L'interprétation et la communication des scores en neuropsychologie : en finir avec la tour de Babel

Auteurs :

Sylvie Willems*; Hichem Slama*; Béatrice Degraeve*; Groupe Label ; Patrick Fery*

Groupe Label : François Radiguer[+] ; George Michael[+]; Hélène Amieva[+] ; Vincent Verdon[+] ; Christine Moroni[+] ; Philippe Azouvi[+] ; Martine Roussel[+] ; Amélie Ponchel[+]

*Membres du premier groupe rédactionnel ayant contribué également à la rédaction de ce document

[+] Membres du deuxième groupe rédactionnel ayant contribué de manière équitable à la révision de ce document

Résumé :

En pratique, la neuropsychologie clinique présente une très grande variabilité dans la façon d'interpréter statistiquement un score obtenu à une tâche psychométrique et de le communiquer via les comptes-rendus écrits. L'absence de consensus touche aussi bien les seuils utilisés pour qualifier le score de « hors norme » que le vocable utilisé pour décrire ce score, mettant tantôt l'accent sur un déficit probable de la fonction cognitive sous-jacente (ex. « performance déficitaire »), tantôt sur la statistique (ex. « performance dans la moyenne »). Dans ce contexte, des experts en neuropsychologie se sont réunis afin de proposer un document de synthèse qui sera soumis à une conférence de consensus. Ce document fait d'abord état du problème lié au manque d'homogénéité avant d'aborder ensuite la question du score seuil. Cette question du score seuil étant liée aux étapes pré-interprétatives, un arbre décisionnel est proposé quant à la méthode statistique (score *z*, percentile…) utilisable selon les différentes caractéristiques de l'échantillon normatif. En outre, différentes précautions sont ensuite évoquées quant à l'utilisation des scores seuils. Bien que la vérification des qualités psychométriques de nos outils et de l'échantillonnage des normes fasse partie des premières précautions qui devraient être rappelées, ce point sort du cadre de la question du choix du seuil ou du vocable et n'est donc volontairement pas abordé de manière approfondie

dans le présent document. Finalement, une recommandation de système d'étiquettes qualitatives pour décrire les scores aux tests est formulée afin de tendre vers plus d'homogénéité dans la manière de qualifier un score par les neuropsychologues.

1. État du problème

L'une des tâches essentielles des neuropsychologues[1] est d'interpréter les résultats de l'évaluation neuropsychologique, puis de les communiquer clairement dans un rapport circonstancié. À cette fin, pour chaque épreuve administrée, il est utile de fournir les scores bruts obtenus par la personne évaluée, mais il est également indispensable de les situer par rapport à des données normatives adaptées avant même de fournir une interprétation clinique. La/le neuropsychologue doit alors mettre des mots et fournir les clés nécessaires à la compréhension de cette comparaison normative. Pour cela, il est indispensable de préciser les seuils de « normalité » utilisés et d'expliquer les catégories de scores avec des qualificatifs qualitatifs clairs pour la/le lectrice.eur (tels que « dans la norme »). Des systèmes de classifications et de qualificatifs ont été proposés (par exemple, Schoenberg & Rum, 2017 ; Schretlen, Testa, & Pearlson, 2010 ; Guilmette et al., 2020). Ils varient en termes de seuil utilisé, de nombre de catégories de scores et de qualificatifs. Ainsi, nous pouvons trouver des propositions à trois catégories (telles que « inférieur à la moyenne », « dans la moyenne » ou « supérieur à la moyenne » ; Brooks et al., 2011) et des propositions à sept ou neuf catégories. Dans le cadre de ses lignes directrices, l'Association Suisse de Neuropsychologie (www.neuropsy.ch) propose, par exemple, une solution en sept catégories :

Rang Percentiles	Qualificatifs associés aux valeurs statistiques
	Très supérieure à la norme
≥95 – ≤98	Clairement supérieure à la norme
>84 – <95	A la limite supérieure de la norme
≥16 – ≤84	Dans la norme
>5 – <16	A la limite inférieure de la norme
≥2–≤5	Clairement inférieure à la norme
0 – <2	Très inférieure à la norme

En Europe francophone, il n'existe cependant pas de consensus. Nous allons voir que plusieurs problèmes découlent directement de ce manque de consensus et de terminologie bien définie.

1.1 Divergence entre neuropsychologues

Bien qu'une majorité de neuropsychologues tentent d'utiliser des termes précis pour qualifier

[1] le terme « neuropsychologue » a été choisi pour faire référence aux clinicien.ne.s ayant une formation et une pratique spécialisée en neuropsychologie.

les scores dans leurs rapports (Leclef et al., 2018), il est commun d'observer une remarquable disparité entre les neuropsychologues qualifiant différemment un même score (Guilmette et al., 2020). Cette inconsistance peut participer à l'incompréhension de la/ du lectrice.eur lorsque plusieurs évaluations sont réalisées par des clinicien.ne.s différent.e.s et lorsque des interprétations différentes sont fournies pour des résultats identiques, d'autant plus que la moitié des neuropsychologues ne fournissent dans leurs écrits ni la définition des termes ni la classification utilisée pour décrire les scores obtenus aux tests. À titre d'exemple, dans une enquête menée auprès de 110 neuropsychologues américains, Guilmette et al. (2008) ont demandé d'attribuer une étiquette descriptive à 12 scores standardisés. Le nombre moyen d'étiquettes différentes associées à un même score était de 14. L'absence de consensus concernait autant les seuils utilisés que la terminologie employée pour qualifier le score. Soulignons que la confusion concernant les seuils n'est pas spécifique à la pratique clinique. Les chercheuses et chercheurs appliquent également parfois des seuils très variables pour qualifier un score de "hors-norme" (avec un score *z* compris entre 1 et 2 ; Beauchamp et al., 2015 ; Meyer, Boscardin, Kwasa, & Price, 2013 ; Schoenberg et al., 2018). En clinique, des différences culturelles sont également relevées avec un seuil régulièrement placé au percentile 5 en Suisse, France, Belgique (soit un score $z \approx 1{,}65$) et au percentile 16 dans les pays germaniques et anglophones (soit au score $z \approx 1$). Ces différences de pratique découlent probablement de multiples facteurs (par exemple, la différence d'outils utilisés et les recommandations associées, la transmission orale de pratique régionale).

1.2 Des qualificatifs ambigus

Concernant la terminologie choisie, certains qualificatifs peuvent participer aussi à la confusion. On constate que certain.e.s neuropsychologues optent pour des étiquettes assez éloignées de la logique de comparaison normative en mettant l'accent sur l'incertitude (par exemple, « performance limite »), la qualité suspectée de la performance ou la déficience présumée (par exemple, « performance faible », « performance déficitaire »), la position, la proximité ou l'éloignement vis-à-vis d'une moyenne (par exemple, « score inférieur/supérieur à la moyenne »), ou encore des comparaisons intra-individuelles impliquant une estimation du score prémorbide attendu compte tenu de l'âge et du niveau d'étude (par exemple, « score inférieur au niveau attendu »).

1.3 Divergences au sein des rapports

Outre ces différences entre professionnels, les neuropsychologues ont également tendance à varier leurs termes au sein d'un même document, tendant à préférer les synonymes aux répétitions sans doute monotones (Leclef et al., 2018). Plus problématique encore, le seuil utilisé au sein d'une même évaluation peut changer d'un test à l'autre en fonction de l'étalonnage utilisé (percentile, écart-type, score T, note standard ou score composite). Ainsi, un score à un test pourrait être considéré comme « hors-seuil » lorsque les normes sont exprimées en percentile, mais « dans les normes » dans le cas de l'utilisation d'un autre type d'étalonnage (par exemple, une note standard associée en réalité à la même probabilité statistique). Ces incohérences découlent probablement d'une méconnaissance de la signification de certains scores standardisés. Par exemple, un.e neuropsychologue peut oublier qu'un score T de 40 est équivalent à un score *z* de -1, et se rapproche donc du percentile 16 en cas de normalité de la distribution. Cette confusion est accentuée par les informations souvent contradictoires fournies dans les manuels des éditeurs des tests (Guilmette et al., 2020), conduisant les neuropsychologues à attribuer des étiquettes différentes au même score.

2. Une nécessité de consensus

Le manque de cohérence dans les terminologies et seuils utilisés est une difficulté ancienne pouvant nuire au crédit de la neuropsychologie clinique auprès des autres professions. C'est pourquoi la SNLF, l'OFPN, le GRECO et des partenaires belges et suisses ont décidé ensemble d'organiser une conférence de consensus afin de dégager des seuils et une terminologie plus homogène. Un groupe de 15 experts en neuropsychologie (voir infra, point 2.2) s'est réuni pour préciser l'objet de la conférence de consensus (voir infra, point 2.1), puis proposer un document de synthèse à soumettre à la conférence de consensus.

2.1 Méthodologie

Un premier groupe restreint s'est réuni fin 2022 sous l'impulsion de Xavier Seron et Sylvie Willems avec Catherine Belin, Hélène Amieva, Philippe Azouvi. Ce groupe a jugé la question des descripteurs essentielle et a pris l'initiative de constituer un comité élargi représentant la profession et différentes associations (SNLF, OFPN, GRECO). Ce comité était composé de George Michael ; Hélène Amieva ; François Radiguer ; Vincent Verdon ; Philippe Azouvi ; Christine Moroni ; Martine Roussel ; Amélie Ponchel ; Sylvie Willems ; Hichem Slama ; Béatrice

Degraeve ; Patrick Fery ; Jacques Grégoire. En mars 2023, ce comité a proposé d'organiser une conférence de consensus présidée par Jacques Grégoire. À cette fin, un premier document a été rédigé entre septembre 2023 et novembre 2023 par un groupe de rédaction restreint composé de Patrick Fery ; Hichem Slama ; Béatrice Degraeve ; Sylvie Willems. Ce document a été révisé par François Radiguer puis par George Michael ; Hélène Amieva ; Vincent Verdon ; Philippe Azouvi ; Christine Moroni ; Martine Roussel ; Amélie Ponchel entre novembre 2023 et janvier 2024. Le document a été retravaillé par le groupe de rédaction en fonction des commentaires puis rediscuté en comité élargi en avril 2024. En juin 2024, un jury a été constitué pour une révision finale du document.

