WCST, Fluence morphologique (Bertoux et al., 2012) (France)		
43	**HVLT**	Shapiro et al., 1999 (États-Unis)	Reproduction visuelle, WMS-R, QI verbal WAIS-R (Shapiro et al., 1999) (États-Unis)		
44	**RAVLT**		Brief Visuo-Spatial Memory test revised (Soble et al., 2020) (États-Unis)		
45	**COWAT**				Corrélation intraclasse (ICC) = .94 à .99 (Ross, 2003) (États-Unis)
46	**RBANS**	Torrence et al., 2016 (États-Unis)		BDAE, Vocabulaire WAIS-R, Matrices de Raven, Benton facial recognition test, TMT, Line cancellation test, Executive interview EXIT, CES-D10, RBMT (Larson et al., 2005) (États-Unis)	
47	**VMI**	Sutton et al., 2011 (États-Unis)	JLO (Malloy et al., 2003) (États-Unis)	COWAT, TMT, MDRS, Behavior Dyscontrol Scale, Boston Naming Test (Malloy et al., 2003) (États-Unis)	*r* = .75 à .88 (Harvey et al., 2017) (États-Unis)

48	**WRMT**			BNT, Animal naming, Similitudes WAIS-R, Cubes WAIS-R, WCST (Soukup et al., 1999) (États-Unis)
49	**Test de Barcelone**			
50	**NCSE/ Cognistat**	Engelhart et al., 1999 (États-Unis)		
51	**LNNB**			
52	**MAS**	Williams, 1999 (États-Unis)		
53	**TOMM**	Heyanka et al., 2015 (États-Unis)	Reliable digit span WAIS-IV (Webber et al., 2018) (États-Unis)	
54	**CERAD**	Strauss et Fritsch, 2004 (États-Unis)		
55	**HRNB**	Ross, 2013 (États-Unis)		
56	**BVMGT**	Decker et al., 2006 (États-Unis)		
57	**WRAT**			

Note. Si aucune information n'est mentionnée, cela signifie que nous n'avons pas trouvé d'étude correspondante pour l'outil concerné.
L'ensemble des études de validité de construit « pures » ont utilisé une méthodologie de type « analyse factorielle », les études de validité convergente et divergente ont utilisé un méthodologie de type « corrélation ».

Tableau A4. *Validité concourante pour chaque outil (ordre décroissant de pourcentage d'utilisation).*

	Outil	Subtest / Sous-tâche	Outil de comparaison	*r*	Pays	Référence
1	**Test de Stroop**	Lecture	Symboles WAIS-IV	.22	Espagne	Periáñez et al., 2021
		Dénomination	Symboles WAIS-IV	.27		
2	**TMT**	Partie A	Symboles WAIS-III	-.63	Espagne	Sánchez-Cubillo et al., 2009
		Partie B	Score de changement WCST	.33		
3	**RL/RI16**	Rappel libre immédiat	Rappel immédiat RAVLT	.34	Italie	Clerici et al., 2017
		Rappel total immédiat	Rappel immédiat RAVLT	.22		
		Rappel libre différé	Rappel différé RAVLT	.52		
		Rappel total différé	Rappel différé RAVLT	.37		
4	**WAIS-IV**	Cubes	Cubes WISC-IV	.82	France	Wechsler, 2011
		Similitudes	Similitudes WISC-IV	.67		
		Mémoire des chiffres	Mémoire des chiffres WISC-IV	.68		
		Matrices	Matrices WISC-IV	.69		
		Vocabulaire	Vocabulaire WISC-IV	.75		
		Arithmétique	Arithmétique WISC-IV	.59		
		Symboles	Symboles WISC-IV	.53		
		Information	Information WISC-IV	.72		
		Code	Code WISC-IV	.75		
		Séquences lettres-chiffres	Séquences lettres-chiffre WISC-IV	.67		
		Compréhension	Compréhension WISC-IV	.77		
		Barrage	Barrage WISC-IV	.36		
		Complètement d'image	Complètement d'image WISC-IV	.32		
5	**d2-r**					
6	**BREF**	Score total	Score total MDRS	.82	France	Dubois et al., 2000
7	**DO80**					
8	**VOSP**	Silhouettes	HVOT	.65	Allemagne	Merten, 2006
9	**WCST**					