Ce jury sera constitué de chercheurs en neuropsychologie et psychométrie (Thierry Lecerf ; Bruno Lenne ; Pierre-Yves Jonin ; Marie Geurten ; Philippe Allain), de clinicien.ne.s psychologues spécialisé.e.s en neuropsychologie (Emilie Favre ; Delphine Fleurion ; Giulia Dormal ; Valérie Vanderaspoilden), de médecins spécialisé.e.s en neuropsychologie (Catherine Belin ; Pascale Pradat-Diehl ; Olivier Godefroy ; Caroline Massot, ; Cécile Donze), et enfin de deux patients partenaires, autrement dit des personnes ayant déjà reçus une évaluation neuropsychologique.

2.2. Objet de la conférence de consensus

Au cours de deux séances préliminaires, le groupe a décidé d'orienter sa réflexion vers les qualificatifs employés pour définir un score. Cette discussion implique la question des seuils associés aux qualificatifs et du seuil pour qualifier un score de « hors norme ». En conséquence, le groupe a également choisi d'initier une réflexion en amont sur la logique de la comparaison normative et les statistiques utilisées. Ainsi, l'orientation du groupe a été de se concentrer exclusivement sur l'étape d'**analyse statistique dite « pré-clinique »** (voir figure 1) d'un score obtenu à une évaluation psychométrique et sur la communication de cette interprétation.

La réflexion n'a donc pas englobé l'**interprétation « clinique »** du résultat en lui-même (une approche jugée inappropriée puisqu'un score décontextualisé ne peut avoir de signification clinique propre en particulier l'interprétation doit tenir compte du profil global de la personne évaluée).

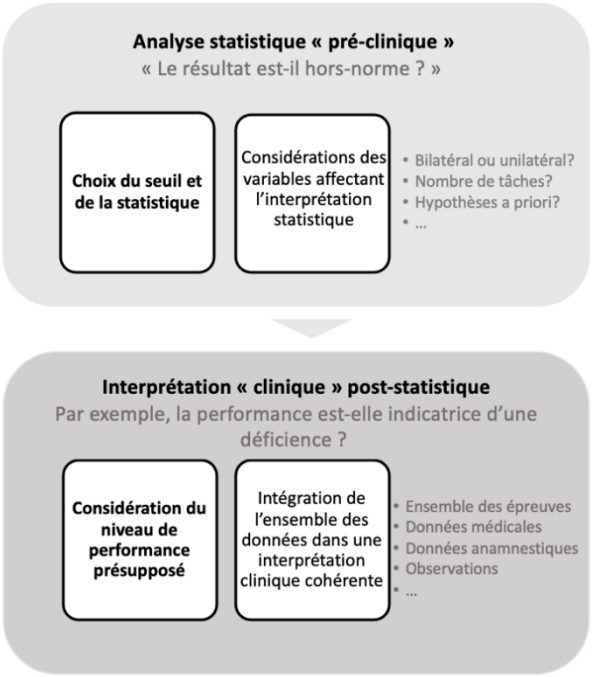

Figure 1. Les étapes de l'interprétation

Autrement dit, l'objectif de ce travail n'est pas de guider et encore moins de restreindre le clinicien dans son interprétation « clinique » post-statistique. Cette étape implique en effet une synthèse experte d'un éventail très large d'informations, allant des différentes données récoltées aux tests neuropsychologiques et questionnaires, aux détails sur le contexte de vie, la situation médicale, comportementale et fonctionnelle de la personne évaluée. Le clinicien peut ainsi évaluer un score comme étant statistiquement dans la norme à l'étape pré-clinique, puis décider par la suite de le considérer comme un indicateur d'une situation à risque d'évolution vers un trouble (si, par exemple, il forme un profil à risque cohérent et validé avec d'autres résultats), comme un indicateur d'une déficience (par exemple, en comparaison avec un niveau attendu plus élevé), ou comme un indicateur d'une dégradation lors d'un suivi longitudinal (par exemple, s'il s'écarte significativement d'un score obtenu précédemment). Bien que le raisonnement clinique dépende fortement du contexte et de la question posée, l'analyse statistique préalable d'un score isolé demeure, dans ses premières étapes, similaire dans les différentes situations cliniques.

En dernier lieu, notons encore que la réflexion n'a pas abordé les étapes préalables à l'interprétation. Ces étapes, nombreuses et cruciales, vont du choix (et de l'analyse critique) d'un outil psychométrique fiable et valide en fonction de la question clinique posée, la qualité

de l'échantillonnage des normes utilisées, à la standardisation de l'administration d'un test. Il est essentiel de rappeler que ces étapes sont des prérequis impératifs à l'interprétation. Par exemple, l'utilisation d'un échantillon normatif non représentatif de la population que l'on cherche à évaluer rendra totalement non valide l'étape d'interprétation statistique du score. Cependant, bien que ces prérequis soient indispensables, ces éléments méthodologiques dépassent la portée spécifique de notre travail et ne seront que succinctement abordés.

3. La logique de la comparaison normative : seuil de décision et risque d'erreur

Avant d'aborder la question du score seuil, les auteurs ont estimé utile de rappeler la logique du raisonnement sous-tendant l'utilisation de normes. Confronté.e au score d'une personne dans une tâche cognitive (ou dans un questionnaire ou encore face à un comportement observé), la/le neuropsychologue doit généralement répondre à la question suivante : ce score correspond-il à ce que l'on pourrait attendre chez cette personne (ce que l'on nomme en statistique l'hypothèse nulle ou H0) ou bien est-il le signe possible d'une altération du fonctionnement cognitif (l'hypothèse alternative ou H1) ? Le plus souvent, la/le neuropsychologue ne dispose pas d'un score de référence exact lui permettant de répondre directement à cette question comme, par exemple, le score antérieur de la personne avant son accident ou sa maladie. Elle/il doit donc estimer la probabilité que ce score reflète une différence par rapport à ce qui est attendu et prendre une décision sur base de cette probabilité. Pour estimer cette probabilité, des outils statistiques sont nécessaires. C'est pour cette raison que les cours de statistiques sont des éléments indispensables à la formation en psychologie, au grand bonheur des étudiants qui mettent souvent plusieurs années à comprendre le sens et l'importance de ces enseignements pour leur pratique. Les outils statistiques généralement utilisés en neuropsychologie clinique se prononcent sur le rejet et non sur l'acceptation de l'hypothèse nulle (H0) (voir Heck et al., 2023 pour une discussion sur l'évaluation de H0 en psychologie à l'aide du facteur de Bayes). Reformulé autrement, ces outils statistiques évaluent la probabilité de l'égalité (H0) et, si celle-ci est suffisamment faible, permettent d'accepter le rejet de l'égalité (c'est-à-dire, d'accepter la différence ou H1). Notons que le rejet de H0 (égalité) en faveur de H1 (différence) ne constitue pas pour autant une preuve de la véracité de H1. Cela peut donner l'impression qu'on joue sur les mots en statistiques… Toutefois, si une personne vous dit : « Je ne vous déteste pas », vous n'allez pas

penser qu'elle vient de vous dire : « Je vous aime ». C'est l'une des particularités du raisonnement scientifique que d'accepter une hypothèse (H1, différence) à travers le rejet d'une autre (H0, égalité), estimée comme trop peu probable. À l'inverse, si à l'issue du test nous ne rejetons pas H0 (égalité), nous ne confirmons pas pour autant cette dernière hypothèse. Si la/le neuropsychologue peut être tenté.e de conclure que la fonction sous-tendant le score est « préservée » (ou « non-altérée »), nous n'avons en réalité aucune preuve de cela. Ne pas rejeter H0 (égalité) ne veut pas dire que H0 est vraie. Ainsi, après un test non-significatif, la seule conclusion que nous pouvons finalement tirer à propos de la fonction testée, c'est que nous n'avons pas suffisamment d'indices pour conclure qu'elle est altérée. Cette subtile nuance crée toute la différence : ne pas dire : « Je vous aime » ne revient pas à dire « Je ne vous aime pas ».

La/le neuropsychologue estime donc le risque de se tromper en décrétant que le score diffère de celui attendu (erreur de type I, également appelé « alpha »). Elle/il fixe également un seuil associé à un risque acceptable de se tromper (le seuil de risque d'erreur). Un avis neuropsychologique concernant un score observé chez une personne évaluée reflète donc une décision basée sur une probabilité de rejet de l'égalité (H0) associée à un seuil de risque d'erreur (voir Michael & Amieva, 2023 pour davantage de précisions).

En neuropsychologie clinique, ces probabilités sont estimées à l'aide de paramètres statistiques issus d'un échantillon de données provenant de la population de référence (échantillon le plus proche possible des caractéristiques de la personne évaluée)[2]. Généralement, ces paramètres sont le score moyen (ou parfois le score médian) de l'échantillon normatif et les variations de scores observées (en général l'écart-type à la moyenne, et plus rarement la médiane des écarts absolus à la médiane). Prenons l'exemple classique du score *z* qui, sur la base de ces paramètres, permet d'estimer la distance (écart standardisé exprimé en nombre d'écarts-types) qui sépare le score observé de la moyenne de l'échantillon. Sur la base théorique de la distribution normale des données (ou distribution de Gauss, voir Michael & Amieva, 2023 pour un rappel concernant la loi normale), le score *z* peut

[2] L'échantillon doit se rapprocher le plus possible des caractéristiques de la personne évaluée afin de ne pas biaiser l'estimation de la probabilité de différence. En réalité, il est souvent composé de personnes partageant seulement un certain nombre de caractéristiques avec la personne évaluée, choisies car elles peuvent affecter significativement les résultats dans des tâches cognitives (ex. âge, niveau d'étude, parfois sexe). La « représentativité » de l'échantillon de référence est ainsi souvent très partielle ce qui peut avoir des conséquences sur les estimations du niveau attendu (Amieva, Michael & Allain, 2011).

être associé à une probabilité d'observer un score donné dans la population de référence. Par exemple, un score de 40 obtenu à une tâche dont la moyenne de l'échantillon de référence est 52 et l'écart-type est 7,29 correspond à un score *z* de -1,65. Sur la courbe de Gauss, la probabilité de se trouver à - 1,65 écart-type en dessous de la moyenne est estimée à environ 5 %, ce qui est habituellement considéré comme peu fréquent. Un.e neuropsychologue qui utilise comme score seuil un score *z* de -1,65 afin de prendre sa décision accepte un risque d'erreur estimé à 5% de qualifier une personne obtenant ce score comme n'appartenant pas à la population de référence. Ce risque d'erreur correspond donc à l'estimation du pourcentage de personnes de la population de référence présentant un score au moins aussi distant de la moyenne que le score de la personne évaluée. Un choix de seuil situé au score *z* de – (OU +) 1,65 implique donc d'accepter un taux de faux positifs de 5%, ce qui correspond à une spécificité de 95%, habituellement jugée comme acceptable. Un risque d'erreur de 5% équivaut en effet au risque maximum accepté en sciences humaines et par la plupart des neuropsychologues clinicien.ne.s en Europe francophone (voir supra). Au-delà de 5%, ces neuropsychologues décident généralement que le risque de faux positif est trop élevé et considèrent le score comme non différent de ce qui serait attendu (non-rejet de l'hypothèse nulle). Notons que dans l'exemple donné, on ne s'intéresse qu'au score de bas niveau (raisonnement dit « unilatéral »). C'est souvent le cas lorsque nous générons l'hypothèse d'un déficit après l'analyse des plaintes de la personne évaluée et du dossier médical. Nous pouvons alors en effet considérer uniquement la moitié de la distribution. Si la/le neuropsychologue s'intéresse à la fois aux scores de bas et de haut niveau (donc aux écarts à droite ET à gauche de la moyenne, par exemple pour un quotient intellectuel), les scores *z* <-1,96 ET >+1,96 seront utilisés comme scores seuils. En effet, les 5 % des scores considérés comme peu fréquents se distribuent des deux côtés de la distribution.

Pour illustrer l'importance du seuil du risque d'erreur fixé, une/un neuropsychologue qui considérerait comme hors-norme un score se situant à un écart-type en dessous de la moyenne (score *z* de -1) se tromperait presque une fois sur six face à une personne issue de la population de référence, puisqu'environ 16% des valeurs observées dans la population de référence sont au-delà de ce seuil (nous avons ainsi un risque de ≈ 16% de faux positifs). Pire, si une/un neuropsychologue qui s'intéresse aux scores se situant à un écart-type en dessous et au-dessus de la moyenne (score *z* de -1 ET +1) se tromperait presque une fois sur trois. Imaginons un monde où les neuropsychologues seraient tenus d'indiquer leur seuil de

décision sur leur carte de visite. Il est probable que peu de patient.e.s consulteraient chez une/un neuropsychologue indiquant sur sa carte de visite : « Je me trompe 1 fois sur 3 face à une personne saine».

La logique statistique décrite précédemment concerne donc la question de la spécificité des résultats, c'est à dire la capacité à détecter les troubles cognitifs en produisant le moins de faux positifs. Pour estimer cette spécificité, le clinicien doit disposer de normes sur un échantillon de personnes sans problématique médicale affectant le système nerveux central (population dite « normale » ou « tout-venante »). A l'inverse, la sensibilité est la capacité à détecter un maximum de personnes « malades » en produisant le moins de faux négatifs (rejets incorrects) (voir Annexe 1). Pour l'estimer, le clinicien doit disposer d'informations concernant les résultats généralement obtenus par la ou les populations cliniques auxquelles la personne évaluée pourrait appartenir d'après les différentes hypothèses posées. Dans la réalité clinique, le/la neuropsychologue dispose rarement de cette information. Cela implique en effet de disposer, pour chaque hypothèse diagnostique, d'un échantillon normatif composé de patient.e.s présentant la pathologie cible mais aussi les caractéristiques similaires à celle de la personne évaluée (âge, niveau de scolarité, sexe). Cette information serait précieuse puisque, sans celle-ci, il n'est pas possible d'estimer la valeur prédictive d'un résultat (c'est-à-dire, la probabilité qu'un score positif/négatif reflète un vrai positif/négatif, voir Annexe 1). Etant donné la rareté d'une telle information, nous ne rentrerons plus en détail sur la question de la sensibilité.

4. Quels outils statistiques choisir ? Quel seuil de décision fixer ?

Le choix de l'outil statistique (score z, percentile, score T, note standard …) et du score seuil associé devant être utilisé est une question complexe. Reprenons le cas du **score z**. Dans le contexte d'un raisonnement unilatéral où l'on ne s'intéresse qu'aux scores de bas niveau (susceptibles de refléter une altération), l'écart est considéré comme significatif si le score z est inférieur ou égal à -1,65 OU supérieur à 1,65 selon le type de score (par exemple, des temps de réponses ou le nombre de réponses erronées). Nous venons de le rappeler, le score seuil doit être adapté pour maintenir le risque d'erreur à 5% en cas de raisonnement bilatéral (score z ≤-1,96 ET ≥1,96). La décision d'un raisonnement bilatéral ou unilatéral n'est pas chose aisée. En effet, selon le contexte de l'examen neuropsychologique (par exemple, dans le cadre du diagnostic d'un trouble neurodéveloppemental), la présence de scores très élevés est

presque aussi informative que la présence de scores très bas et l'interprétation clinique tient généralement compte des forces et des faiblesses cognitives des patients.

Hormis ces premières considérations, pouvons-nous être certains que notre score z (de ± 1,65 en unilatéral, ou ± 1,96 en bilatéral) permettra de maintenir le risque d'erreur à 5%, compte tenu de deux caractéristiques essentielles des échantillons de référence : leur effectif et la distribution des données au sein de l'échantillon ? La réponse est malheureusement négative. En effet, lorsque la distribution des scores est normale (condition requise pour l'utilisation du score z), plus la taille de l'échantillon de référence est petite et plus le risque de faux positifs est élevé. À titre d'exemple, pour un score z de ± 1,65 en unilatéral le risque théorique de 5 % devient en réalité de 6,25 % avec un échantillon de 20 personnes, de 7,57 % avec un échantillon de 10 personnes et de 10,37 % avec un échantillon de 5 personnes. Seuls les échantillons de plus de 50 personnes maintiennent le risque de faux positif au niveau acceptable souhaité (<5,53%) (Crawford & Garthwaite, 2005), pour autant que l'échantillonnage ait été réalisé selon les règles de l'art. Certains auteurs se sont basés sur l'intervalle de confiance de 95% du score seuil et ont ainsi montré la nécessité de disposer d'effectifs très larges, estimés à environ 500 contrôles pour contenir le taux de faux positifs inférieur à 7% (Godefroy et al., 2014).

Par ailleurs, s'agissant de tests cognitifs, il est fréquent que la distribution des scores ne suive pas la loi normale. Par exemple, le nombre de réponses correctes à une tâche de reconnaissance lors d'un apprentissage d'une liste de mots ou encore le nombre total de rappels corrects à une tâche couplant un essai de rappel libre et un essai de rappel indicé sont souvent proches du plafond chez les personnes saines. Deux facteurs caractérisent les distributions : l'asymétrie et l'aplatissement. Pour un même effectif de l'échantillon de référence, en fixant le seuil à 5%, le taux de faux positifs associé à l'utilisation du score z augmente en fonction de la « sévérité » de l'asymétrie[3]. Par exemple, lorsque l'échantillon comporte 10 individus et que la distribution s'écarte fortement d'une distribution normale tant en ce qui concerne la symétrie que l'aplatissement, ce taux peut atteindre 10,65%

[3] Une distribution est **symétrique** si l'étalement de sa partie gauche est équivalent (ou en miroir) à celui de sa partie droite. Dans le cas contraire, la distribution est asymétrique, l'étalement étant plus long à gauche (c'est le cas lorsque la majorité des scores sont proches du score maximum, par exemple ; la partie droite est moins étalée que la partie gauche) ou à droite (c'est le cas lorsque la majorité des scores sont proches de 0, par exemple ; la partie gauche est moins étalée que la partie droite).

(Crawford & Garthwaite, 2005) (voir Annexe 2).

En conclusion, malgré ses apparentes qualités, le score z est associé à un taux de faux positifs supérieur à 5% lorsque l'effectif de l'échantillon est inférieur à 50, même si la distribution des scores au sein de l'échantillon est normale. Ces taux sont encore plus élevés lorsque la distribution est asymétrique et davantage encore lorsqu'elle est à la fois asymétrique et leptokurtique ou platykurtique.

Un autre outil statistique communément utilisé en clinique neuropsychologique est le **(per)centile** qui présente deux avantages en comparaison du score z (pour le calcul du percentile, voir l'Annexe 3). Premièrement, il peut être utilisé que la distribution des scores dans l'échantillon de référence soit normale ou non. En effet, rappelons qu'il ne dépend pas d'une estimation théorique du risque d'erreur puisqu'il représente le pourcentage réel de scores de l'échantillon normatif qui se situent en dessous d'un score donné. Deuxièmement, il en découle que la valeur du percentile nous renseigne beaucoup plus directement à la fois sur le fait que le score soit hors seuil ou non et sur sa rareté. Par exemple, si un score se situe au percentile 4, cela signifie à la fois qu'il s'écarte significativement de la distribution des scores du groupe de référence (si la/le neuropsychologue fixe le seuil au percentile 5), et que seuls 4% de l'échantillon de référence ont un score aussi bas (ou élevé). Malgré ces avantages indéniables, il y a plusieurs limites à l'utilisation du percentile. La première est que les scores qui ne font pas partie de l'échantillon de référence seront par défaut situés au percentile 0 ou 100 (souvent exprimé en « percentile <1 » et « percentile >99 », respectivement), quel que soit leur écart par rapport au score le plus bas de l'échantillon de référence. Par exemple, si le score le plus faible obtenu par les participants de l'échantillon normatif est de 22 (le maximum étant 30), le percentile sera le même que la personne évaluée obtienne 21, 12 ou 5/30, alors que l'écart par rapport à la distribution ne l'est pas. Cela n'affecte pas la décision de considérer le score comme hors seuil, mais cela affecte la décision quant au degré de sévérité de l'éventuelle altération cognitive sous-jacente. Enfin, l'effectif de l'échantillon et la fréquence des différents scores dans l'échantillon vont limiter la possibilité du calcul du percentile. Par exemple, si le seuil est percentile ≤5, alors le score le plus bas de l'échantillon est au percentile 5 avec un échantillon de 10 personnes dont une seule présente le score le plus bas ((0+(1*0,5)/10 = 0,05; 0,05 *100 = 5). Si le seuil est ≤2,5, alors il faut un échantillon d'au moins 20 personnes ((0+(1*0,5)/20 = 0,025; 0,025 * 100 = 2,5). Au moins 10 ou 20 personnes sont donc nécessaires dans le sous-groupe de référence, mais cette condition n'est pas rencontrée

dans toutes les études de normalisation basées sur le percentile. En conclusion, le percentile est moins équivoque dans son interprétation qui ne nécessite ni calcul, ni table de conversion, ni réassurance sur la symétrie et l'aplatissement. Pour ces raisons, la plupart des neuropsychologues reconnaissent que la simplicité des percentiles est indéniable. Malheureusement, plusieurs situations peuvent limiter leur possibilité d'utilisation, comme une petite taille d'échantillon.