10	**MEM-IV**	Mémoire logique I	Histoires CMS	.38		
		Mémoire logique II	Histoires rappel différé CMS	.38		
		Mots couplés I	Mots couplés CMS	.42		
		Mots couplés II	Mots couplés rappel différé CMS	.26		
		Dessins I	Localisation de points CMS	.37		
		Dessins II	Localisation de points rappel différé CMS	.26	États-Unis	Wechsler, 2011
		Reproduction visuelle I	Localisation de points CMS	.45		
		Reproduction visuelle II	Localisation de points rappel différé CMS	.26		
		Addition spatiale	Mémoire des chiffres CMS	.48		
		Mémoire des symboles	Mémoire des chiffres CMS	.49		
11	**CVLT**	Liste A Total	Quotient mnésique verbal WMS-R	.66		
		Liste B	Quotient mnésique verbal WMS-R	.36		
		Rappel immédiat libre	Quotient mnésique verbal WMS-R	.66		
		Rappel immédiat indicé	Quotient mnésique verbal WMS-R	.65	Québec	Nolin, 1999
		Rappel différé libre	Quotient mnésique verbal WMS-R	.50		
		Rappel différé indicé	Quotient mnésique verbal WMS-R	.69		
		Reconnaissance	Quotient mnésique verbal WMS-R	.27		
12	**DMS48**					
13	**MMSE**					
14	**Test des portes**					
15	**Test de l'horloge**					

16	**WISC-V**	Similitudes	Similitudes WAIS-IV	.75	France	Wechsler, 2016
		Vocabulaire	Vocabulaire WAIS-IV	.64		
		Information	Information WAIS-IV	.57		
		Cubes	Cubes WAIS-IV	.78		
		Puzzles visuels	Puzzles visuels WAIS-IV	.83		
		Matrices	Matrices WAIS-IV	.67		
		Arithmétique	Arithmétique WAIS-IV	.60		
		Mémoire des chiffres	Mémoire des chiffres WAIS-IV	.84		
		Symboles	Symboles WAIS-IV	.61		
		Code	Code WAIS-IV	.33		
17	**TEA / TAP**					
18	**MoCA**	Score total	Score total MDRS	.77	Canada	Lam et al., 2013
19	**NEPSY-II**	Catégorisation	Identification de concepts WISC-IV	.24	États-Unis	Korkman et al., 2012
		Compréhension de consignes	Compréhension WISC-IV	.43		
		Mémoire des figures immédiat	Localisation de points 1 CMS	.46		
		Mémoire des figures différé	Localisation de points 2 CMS	.36		
		Mémoire des visages immédiat	Visages 1 CMS	-.06		
		Mémoire des visages différé	Visages 2 CMS	.16		
		Mémoire narrative	Mémoire WNV	.24		
		Interférence listes de mots	Mémoire WNV	.23		
20	**Test de Hayling**					
21	**BECS-GRECO**					
22	**MDRS**	Score total	Score total MMSE	.85	Argentine	Luis Fernandez et Scheffel, 2003
23	**PASAT**	Score total	Score total SDMT	.54	Pays-Bas	Sonder et al., 2014