Une troisième manière de déterminer si le score observé correspond au score attendu, moins répandue en neuropsychologie, consiste à le comparer à un score seuil (**cut-off score**). Dans ce cas, les données normatives indiquent le score brut en dessous duquel (ou au-dessus duquel) le score observé peut être considéré comme s'écartant de manière significative du score attendu. L'avantage de cette méthode est de nous informer sur la probabilité de faux positifs (via la spécificité associée au score seuil), mais également sur la probabilité de faux négatifs (via la sensibilité) (voir Annexe 1, pour un rappel de ces concepts). Comme mentionné dans la section précédente, ces informations sont néanmoins encore rares en neuropsychologie clinique, car elles nécessitent deux échantillons normatifs (un échantillon « tout-venant » et un échantillon « clinique »). En outre, l'information ainsi obtenue est spécifique aux populations investiguées. Dès lors, si un test est utilisé pour identifier un vieillissement cognitif pathologique, la sensibilité peut être calculée en observant le nombre de personnes souffrant d'un trouble neurocognitif majeur lié à la maladie d'Alzheimer identifiées par le score seuil du test. Toutefois, l'indice de sensibilité associé à ce score seuil sera spécifique uniquement pour cette population clinique et ne pourra être utilisé dans aucun autre contexte (par exemple, d'un trouble neurocognitif majeur lié à une maladie de Parkinson).

En résumé, le score z est limité dans son efficacité à maintenir le risque de faux positifs proche de 5 %, car il est affecté par la taille de l'effectif et/ou la forme de la distribution des données normatives. Le percentile est quant à lui parfois limité dans sa capacité à estimer la fréquence du score dans la population dont l'échantillon de référence est issu, par exemple, quand ce dernier est de petite taille. Le cut-off proposé sur base du calcul de la spécificité et de la sensibilité est quant à lui encore peu fréquent dans les études normatives ou nos manuels de test.

Afin de prendre en compte l'impact de la taille de l'échantillon sur le risque de faux positifs lorsque la distribution est normale, un **test (Student) t** modifié a été développé

permettant de comparer le score d'une seule personne à la moyenne d'un échantillon de référence (Sokal & Rohlf, 1995 ; Crawford & Howell, 1998)[4]. Un tableur Excel permettant de calculer ce score t modifié et la probabilité associée est communiqué dans l'article de Fery et Claes (soumis), toutefois nous en expliquons ici la logique. Pour décider si le score d'une personne est significativement différent de la moyenne de l'échantillon, il faut d'abord choisir entre l'hypothèse d'un écart significatif dans une direction précise (test unilatéral) ou aucune hypothèse spécifique (test bilatéral, voir remarques ci-avant). Ensuite, on utilise le nombre de degrés de liberté (l'effectif de l'échantillon moins un) et le seuil de la valeur p (ici 0,05) pour prendre cette décision. Ensuite, il s'agit de regarder où se situe la valeur t obtenue dans une table du t de Student compte tenu du nombre de degrés de liberté, du seuil et du fait que le test soit unilatéral ou bilatéral. Par exemple, dans la situation où la valeur t est -2,818, le nombre de degrés de liberté est 9, et le test est unilatéral, la valeur p associée est 0,01. Cela signifie que seul 1 % des individus de la population dont l'échantillon de référence est issu obtiendraient un score aussi bas (ou élevé) que celui de la personne évaluée. Il peut donc être inféré que le score de la personne évaluée s'écarte de manière significative de la moyenne du groupe de référence et que la fréquence de ce score est rare dans la population. Outre ces qualités, le t modifié maintient le risque de faux positifs autour de 5%, y compris pour des échantillons de petite taille : 5% pour un effectif égal à 20 ou à 10, ou 5,01% pour un effectif égal à 5 (Crawford & Garthwaite, 2005). Malheureusement, le test t modifié ne parvient pas à maintenir le risque de faux positifs proche de 5% quelle que soit la forme de la distribution. Il est en effet particulièrement affecté par l'asymétrie (voir Annexe 4). Brièvement, lorsque la distribution est asymétrique, le t modifié et le seuil ≤0,025 sont indiqués pour maintenir le risque à 5%. Lorsque la distribution est asymétrique et anormalement aplatie, le t modifié et le seuil ≤0,02 sont alors indiqués.

Dans la pratique courante, étant donné que nous disposons peu souvent de ces informations, il apparaît raisonnable de proposer d'adopter un comportement prudent lors de l'interprétation de valeurs p situées entre 0,02 et 0,05 et d'interpréter avec plus de certitude uniquement les scores sous le **seuil ≤0,02**. Cette remarque vaut bien évidemment

[4] La formule de ce *t* modifié est la suivante (Crawford & Howell, 1998) : (Score – Moyenne) / (écart standard * racine (N+1/N))

pour les scores z pour lesquels un risque théorique de 2% pourrait être privilégié, cela reviendrait à opter pour un seuil plus conservateur de **score z 2,05 en unilatéral (et 2,33 en bilatéral)**, et interpréter avec prudence tous scores z situés entre celui-ci et 1,65 (et 1,96 en bilatéral).

Deux dernières remarques méritent d'être mentionnées concernant l'utilisation du t modifié et du score z. Premièrement, le maintien des scores extrêmes dans l'échantillon de référence (méthode préconisée par Crawford et al., 2006), peut tirer la moyenne vers le haut et vers le bas et augmenter le risque de faux positif ou de faux négatif (voir l'Annexe 5). Deuxièmement, le test t et le score z ne sont pas calculables lorsque l'écart standard est égal à 0 (ce qui est par exemple le cas lorsque l'échantillon tout-venant ne commet par exemple aucune erreur à un test), ce qui place alors la/le clinicien·ne dans la nécessité d'estimer relativement subjectivement si le score est hors seuil ou non (voir l'Annexe 6).

Le test t modifié semble remplir les différentes conditions nécessaires pour guider la décision de considérer un score observé comme s'écartant significativement du score attendu et pour estimer la fréquence du score observé dans la population dont l'échantillon est issu. Il est aussi plus robuste que le score z, car il maintient davantage le taux de faux positifs à 5%. Toutefois, il est assez peu utilisé dans la pratique clinique, peut-être en raison du temps nécessaire à son application. Par contre, le score z et le percentile sont très répandus. Dès lors, afin de guider la/le neuropsychologue dans les critères de décision à appliquer selon le type de données normatives à sa disposition, nous proposons l'arbre décisionnel suivant (voir figure 2). Il s'agit d'abord de déterminer l'outil statistique qui va être utilisé pour estimer si le score est hors seuil ou non. La plupart du temps, cet outil statistique est directement fixé par les données normatives communiquées (c'est le cas des percentiles et des cut-off scores). Dans d'autres cas, lorsque la moyenne et l'écart-type sont disponibles pour chaque sous-groupe de référence, deux options sont possibles. Si l'effectif de chaque sous-groupe est bien renseigné, le test t modifié est recommandé. Dans le cas contraire, le score z est utilisé par défaut. Dans ce cas, étant donné l'inconnue quant à la taille de l'effectif, nous préconisons l'utilisation d'un seuil conservateur (un score z de ±2,05 en unilatéral, celui-ci correspondant à la p valeur égale à 0,02 et qui est le seuil utilisé lorsque la distribution s'écarte fortement d'une distribution normale tant en termes de symétrie que d'aplatissement). Dans le cas du test t modifié, un seuil à p ≤≤0,05 pourrait être conservé uniquement en cas d'informations nous rassurant sur la symétrie et l'absence d'aplatissement anormale. Dans le cas contraire,

un seuil de ≤≤0,02 est plus prudent pour garder la probabilité d'erreur au seuil des 5%. À nouveau, rappelons qu'un raisonnement bilatéral impose d'adapter nos critères.

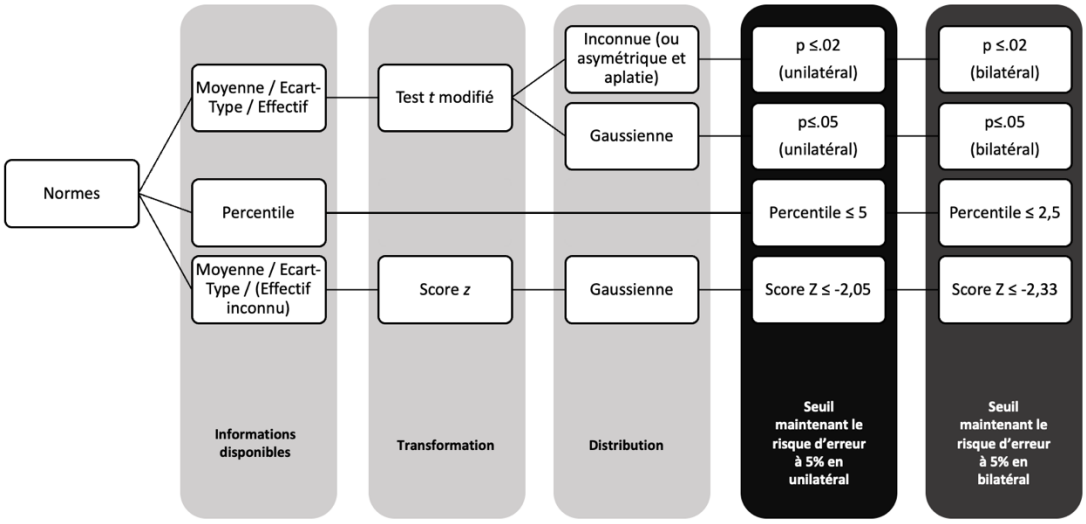

Figure 2 : Arbre décisionnel selon l'outil statistique utilisé.

Cet arbre décisionnel est construit en supposant que les conditions suivantes soient rencontrées : Premièrement, il suppose un score qui est de bas niveau s'il est inférieur à la moyenne (un nombre de réponses correctes par exemple). Si le score est d'autant plus faible qu'il a une valeur élevée (comme un temps de réaction ou un nombre d'erreurs), le signe du score *z* doit être adapté et considéré comme positif. Pour le percentile, lorsque le score est d'autant plus faible qu'il a une valeur élevée, nous proposons dans nos rapports de rendre sa lecture plus aisée en déduisant le percentile de 100 (P100-P95=P5), de telle sorte que dans tous les cas ce sont les seuils P ≤5 et P ≤2,5 qui servent de critère de décision.