24	**WPPSI-IV**	Information	Information WISC-IV	.66		
		Similitudes	Similitudes WISC-IV	.69		
		Vocabulaire	Vocabulaire WISC-IV	.70		
		Compréhension de situations	Compréhension de situations WISC-IV	.48		
		Cubes	Cubes WISC-IV	.70	France	Wechsler, 2012
		Matrices	Matrices WISC-IV	.67		
		Identification de concepts	Identification de concepts WISC-IV	.33		
		Symboles	Symboles WISC-IV	.60		
		Barrage	Barrage WISC-IV	.51		
		Code	Code WISC-IV	.54		
25	**Mini-SEA**					
26	**NPI**					
27	**D-KEFS**	(Voir subtest TMT, Sánchez-Cubillo et al., 2009)				
28	**BADS**	Programmation de l'action	Porteus maze	-.34		
		Test de la clé	Porteus maze	-.28		
		Test des six éléments modifiés	Porteus maze	-.30		
		Changement de carte	TMT	-.30	Australie	Norris et Tate, 2000
		Plan du zoo	Porteus maze	-.41		
		Jugement temporel	Cognitive estimation test	.11		
29	**Figure complexe de Rey**	Copie	Developmental Test of Visual Perception	.42		
30	**Token test**				États-Unis	Smith et al., 2007
31	**SDMT**	Score total	Score total PASAT	.54	Pays-Bas	Sonder et al., 2014
32	**RBMT**	Dépistage	WRMT Words	.60	France	Wilson et al., 2010
		Profil	WRMT Words	.63		
33	**BCCOGSEP**					
34	**BREP**					
35	**HVOT**	Score total	Silhouettes VOSP	.64	Allemagne	Merten, 2005

36	**JLO**					
37	**BNT**					
38	**BSRT**					
39	**BDAE**					
40	**FTT**					
41	**CPT**					
42	**SEA**					
43	**HVLT**	Rappel total immédiat	Mémoire logique WMS-R, Rappel immédiat WMS-R	.75	États-Unis	Shapiro et al., 1999
		Rappel différé	Mémoire logique WMS-R Rappel différé	.77		
44	**RAVLT**					
45	**COWAT**					
46	**RBANS**	Indice d'attention	Executive interview EXIT	-.37	États-Unis	Larson et al., 2005
		Indice de langage	Vocabulaire WAIS-R	.51		
		Indice visuo-spatial	Matrices de Raven	.66		
		Indice de mémoire immédiate	RBMT	.68		
		Indice de mémoire différée	RBMT	.72		
47	**VMI**	Score total	Figure complexe de Rey	.36		
48	**WRMT**	Reconnaissance de visages	Benton facial recognition test	.51	États-Unis	Soukup et al., 1999
49	**Test de Barcelone**					
50	**NCSE / Cognistat**					
51	**LNNB**					
52	**MAS**					
53	**TOMM**	Score total	Word choice test	.30	États-Unis	Webber et al., 2018
54	**CERAD**					
55	**HRNB**					

56	**BVMGT**	Copie	WISC-III	.55	États-Unis	Volker et al., 2010
		Rappel	WISC-III	.34		
57	**WRAT**	Blue word-reading list	Wechsler Test of Adult Reading	.78	États-Unis	Mullen et Fouty, 2014
		Green word-reading list	Wechsler Test of Adult Reading	.75		

Note. Si aucune information n'est mentionnée, cela signifie que nous n'avons pas trouvé d'étude correspondante pour l'outil concerné.
La validité concourante est exprimée en corrélation *r* de Pearson.

Tableau A5. *Fidélité test-retest pour chaque outil (ordre décroissant de pourcentage d'utilisation).*

	Outil	Subtest / Sous-tâche	*r*	Pays	Référence
1	**Test de Stroop**	Lecture – Temps	.75	Québec	Tremblay, 2004
		Lecture – Erreurs	.15		
		Dénomination – Temps	.82		
		Dénomination – Erreurs	.26		
		Interférence – Temps	.82		
		Interférence – Erreurs	.44		
2	**TMT**	Partie A	.35	États-Unis	Basso et al., 1999
		Partie B	.64		
3	**RL/RI16**				
4	**WAIS-IV**	Cubes	.54	France	Wechsler, 2011
		Similitudes	.66		
		Mémoire des chiffres	.69		
		Matrices	.69		
		Vocabulaire	.63		
		Arithmétique	.61		
		Symboles	.67		
		Puzzles visuels	.64		
		Information	.70		
		Code	.69		
		Séquences lettres-chiffres	.65		
		Balances	.66		
		Compréhension	.62		
		Barrage	.62		
		Complètement d'image	.63		
5	**d2-r**	CCT	.89	France	Brickenkamp et al., 2015
		CC	.90		
		E%	.35		