5. Au-delà du score seuil : considérations lors de l'interprétation

5.1. Le nombre de tests administrés

Les évaluations neuropsychologiques complètes impliquent souvent l'administration de plusieurs tests qui fournissent de nombreux scores. À titre d'exemple, la batterie NAB (*Neuropsychological Assessment Battery* ; Stern & White, 2003, voir aussi Crum et al., 2023), comprend 36 tests, produit 52 scores primaires, 5 scores composites, 1 score récapitulatif, et plus de 52 scores secondaires. Le RL/RI-16, largement employé dans la pratique clinique, génère 11 scores différents (le rappel immédiat, les 4 rappels libres et 4 rappels totaux, la reconnaissance et fausse reconnaissance). Si la multiplicité des tests dans les évaluations neuropsychologiques est essentielle pour obtenir un état des lieux exhaustif et approfondi du fonctionnement cognitif d'un individu, elle n'est pas sans conséquences sur les résultats observés. En effet, à mesure que le nombre de tests administrés et le nombre de scores

analysés augmentent, la probabilité d'obtenir des résultats s'écartant significativement de la moyenne de l'échantillon normatif augmente également (e.g., Brooks et al., 2007, 2008 ; Godefroy et al., 2014, Ingraham & Aiken, 1996). Ce phénomène d'apparition de scores hors-seuil dans des ensembles de plusieurs tests a été illustré à plusieurs reprises (voir Annexe 7). Il peut être attribué à divers facteurs, parmi lesquels figure la variabilité normale des scores. Les performances cognitives d'un individu sont en effet susceptibles de fluctuer avec la fatigue, la motivation, l'engagement dans la tâche, les conditions émotionnelles, médicales (traitements, douleurs, fluctuations des symptômes…) ou d'autres facteurs transitoires. Avec l'augmentation du nombre de tests administrés, ces fluctuations normales peuvent se détériorer et provoquer des déclins de performance. Notons que certains tests peuvent être plus sensibles à certains facteurs que d'autres : par exemple, certains tests pourraient être particulièrement affectés par la fatigue alors que d'autres pourraient être plus sensibles au niveau d'engagement ou d'anxiété.

Sur le plan statistique, le cumul des chances d'erreur (ou « inflation du risque alpha ») contribue également à l'accroissement du risque d'erreur total avec le nombre de tests administrés (Binder et al., 2009 ; Decker et al., 2012 ; Roussel & Godefroy, 2016). En effet, chaque test individuel comporte une certaine marge d'erreur et lorsque plusieurs tests sont administrés, ces marges d'erreur s'additionnent. Concrètement, lorsque nous effectuons plusieurs tests, chaque test individuel comporte un risque de commettre une erreur de type I (« alpha »), dont le niveau admis est généralement de 5% (maximum). Ainsi, si nous effectuons 10 tests indépendants, avec un niveau de signification alpha de 0.05 chacun, le risque *global* d'obtenir au moins une erreur de type I augmente considérablement : 1- (1 - 0.05)^10 = 1-(0.95)^10 ≈ 40%, ce qui est bien supérieur à 5%. À mesure que le nombre de tests effectués augmente, le risque de classer à tort un individu augmente également par la multiplication des opportunités d'erreur, liées à l'accumulation des marges d'erreur individuelles, comme l'illustre le tableau 1.

Tableau 1. L'inflation du risque d'erreur en fonction du nombre de tests

Nombre de tests (seuil alpha = 5%)	Risque global d'erreur
1	5%
2	9.75%
5	22.62
10	40.13%
50	92.31%
k	$1- (1 - 0.05)^k$

Pour atténuer ce risque, plusieurs solutions peuvent être envisagées. Par exemple, en recherche, l'application de certaines procédures de correction (par exemple, la correction de Bonferroni) permet d'ajuster le seuil de signification en fonction du nombre de tests effectués. Cela peut toutefois augmenter le risque de perte de sensibilité avec plus de « faux négatifs » (ou erreur de type II, se produisant lorsqu'on ne rejette pas une hypothèse nulle qui est en réalité fausse). En clinique, lorsque plusieurs scores évaluent la même fonction, certaines batteries reposent sur des procédures d'agrégation de scores multiples en un score global (comme la somme des scores, la moyenne de scores *z* ou le dénombrement des scores hors-seuil avec ajustement de seuil ; voir Roussel & Godefroy, 2016). Toutefois, on peut alors regretter la perte d'informations spécifiques à un test. Une autre solution à nouveau raisonnable pourrait être de choisir un seuil de risque d'erreur plus conservateur lorsque de multiples tests et scores sont utilisés (encore plus sans hypothèses a priori). Cet argument conforte l'utilisation de scores seuils plus conservateurs ainsi que proposé dans la figure 1.

Avant de poursuivre, notons que l'ajustement des seuils de décision, basé sur le nombre de scores analysés, ne signifie pas que l'interprétation clinique des résultats doit être dissociée de la réalité clinique de la personne évaluée, y compris sa plainte et ses difficultés au quotidien. La prudence avec laquelle le neuropsychologue fixe son seuil et l'interprétation qu'il fait des résultats (de façon holistique et personnalisée) sont deux étapes différentes du processus d'évaluation. Un score qui n'est pas considéré comme « hors-seuil » mais qui est faible pourrait ne pas être problématique pour certains patients, tout en représentant une baisse significative pour d'autres ou engendrer des difficultés importantes dans leur vie quotidienne, en fonction des exigences de leur vie professionnelle, de leur mode de vie, ou

d'un niveau antérieur élevé de performance. En tout état de cause, l'interprétation clinique ne se résume pas à la simple lecture des scores obtenus aux tests, mais les utilise dans une appréciation plus globale et nuancée. In fine, ce n'est pas tant le test qui est « inférentiel » mais bien le clinicien. Le seuil de décision statistique ne présuppose pas automatiquement la future décision du clinicien, son jugement et ses préconisations. Dans cet esprit, les éventuelles hypothèses préalables que pourrait avoir le clinicien concernant les difficultés attendues chez un patient influencent l'interprétation clinique en plus des seuils de décisions statistiques.

5.2. Hypothèses a priori

Le risque d'erreur en statistique dépend également des prédictions que l'on peut émettre en fonction de connaissances antérieures. Lorsque des prédictions sont réalisées en fonction de connaissances préalables, par exemple sur base des troubles cognitifs attendus dans une pathologie, la/le neuropsychologue possède des "a priori" statistiques. À l'inverse, lorsqu'elle/il ne possède pas d'a priori, elle/il se trouve dans une situation qualifiée de post-hoc, c'est-à-dire où la décision est prise uniquement sur base de ce qui est observé, sans attentes préalables. Le fait de posséder des hypothèses a priori permet théoriquement de diminuer le risque d'erreur et de garder un score seuil plus libéral. Par exemple, si la/le neuropsychologue doit se prononcer sur la possibilité d'une pathologie neurodégénérative, elle/il peut se baser sur les connaissances antérieures liées à cette pathologie ainsi que sur les données d'hétéro-anamnèse pour ne pas avoir à corriger son risque d'erreur. Ainsi face à une association spécifique de scores sous un seuil plus libéral (par exemple, un score *z* de 1,65) mais consistante avec le tableau clinique, il sera considéré comme moins risqué de ne pas corriger le seuil de décision puisque des attentes a priori concernant la pathologie sont disponibles et rencontrées. Le même raisonnement est valable lorsque le clinicien dispose déjà d'éléments significatifs (c'est-à-dire, diagnostic établi, données antérieures, rapport d'imagerie cérébrale …), qui lui permettent de formuler des hypothèses précises sur les altérations cognitives attendues chez la personne évaluée. Bien entendu, cette logique de non-correction du seuil ne pourra être appliquée que lorsque des connaissances suffisantes sont disponibles a priori. Dans le cas contraire, la/le neuropsychologue devra faire preuve de prudence lorsqu'elle/il multiplie les tâches et les scores, et privilégier l'utilisation d'un seuil de décision (post-hoc) plus strict.

5.3. Qualité des tests et des échantillons normatifs

Bien entendu, un seuil de risque d'erreur conservateur ne garantit pas la réduction des erreurs à lui seul. À titre d'exemple, une décision basée sur un test neuropsychologique de mauvaise qualité ou sur un échantillon normatif inadapté, ne pourra être elle-même valide. La qualité du test neuropsychologique et la pertinence de son utilisation est déterminée au regard de multiples données nous apportant des éléments de preuves de validité, de fidélité, spécificité et sensibilité spécifiques aux situations et populations cliniques dont sont issues ces données (e.g., Laveault & Grégoire, 2023). Aucun outil statistique de prise de décision n'a d'impact sur ces éléments qui dépendent de la construction du test et de la validité de son utilisation. À l'inverse, un test possédant de faibles qualités psychométriques risquera d'entraîner une prise de décision inadaptée. En réalité, les erreurs de mesure sont inhérentes à tout processus d'évaluation. Malgré les efforts déployés pour concevoir des instruments fiables et valides, il existe toujours une possibilité que les résultats obtenus ne reflètent pas parfaitement les véritables capacités ou caractéristiques de la personne évaluée. Un regard critique sur les outils disponibles, reconnaissant leurs limites et leurs marges d'erreurs intrinsèques, est essentiel pour une interprétation adéquate des résultats.

En plus de la qualité psychométrique des outils, la validité de la démarche est aussi tributaire de la qualité de l'échantillon normatif à partir duquel l'épreuve a été normalisée. En particulier, si l'interprétation ne peut pas se faire dans le contexte d'un niveau prémorbide ou antérieur, certains points de vigilance doivent également être observés. Comment les personnes constituant l'échantillon de référence ont-elles été recrutées ? Le nombre de sujets est-il suffisamment important pour se prémunir des effets d'échantillonnage ? Intègre-t-il suffisamment de variabilité au plan socio-démographique, incluant les groupes généralement moins représentés (comme les personnes très âgées ou avec un faible niveau scolaire) ? Au final, l'échantillon est-il représentatif de la population qu'il est supposé représenter ? Certains critères d'exclusion ont-ils été conçus pour exclure les personnes présentant une pathologie dans l'échantillon normatif (s'agit-il d'un échantillon de sujets sains ou tout venants) ? Quelle est la fréquence des scores faibles au sein de l'échantillon normatif ? Les scores obtenus par l'échantillon normatif sont-ils récents et reflètent-ils les caractéristiques actuelles de la population ? La distribution des performances dévie-t-elle de la normale ?

Même si les scores sont toujours intégrés dans le contexte clinique global de la personne évaluée, en tenant compte de l'anamnèse, des observations comportementales, de l'examen

médical et d'autres informations pertinentes, une posture prudente consiste à se montrer particulièrement vigilant au risque de faux positif lorsque l'échantillon normatif présente certaines limites ou caractéristiques. À titre d'exemple, une prise de décision basée sur un échantillon normatif issu d'une autre culture que celle de la personne évaluée sera à considérer avec beaucoup de prudence (Fujii, 2018 ; Franzen et al., 2022 ; Pedraza & Mungas, 2008).

6. Choix du vocable pour communiquer les scores

Le raisonnement sous-tendant l'interprétation d'un score est donc complexe et explique d'autant plus la difficulté de le communiquer en termes simples. Le rapport écrit est en effet consulté par différents intervenants et chaque fois que cela est possible par la personne évaluée. Les rapports peuvent être ainsi lus par des professionnels très variés allant des médecins (spécialisés ou non spécialisés) aux juges dans le contexte d'une expertise médico-légale. Le rapport doit donc être suffisamment pédagogique pour être lu par ce public très varié. À défaut de clarté, il risque d'être lu sans être compris, ou de ne pas être lu et de ne pas participer aux décisions concernant la personne évaluée. Or, plusieurs études ont montré que les équipes interdisciplinaires estiment nos rapports difficilement lisibles, bien que précieux pour la planification des soins (p. ex., Postal et al., 2018). Les neuropsychologues sont, comme tous les professionnels dont la pratique repose sur des savoirs spécialisés, confrontés au problème de la transmission de leurs observations et de leurs conclusions dans un langage compréhensible pour les personnes auxquelles elles sont destinées. Dans ce contexte, l'un des principaux défis à relever touche à la communication des résultats eux-mêmes. C'est ainsi que, au-delà du score seuil qui doit être clarifié et communiqué dans les rapports, la/le neuropsychologue doit qualifier verbalement le score observé en commentant sa fréquence dans l'échantillon de référence.

Le vocable utilisé dans cette optique ne fait pas l'objet d'une définition consensuelle. Par ailleurs, dans un grand nombre de cas, les termes introduisent une ambiguïté entre le raisonnement statistique autour d'un score et l'interprétation clinique concernant l'intégrité de la fonction sous-jacente. Par exemple, l'utilisation du terme « performance déficitaire » pourrait suggérer une altération de la fonction cognitive sous-jacente, sur la seule base d'un score s'écartant des attentes normatives, sans tenir compte du profil global des résultats aux différents tests, ainsi que du contexte et des spécificités de la personne. Une telle approche

ne serait bien évidemment pas considérée comme une méthode acceptable pour parvenir à des conclusions cliniques. En effet, il est bien connu qu'un grand nombre de facteurs sont susceptibles d'occasionner des fluctuations « normales » de la performance, aussi bien à un niveau interindividuel qu'intra-individuel, sans oublier les facteurs statistiques (cf. supra). Pour toutes ces raisons, le groupe de consensus de l'Académie des Neuropsychologues américains (AACN ; Guilmette et al., 2020), suggère d'utiliser des qualificatifs qui concernent les scores en tant que tels, et leur comparaison aux normes et non l'état postulé de la fonction cognitive. Ainsi, le qualificatif pourrait refléter la logique de la comparaison normative, autrement dit la probabilité estimée d'obtenir ce score dans la population ou l'incertitude associée au rejet de l'hypothèse nulle (cf. supra). Mentionner qu'un score suggère une altération pathologique ne devrait se faire qu'au terme d'une interprétation globale de l'ensemble des données à disposition, autrement dit dans la conclusion du rapport. C'est donc pour la conclusion que la/le neuropsychologue devrait réserver l'utilisation de termes tels que « déficitaire », « altéré » ou « pathologique » pour qualifier un processus cognitif (et non un score).

Une classification et une nomenclature simple et explicite devraient être utilisées pour favoriser une communication consensuelle et sans ambiguïté. Dans cet objectif, le groupe de consensus de l'AACN (Guilmette et al., 2020) propose des qualificatifs mettant l'accent sur la probabilité d'obtenir le score : « score exceptionnellement bas » ; « score inférieur à la moyenne » ; « score dans la moyenne faible » ; « score dans la moyenne » ; etc. avec une division en 7 catégories. Toutefois, leur distinction entre une catégorie « score dans la moyenne » de deux autres « score dans la moyenne faible » et « score dans la moyenne élevée » pour des scores qui ne sont dans aucun cas « hors-seuil » nous semble peu informative et source de confusion. Ces catégories pourraient en effet conduire à une augmentation de la prise de risque dans l'interprétation statistique. Dans ce contexte, il nous semble plus prudent d'utiliser une division en **5 catégories** visant à donner un statut singulier uniquement aux scores associés à des probabilités théoriques en deçà des 5% et 2 % de risque d'erreur. Soulignons encore que la probabilité de 5% n'impliquant pas ipso facto un score hors seuil (cf. supra). Par ailleurs, un score associé à un risque d'erreur de 10% peut être interprété ultérieurement comme reflétant une déficience dans une démarche diagnostique, ou un déclin cognitif lors d'un suivi longitudinal des performances cognitives d'un même patient.

Un autre point de considération concerne l'emploi des termes « faible » et « fort » qui s'écarte de cette logique visant à éviter, à ce stade du raisonnement, toute interprétation qualitative ou clinique. En outre, cette terminologie peut prêter à confusion selon qu'il s'agit de réponses correctes, d'erreurs ou des temps de réaction. Pour cette raison, une transformation des données plaçant la « faible performance » du côté bas et la « bonne performance » du côté haut permettrait une plus grande lisibilité. En conséquence, nous proposons qu'un score z de +1,96 pour un taux d'erreurs, soit transformé en score z de -1,96. Nous prônons également l'utilisation de la terminologie moins ambiguë « score de haut niveau » et « score de bas niveau ». À nouveau, la terminologie « haut » et « bas » n'implique pas ipso facto un score hors seuil. Cette terminologie a l'avantage de découpler la position du score de sa valeur numérique. Concrètement, l'expression « score de bas niveau » sied tant pour un nombre élevé d'erreurs, une vitesse de réponse lente que pour un nombre bas de réponses correctes. De même, l'expression « score de haut niveau » sied tant pour un temps de réponse rapide que pour un nombre élevé de réponses correctes et un nombre faible d'erreurs.

Une proposition alternative a été discutée : score « faible » versus score « élevé ». Selon le groupe de travail, cette expression entretient potentiellement une ambiguïté lorsque la valeur numérique du score est en contradiction avec la qualification qui lui sera attribuée. Par exemple, il n'y a pas d'ambiguïté lorsqu'un nombre bas de réponses correctes est qualifié de score faible (effectivement, numériquement, le score est faible). En revanche, il peut y en avoir une lorsqu'un temps de réponse élevé ou un nombre élevé d'erreurs sont qualifiés de faibles (ou un nombre bas d'erreurs est qualifié de score élevé).

Un autre point qui a été soulevé concerne la compréhension de ces expressions par des personnes qui lisent les évaluations cognitives mais dont la neuropsychologie n'est pas la discipline (par exemple, un magistrat). Si nous observons le monde qui nous entoure, il y a au moins un domaine dans lequel l'expression « haut niveau » est utilisée. Le sport désigne un athlète comme de « haut niveau » quelle que soit la valeur numérique élevée ou basse de ses scores. Un·e athlète dont le record de saut à la perche est élevé en nombres de mètres est qualifié·e « de haut niveau ». De même un·e athlète dont le record au 100 mètres est bas en nombre de secondes est qualifié·e « de haut niveau ». La communauté semble comprendre cette expression dans ce contexte.

Enfin, tout comme le groupe de consensus de l'AACN, nous proposons de mettre l'accent sur la probabilité d'obtenir le score dans la population. Toutefois, nous proposons de

ne pas reprendre le terme « exceptionnel » étant donné son double sens dans la langue française et le risque de confusion associé. Ainsi, nous proposons les vocables « fréquent », « peu fréquent », « très peu fréquent ».

Tableau 2. Proposition de qualificatifs associés à 5 catégories de scores

Fréquence estimée de la population ayant ce score ou moins (*****= risque d'erreur associé à la décision de qualifier un score de hors-normes)	**Label**	**Seuil**
≧ 98%	Score **très peu fréquent** de haut niveau	
Entre ≧ 95% et ＜ 98%	Score **peu fréquent** de haut niveau	
Entre ＞ 5 % et ＜ 95 %	Score **fréquent**	Score à interpréter ultérieurement en fonction du niveau de performance présupposée
Entre ＞ 2% et ≦ 5 %	Score **peu fréquent** de bas niveau	Score jugé hors norme avec un **seuil libéral**
≦ 2%	Score **très peu fréquent** de bas niveau	Score jugé hors norme avec un **seuil conservateur**

**Note. ***Valeur p associée au t modifié, probabilité associée au score *z* et percentile.
Notons que la valeur p associée au t modifié est un estimateur plus précis du risque d'erreur que la probabilité associée au score *z* puisque ce premier tient compte de l'effectif de l'échantillon normatif. Aussi, le percentile n'est pas un réel estimateur probabiliste du risque d'erreur, mais plutôt une fréquence du score dans l'échantillon normatif.
Lorsque l'on s'intéresse aux scores bas et élevés (comme dans le cas d'un Quotient Intellectuel), le risque d'erreur pour les scores de haut niveau est similaire à celui des scores de bas niveau.

En conclusion, le groupe de travail propose l'adoption d'un ensemble de 5 qualificatifs avec une communication explicite du score seuil utilisé. Comme illustré dans les tableaux en annexe (voir Annexe 8), ces qualificatifs peuvent être présentés dans le rapport dans une table résumant les outils statistiques associés. En outre, pour faciliter la lecture, il nous apparaît opportun d'accompagner cette table d'une brève explication du principe de la comparaison normative (Willems & Seron, 2023).