6	**BREF**	Score total	.85	Italie	Appollonio et al., 2005
7	**DO80**				
8	**VOSP**	Silhouettes	.88	Royaume-Uni	Bird et al., 2004
9	**WCST**	Erreurs total	.50	États-Unis	Basso et al., 1999
		Réponses persévératives	.50		
		Erreurs persévératives	.52		
		Pourcentage réponse conceptuelle	.54		
		Catégories complétées	.54		
		Nombre d'échecs à maintenir le set	-.02		
		Apprentissage	.36		
		Essais totaux	.30		
10	**MEM-IV**	Mémoire logique I	.73	France	Wechsler, 2012
		Mémoire logique II	.68		
		Mots couplés I	.76		
		Mots couplés II	.76		
		Dessins I	.73		
		Dessins II	.72		
		Reproduction visuelle I	.62		
		Reproduction visuelle II	.59		
		Addition spatiale	.74		
		Mémoire des symboles	.71		
11	**CVLT**	Essais 1 à 5	.64	États-Unis	Alioto et al., 2017
		Rappel libre immédiat	.67		
		Rappel libre indicé immédiat	.66		
		Rappel libre différé	.69		
		Rappel libre indicé diffré	.62		
		Reconnaissance	.57		
12	**DMS48**				
13	**MMSE**	Score total	.32	République Tchèque	Kopecek et al., 2016

14	**Test des portes**				
15	**Test de l'horloge**				
16	**WISC-V**	Similitudes	.67	France	Wechsler, 2016
		Vocabulaire	.65		
		Information	.70		
		Compréhension	.60		
		Cubes	.67		
		Puzzles Visuels	.66		
		Matrices	.65		
		Balances	.66		
		Arithmétique	.67		
		Mémoire des chiffres	.65		
		Mémoire d'images	.66		
		Séquences lettres-chiffres	.63		
		Symboles	.65		
		Code	.63		
		Barrage	.62		
17	**TEA / TAP**	Alerte phasique	.37	Europe	Zimmermann et Fimm, 2010
		Attention divisée I	.52		
		Balayage visuel	.27		
		Comparaison intermodale	.21		
		Examen du champ visuel	.71		
		Flexibilité (lettres-chiffres)	.62		
		Go/Nogo (2 stimuli, 1 cible)	.40		
		Incompatabilité	.31		
		Mémoire de travail	.46		
		Négligence	.77		
18	**MoCA**	Score total	.47	République Tchèque	Kopecek et al., 2016
19	**NEPSY-II**	Attention auditive	.56	États-Unis	Korkman et al., 2012

Fluidité de dessins	.61
Dénomination – Total durée	.81
Inhibition – Total durée	.79
Inhibition – Erreurs	.64
Changement – Total durée	.83
Statue	.79
Catégorisation	.67
Réponses associées	.67
Horloge	.72
Compréhension de consignes	.78
Processus phonologiques	.83
Productions de sons	.76
Dénomination rapide	.85
Production de mots – Sémantique	.84
Production de mots – Lettre initial	.54
Mémoire des figures – Contenu	.74
Mémoire des figures – Spatial	.62
Mémoire des figures – Total	.70
Mémoire des figures différée – Contenu	.66
Mémoire des figures différée – Spatial	.67
Mémoire des figures différée – Total	.67
Mémoire des visages	.62
Mémoire des visages différée	.66
Mémoire des prénoms	.67
Mémoire des prénoms différée	.53
Mémoire des prénoms – Total	.70
Mémoire narrative – Rappel libre et indicé	.76
Mémoire narrative – Rappel libre	.68