Cette proposition de catégories de scores et des qualificatifs associés devrait être adaptée dans plusieurs contextes. Prenons le cas des épreuves donnant un éventail de scores bruts très restreint avec des résultats élevés obtenus par la plupart des personnes en bonne santé. C'est le cas de tests ou d'indices conçus pour identifier des difficultés (et non des performances supérieures à la moyenne), où les meilleurs résultats sont associés à une large marge de probabilité. À titre d'exemple, on observe ces fourchettes restreintes de scores pour le nombre d'erreurs dans certaines épreuves (ex. le test de dénomination du Stroop où une absence d'erreur est observée pour 90 % et plus de l'échantillon normatif), dans un grand nombre de tâches de reconnaissance (ex., tâche de reconnaissance du Brief Visuospatial Memory Test) et dans d'autres tâches habituellement très bien réussies avec un effet plafond. Dans ce cas, nous recommandons aux neuropsychologues de ne pas utiliser la terminologie « de haut niveau » pour désigner les scores supérieurs. Seuls les scores associés à une probabilité inférieure à 2% pourraient être dit « très peu fréquents » alors que les scores associés à une large fourchette de probabilités (ex., percentile 10-99) pourraient être simplement qualifiés de fréquents, signifiant ainsi simplement qu'ils sont conformes aux attentes. Cette proposition a comme avantage d'utiliser un langage commun avec les labels proposés pour les autres tests.

Enfin, nous n'aborderons pas ici la question du score cut-off mais le lecteur intéressé peut se tourner vers l'annexe 1.

7. Illustrations cliniques

Afin d'illustrer la logique développée dans ce document, envisageons quelques situations cliniques fréquentes en neuropsychologie.

Illustration 1 - suspicion de trouble neurocognitif majeur ayant pour étiologie une maladie d'Alzheimer

Un patient francophone de 80 ans, ancien enseignant du secondaire ayant réalisé des études universitaires, est envoyé chez une neuropsychologue par une neurologue afin de contribuer à une mise au point diagnostique dans le cadre d'une suspicion de trouble neurocognitif majeur (ayant pour étiologie une probable maladie d'Alzheimer). La demande fait suite à des plaintes de la personne évaluée et de son entourage concernant des pertes de mémoire de plus en plus fréquentes et invalidantes. Un bilan neuropsychologique investiguant de

nombreuses fonctions cognitives est réalisé. Le bilan est composé de 12 tâches et comprend 55 scores associés à des normes.

Question 1 : unilatéral ou bilatéral ? Au vu de la question posée, l'identification de scores (très) peu fréquents de haut niveau n'est pas considérée comme pertinente et la neuropsychologue décide de se concentrer uniquement sur les scores (très) peu fréquents de bas niveau. Elle décide donc de se situer en unilatéral sur le plan statistique.

Question 2 : a priori ou post-hoc ? Au vu de la littérature abondante sur les troubles cognitifs associés à la maladie d'Alzheimer, la neuropsychologue peut émettre des prédictions fortes sur les sphères cognitives à risque de détérioration dans une maladie d'Alzheimer débutante (de forme amnésique typique au vu des plaintes de la personne évaluée et de son entourage) : la mémoire épisodique (très à risque), les fonctions exécutives (à risque) et la mémoire sémantique (à risque). Le fait de posséder des a priori lui permettrait de ne pas devoir corriger son seuil de risque d'erreur. Elle décide donc de garder un seuil libéral de risque d'erreur pour ces fonctions malgré les nombreux tests réalisés. Pour les fonctions cognitives non concernées par ces a priori, la neuropsychologue décide par contre d'adopter d'emblée un seuil de risque d'erreur plus conservateur vu le nombre de scores concernés.

Question 3 : informations normatives disponibles ? Pour certains tests, la neuropsychologue dispose de la moyenne, de l'écart-type et de la taille de l'échantillon normatif (10 sujets par cellule de l'échantillon normatif). Elle dispose d'informations la rassurant sur la normalité et l'aplatissement de la courbe des données. Pour ces tests, la neuropsychologue décide donc d'utiliser un test *t* modifié, en unilatéral avec un seuil de risque d'erreur libéral ($p \leq\leq 0,05$). Plusieurs scores du test de mémoire épisodique obtiennent un $p \leq\leq 0,05$ en unilatéral conformément à ses hypothèses a priori et sont donc qualifiés comme des scores peu fréquents de bas niveau hors seuil.

Pour d'autres tests, la neuropsychologue dispose uniquement de la moyenne et de l'écart type. La neuropsychologue décide ici d'utiliser un score *z* avec un seuil de risque d'erreur conservateur en unilatéral ($z \leq\leq -2,05$). Plusieurs scores obtenus avec des tâches évaluant les fonctions exécutives donnent des $z \leq\leq -2,05$ et sont donc considérés comme des scores très peu fréquents de bas niveau hors seuil.

Enfin, pour certains tests la neuropsychologue dispose uniquement des percentiles obtenus sur des échantillons de plus de 30 participants. Pour ces tests, la neuropsychologue décide d'utiliser un percentile en unilatéral (percentile $\leq\leq 5$) puisqu'il s'agit de percentiles réels non

influencés par la distribution des données. La tâche de fluences sémantiques obtient un score équivalent à un percentile 4 et est donc considéré comme score peu fréquent de bas niveau hors seuil.

Pour rédiger ses conclusions, la neuropsychologue tiendra compte de ces scores (très) peu fréquents de bas niveau hors seuil mais intégrera également un nombre important d'informations supplémentaires, qualitatives et quantitatives, jugées pertinentes pour la question posée (ex : plaintes spontanées, chronologie et évolution des difficultés cognitives dans le temps, comportement de la personne évaluée durant l'examen, informations disponibles sur l'état émotionnel de la personne évaluée, hétéro-anamnèse, antécédents médicaux, médication en cours, bilans neuropsychologiques réalisés par le passé).

Illustration 2 - difficultés scolaires avec suspicion de trouble de déficit de l'attention avec ou sans hyperactivité (TDAH)

Une patiente francophone de 9 ans, scolarisée en 3ème de l'enseignement normal (CE2), est envoyée chez un neuropsychologue par un pédopsychiatre afin d'améliorer la compréhension de difficultés scolaires pour lesquelles une suspicion de TDAH est avancée. Un bilan neuropsychologique investiguant de nombreuses fonctions cognitives est réalisé. Le bilan est composé de 16 tâches (en plus de l'évaluation du QI) et comprend 74 scores associés à des normes. Le test de QI à lui seul comprend 10 tâches et 16 scores additionnels.

Question 1 : unilatéral ou bilatéral ? Au vu de la question posée, l'identification de scores (très) peu fréquents de haut niveau est considérée comme pertinente pour la question clinique posée (par exemple, l'évaluation intellectuelle où l'identification de forces est importante pour la question scolaire) et le neuropsychologue décide de se concentrer à la fois sur les scores (très) peu fréquents de bas et de haut niveau. Il décide donc de se situer en bilatéral sur le plan statistique.

Question 2 : a priori ou post-hoc ? Au vu de la littérature abondante sur les troubles cognitifs associés au TDAH, le neuropsychologue peut émettre des prédictions fortes sur les sphères cognitives à risque d'altération dans un TDAH : l'administrateur central de la mémoire de travail (à risque), les fonctions exécutives incluant notamment l'inhibition et les signes d'impulsivité (très à risque), ainsi que la planification (à risque), les fonctions attentionnelles et particulièrement les fluctuations des temps de réponse dans les tâches longues et monotones (très à risque) et l'attention divisée (à risque). Le fait de posséder des hypothèses a priori permet au neuropsychologue de ne pas devoir corriger son seuil de risque d'erreur. Il

décide donc de garder un seuil de 5% de risque d'erreur (en bilatéral) pour ces fonctions malgré les nombreux tests réalisés. Pour les fonctions cognitives non concernées par ces a priori, le neuropsychologue décide d'adopter un seuil de risque d'erreur similaire mais de considérer ces scores avec nettement plus de prudence puisque ne disposant pas d'a priori les concernant.

Question 3 : informations normatives disponibles ? Pour certains tests, le neuropsychologue dispose de la moyenne, de l'écart-type et de la taille de l'échantillon normatif. Par contre, il ne dispose pas d'information sur la normalité et l'aplatissement de la courbe des données. Pour ces tests, le neuropsychologue décide d'utiliser un test *t* modifié, en bilatéral avec un critère conservateur ($p \leq\leq 0{,}02$). Plusieurs scores de tâches attentionnelles monotones évaluant les fluctuations de réponse obtiennent un $p \leq\leq 0{,}02$ en bilatéral et sont donc considérés comme des scores très peu fréquents de bas niveau hors seuil.

Pour certains tests, le neuropsychologue dispose uniquement de la moyenne et de l'écart type. Pour ces tests, le neuropsychologue décide d'utiliser un score *z* en bilatéral ($z \leq\leq -2{,}33$ ou $z \geq\geq 2{,}33$). Plusieurs scores dans les tâches exécutives d'inhibition et de planification obtiennent un $z \leq\leq -2{,}33$ et sont donc considérés comme des scores très peu fréquents de bas niveau hors seuil. Pour les tâches évaluant des fonctions non concernées par les a priori, le neuropsychologue décide d'utiliser le même seuil mais de considérer ces scores avec prudence. Certains scores de tâches d'attention sélective obtiennent un $z \leq\leq -2{,}33$ et sont donc considérés comme des scores très peu fréquents de bas niveau potentiellement hors seuil. Pour le test de QI, le neuropsychologue dispose de la moyenne et de l'écart-type pour les indices et les tâches. Étant donné que des scores peu fréquents dans les tests de QI ne font pas partie des a priori dans le TDAH, le neuropsychologue décide d'utiliser un seuil de risque d'erreur similaire (score $z \leq\leq -2{,}33$ ou $z \geq\geq 2{,}33$) mais de considérer également les scores avec prudence. La patiente obtient plusieurs tâches et indices avec des scores $z \geq\geq 2{,}33$ qui sont donc considérés comme des scores très peu fréquents de haut niveau potentiellement hors seuil.

Enfin, pour certains tests, le neuropsychologue dispose uniquement des percentiles. Pour ces tests, le neuropsychologue décide d'utiliser un percentile en bilatéral, avec un seuil de risque d'erreur de 5% (percentile $\leq\leq 2{,}5$ ou $\geq\geq 97{,}5$). Pour les tâches évaluant des fonctions non concernées par les a priori, le neuropsychologue décide d'utiliser un seuil de risque d'erreur similaire mais en les considérant avec prudence. Aucune tâche disposant uniquement de

percentiles n'est considérée comme ayant des scores très peu fréquents de bas ou de haut niveau.