		Imitation position de mains – Répétition	.75		
		Imitation position de mains – Rappel	.72		
		Imitation position de mains – Total	.49		
		Liste de mots	.66		
		Répétition de phrases	.77		
		Précision visuomotrice	.71		
		Reconnaissance d'affects	.55		
		Théorie de l'esprit	.77		
		Flèches	.65		
		Cubes	.76		
		Copie de figures – Motricité	.63		
		Copie de figures – Global	.66		
		Copie de figures – Local	.59		
		Copie de figures – Total	.72		
		Puzzles géométriques	.71		
		Puzzles d'images	.80		
20	**Test de Hayling**				
21	**BECS-GRECO**				
22	**MDRS**	Score total	.82		
		Attention	.77		
		Persévération	.66	États-Unis	Schmidt et al., 2005
		Conceptualisation	.70		
		Mémoire	.80		
23	**PASAT**	Score total	.83	Pays-Bas	Sonder et al., 2014
24	**WPPSI-IV**	Information	.70	France	Wechsler, 2014

		Similitudes	.62		
		Vocabulaire	.66		
		Compréhension de situations	.60		
		Compréhension de mots	.57		
		Dénomination d'images	.69		
		Cubes	.47		
		Assemblage d'objets	.65		
		Matrices	.62		
		Identification de concepts	.61		
		Reconnaissance d'images	.69		
		Mémoire spatiale	.54		
		Symboles	.63		
		Barrage	.44		
		Code	.63		
25	**Mini-SEA**				
26	**NPI**	Fréquence	.79	États-Unis	Cummings et al., 1997
		Sévérité	.86		
27	**D-KEFS**	Design fluency	.32	États-Unis	Shunk et al., 2001
		Twenty questions test	.24		
		Word contexte test	.70		
		Proverb test	.76		
28	**BADS**	Programmation de l'action	.67	Royaume-Uni	Wilson et al., 1998
		Test des clés	.71		
		Cartes	-.08		
		Plan du zoo	.39		
		Jugement temporel	.64		
29	**Figure complexe de Rey**	Copie	.56	États-Unis	Nakhutina et al., 2010
		Rappel	.70		
30	**Token test**				

31	**SDMT**	Score total	.70	Portugal	Pereira et al., 2015
32	**RBMT**				
33	**BCCOGSEP**				
34	**BREP**				
35	**HVOT**				
36	**JLO**				
37	**BNT**	Score total	.86	États-Unis	Sachs et al., 2012
38	**BSRT**	Rappel total	.72	Espagne	Campo et al., 2004
		Récupération à long terme	.70		
		Récupération à court terme	.61		
		Encodage à long terme	.65		
		Récupération à long terme consistante	.70		
		Récupération à long terme aléatoire	.41		
		Intrusions	.34		
		Rappel indicé	.50		
		Choix multiple	.23		
		Choix multiple dfféré	.35		
		Rappel différé	.65		
		Essai 1	.40		
		Dernière récupération à long terme consistante	.62		
39	**BDAE**				
40	**FTT**				
41	**CPT**	Pourcentage omissions	.09	Canada	Soreni et al., 2009
		Pourcentage fausses alertes	.72		
		Temps de réaction	.76		
		Erreur standard – Temps de réaction	.63		
42	**SEA**				

43	**HVLT**	Rappel total	.80	États-Unis	O'Neil-Pirozzi et al., 2012
		Rappel différé	.82		
		Pourcentage de rétention	.69		
		Reconnaissance	.64		
44	**RAVLT**	Essai 1 à 3	.63	Pays-Bas	Van der Elst et al., 2007
		Essai 1 à 5	.69		
		Rappel différé	.67		
45	**COWAT**		.74	États-Unis	Ruff et al., 1996
46	**RBANS**	Score total	.90	Espagne	De la Torra et al., 2014
47	**VMI**	Score 1	.62	États-Unis	Harvey et al., 2017
		Score 2	.56		
48	**WRMT**	Mots	.69	Royaume-Uni	Bird et al., 2003
		Visages	.76		
49	**Test de Barcelone**				
50	**NCSE / Cognistat**				
51	**LNNB**				
52	**MAS**				
53	**TOMM**				
54	**CERAD**				

55	**HRNB**	Category	.85		
		Tactual performance test – Total	.83		
		Tactual performance test – Memory	.71		
		Tactual performance test – Location	.60		
		Seashor rythm	.70		
		Speech sounds perception	.80		
		Trails A	.79	États-Unis	Dikmen et al., 1999
		Trails B	.89		
		Tapping D	.77		
		Tapping ND	.78		
		Grip D	.90		
		Grip ND	.91		
		Impairment index	.81		
		Average impairment rating	.92		
56	**BVMGT**				
57	**WRAT**	Score total	.90	États-Unis	Ashendorf et al., 2009

Note. Si aucune information n'est mentionnée, cela signifie que nous n'avons pas trouvé d'étude correspondante pour l'outil concerné.
Les corrélations qui étaient exprimées en *Z* de Fisher ont préalablement été modifiées en *r* de Pearson.

Résumé

Le contexte actuel dans lequel s'inscrit la neuropsychologie clinique est marqué par la volonté de développer une démarche *evidence-based.* Pour les neuropsychologues, l'application de l'Evidence-Based Practice (EBP) doit idéalement s'opérer dès l'étape de l'évaluation cognitive en veillant à considérer les trois piliers de la démarche de façon intégrée.

Dans ce cadre, nous nous sommes questionnés sur les pratiques évaluatives des neuropsychologues francophones ainsi que sur les qualités des outils qu'ils emploient. En effet, nous avions pour objectif de faire le point sur l'état actuel de validation des outils d'évaluation fréquemment utilisés par les neuropsychologues français, tout en mettant en évidence les variables qui influencent les professionnels dans le choix de ceux-ci. Pour ce faire, nous avons répertorié plusieurs caractéristiques psychométriques (données normatives, validité et fidélité) relatives à une liste d'outils préalablement définie. En parallèle, nous avons cherché à déterminer quelles données psychométriques relatives aux outils sont liées à leur fréquence d'utilisation. Enfin, nous avons mené une méta-analyse sur base des indices de fidélité test-retest afin de déterminer si les outils d'évaluation cognitive fournissent des mesures stables. Au travers de cette analyse, nous avons également évalué la présence d'éventuels modérateurs susceptibles d'influencer l'indice test-retest.

Nos résultats ont permis de mettre en lumière divers éléments. D'une part, les outils les plus fréquemment employés s'accompagnent de normes de bonne qualité. D'autre part, nous avons pu mettre en évidence un manque de recherche dans le domaine psychométrique en Europe francophone, particulièrement lorsqu'il s'agit d'évaluer la validité ainsi que la fidélité des mesures fournies par nos outils. De plus, certaines données suggèrent que les neuropsychologues considèrent peu les qualités psychométriques de leurs outils. Nous avons émis plusieurs pistes explicatives à ce sujet. Enfin, notre méta-analyse a révélé que les outils d'évaluation présentent globalement une bonne fidélité test-retest, qui varie selon certains modérateurs. À ce sujet, cependant, nos résultats ne rejoignent pas l'ensemble des études précédemment menées.

Pour clôturer ce travail, nous avons pris le temps de proposer quelques pistes à mettre en place pour tendre davantage vers une démarche évaluative *evidence-based*. Nous avons également discuté des limites de notre étude, qui, nous le pensons, mérite d'être perfectionnée afin de mieux cibler les pratiques des neuropsychologues francophones.