Pour rédiger ses conclusions, le neuropsychologue tiendra compte de ces scores peu ou très peu fréquents de bas et de haut niveau hors seuil mais intégrera également un nombre important d'informations supplémentaires, qualitatives et quantitatives, jugées pertinentes pour la question posée (ex : plaintes spontanées, bulletins scolaires, informations données par les enseignant.e.s et journaux de classe, chronologie et évolution des difficultés cognitives dans le temps, développement dans la petite enfance, comportement de la patiente durant l'examen, informations disponibles sur l'état émotionnel de la patiente, hétéro-anamnèse auprès des parents, antécédents médicaux et familiaux, éventuelle médication en cours, bilans neuropsychologiques réalisés par le passé, questionnaires sur le TDAH et sur les comorbidités éventuelles). Les scores potentiellement hors seuil seront également intégrés avec prudence dans les conclusions.

Illustration 3 - bilan de première ligne chez un patient avec plaintes spontanées, sans aucun antécédent ni suspicion de syndrome clinique identifié

Un patient francophone de 37 ans, disposant d'un diplôme d'enseignement supérieur de type court[5] en informatique consulte une neuropsychologue parce qu'il a vu une émission à la télévision sur la neuropsychologie et souhaiterait mieux se comprendre. Le patient n'a pas de plaintes cognitives clairement identifiées si ce n'est qu'il se sent un peu différent des autres (mais son propos n'est pas clair quant à ses différences) et qu'il a vécu trois épisodes de burnout ces quinze dernières années. Un bilan neuropsychologique investiguant de nombreuses fonctions cognitives est réalisé. Le bilan est composé de 18 tâches (hors test de QI) et comprend 79 scores associés à des normes. Le test de QI à lui seul comprend 10 tâches et 15 scores additionnels.

Question 1 : unilatéral ou bilatéral ? Au vu du caractère peu spécifique de la question posée, l'identification de scores (très) peu fréquents de haut niveau est considérée comme pertinente pour la question clinique posée (par exemple, au niveau de l'évaluation intellectuelle où l'identification de forces pourrait orienter la conclusion). La neuropsychologue décide de se concentrer à la fois sur les scores (très) peu fréquents de bas et de haut niveau. Elle décide donc de se situer en bilatéral sur le plan statistique.

[5] Équivalant d'une licence professionnelle en France.

Question 2 : a priori ou post-hoc ? Le caractère général de la question posée ne permet pas d'avoir des a priori suffisamment forts sur le plan statistique. La neuropsychologue décide d'adopter un seuil de risque d'erreur conservateur et de considérer les scores obtenus avec prudence vu le risque important de faux positifs.

Question 3 : informations normatives disponibles ? Pour certains tests, la neuropsychologue dispose de la moyenne, de l'écart-type et de la taille de l'échantillon normatif. La neuropsychologue décide d'utiliser un test *t* modifié, en bilatéral. Certains scores de tâches attentionnelles monotones évaluant les fluctuations de réponse obtiennent un p ≤≤0,02 en bilatéral et sont donc considérés comme des scores très peu fréquents de bas niveau potentiellement hors seuil.

Pour certains tests, la neuropsychologue dispose uniquement de la moyenne et de l'écart type. Pour ces tests, la neuropsychologue décide d'utiliser un seuil de risque d'erreur en bilatéral ($z \leq$ - 2,33 ou $z \geq$2,33). Certains scores de tâches exécutives et de théorie de l'esprit obtiennent un $z \leq$-2,33 et sont donc considérés comme des scores très peu fréquents de bas niveau potentiellement hors seuil. Pour le test de QI, la neuropsychologue dispose de la moyenne et de l'écart-type pour les indices et les tâches. La neuropsychologue décide d'utiliser un seuil de risque d'erreur en bilatéral ($z \leq$ - 2,33 ou $z \geq\geq$2,33). Le patient obtient plusieurs tâches et indices avec des scores $z \geq\geq$2,33 qui sont donc considérés comme des scores très peu fréquents de haut niveau potentiellement hors seuil.

Enfin, pour certains tests la neuropsychologue dispose uniquement des percentiles. Pour ces tests, la neuropsychologue décide d'utiliser un seuil de risque d'erreur de 5% en bilatéral (percentile ≤2,5 ou ≥≥97,5). Aucune tâche disposant uniquement de percentiles n'est considérée comme ayant des scores très peu fréquents de bas ou de haut niveau.

Pour rédiger sa conclusion, la neuropsychologue tiendra compte de ces scores très peu fréquents de bas et de haut niveau potentiellement hors seuil, avec prudence vu le risque plus important de faux positifs en raison du manque d'a priori. Elle intégrera également un nombre important d'informations supplémentaires, qualitatives et quantitatives, jugées pertinentes pour la question posée (ex : plaintes spontanées, chronologie et évolution des difficultés cognitives dans le temps, comportement de la personne évaluée durant l'examen, informations disponibles sur l'état émotionnel de la personne évaluée, antécédents médicaux et familiaux, éventuelle médication en cours, questionnaires de dépistage de troubles et syndromes psychiatriques).

Références

Beauchamp, M. H., Brooks, B. L., Barrowman, N., Aglipay, M., Keightley, M., Anderson, P., ... Zemek, R. (2015). Empirical derivation and validation of a clinical case definition for neuropsychological impairment in children and adolescents. Journal of the International Neuropsychological Society, 21(8), 596–609.

Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To err is human: "Abnormal" neuropsychological scores and variability are common in healthy adults. Archives of Clinical Neuropsychology, 24, 31–46.

Bridges, A. J., & Holler, K. A. (2007). How many is enough? Determining optimal sample sizes for normative sudies in pediatric neuropsychology. Child Neuropsychology, 13, 528–538.

Brooks, B. L., Iverson, G. L., & White, T. (2007). Substantial risk of "accidental MCI" in healthy older adults: Base rates of low memory scores in neuropsychological assessment. Journal of the International Neuropsychological Society, 13, 490–500.

Brooks, B. L., Iverson, G. L., Holdnack, J. A., & Feldman, H. H. (2008). The potential for misclassification of mild cognitive impairment: A study of memory scores on the Wechsler Memory Scale-III in healthy older adults. Journal of the International Neuropsychological Society, 14, 463–478.

Brooks, L. B., Sherman, E. M. S., Iverson, G. L., Slick, D. J. & Strauss E. (2011). Psychometric Foundations for the Interpretation of Neuropsychological Test Results. In: The little Black Book of Neuropsychology, Schoenberg, M. R. & Scott, J. G. (Eds). Springer, New York.

Crawford, J. R., & Garthwaite, P. H. (2005). Testing for suspected impairments and dissociations in single-case studies in neuropsychology: evaluation of alternatives using Monte Carlo simulations and revised tests for dissociations. Neuropsychology, 19(3), 318-331.

Crawford, J. R., & Garthwaite, P. H. (2008). On the "optimal" size for normative samples in neuropsychology: capturing the uncertainty when normative data are used to quantify the standing of a neuropsychological test score. Child Neuropsychology, 14(2), 99-117.

Crawford, J. R., Garthwaite, P. H., Azzalini, A., Howell, D. C., & Laws, K. R. (2006). Testing for a deficit in single-case studies: Effects of departures from normality. Neuropsychologia, 44(4), 666-677.

Crawford, J. R., Garthwaite, P. H., & Slick, D. J. (2009) On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores. The Clinical Neuropsychologist, 23, 1173-1195.

Crawford, J. R., & Howell, D. C. (1998). Comparing an individual's test score against norms derived from small samples. The Clinical Neuropsychologist, 12, 482-486

Crum, T. A., Gontkovsky, S. T., Teichner, G., & Stern, R. A. (2023). Neuropsychological Assessment Battery (NAB) in Clinical Practice. The SAGE Handbook of Clinical Neuropsychology: Clinical Neuropsychological Assessment and Diagnosis, 368.

Decker, S. L., Schneider, W. J., & Hale, J. B. (2012). Estimating base rates of impairment in neuropsychological test batteries: A comparison of quantitative models. Archives of Clinical Neuropsychology, 27(1), 69-84.

De Rotrou, J., Wenisch, E., Chausson, C., Dray, F., Faucounau, V., & Rigaud, A. S. (2005). Accidental MCI in healthy subjects: a prospective longitudinal study. European Journal of Neurology, 12(11), 879-885.

Fery, P., Claes, T. (soumis) AUTONORMES : une application d'analyse statistique des scores obtenus à des tâches évaluant les fonctions cognitives.

Fujii, D. E. (2018). Developing a cultural context for conducting a neuropsychological evaluation with a culturally diverse client: The ECLECTIC framework. The Clinical Neuropsychologist, 32(8), 1356-1392.

Franzen, S., European Consortium on Cross-Cultural Neuropsychology (ECCroN), Watermeyer, T. J., Pomati, S., Papma, J. M., Nielsen, T. R., ... & Bekkhus-Wetterberg, P. (2022). Cross-cultural neuropsychological assessment in Europe: Position statement of the European consortium on Cross-Cultural Neuropsychology (eccron). The Clinical Neuropsychologist, 36(3), 546-557.

Godefroy, O., Diouf, M., Bigand, C., & Roussel, M. (2014) Troubles neuro- cognitifs d'intensité légère ou performances normales basses ? Rev Neuropsychol, 6, 159-162.

Godefroy, O., Gibbons, L., Diouf, M., Nyenhuis, D., Roussel, M., Black, S., Bugnicourt, J.M.; GREFEX study group (2014). Validation of an integrated method for determining cognitive ability: Implications for routine assessments and clinical trials. Cortex, 54, 51-62

Guilmette, T. J., Hagan, L., & Giuliano, A. J. (2008). Assigning qualitative descriptors to test scores in neuropsychology: Forensic implications. The Clinical Neuropsychologist, 22(1), 122–139.

Guilmette, T. J., Sweet, J. J., Hebben, N., Koltai, D., Mahone, E. M., Spiegler, B. J., ... & Conference Participants. (2020). American Academy of Clinical Neuropsychology consensus conference statement on uniform labeling of performance test scores. The Clinical Neuropsychologist, 34(3), 437-453.

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P. C., Derks, K., Dienes, *z*., ... & Hoijtink, H. (2023). A review of applications of the Bayes factor in psychological research. Psychological Methods, 28(3), 558.

Ingraham, L. J., & Aiken, C. B. (1996). An empirical approach to determining criteria for abnormality in test batteries with multiple measures. Neuropsychology, 10(1), 120.

Laveault, D., & Grégoire, J. (2023). Introduction aux théories des tests en psychologie et en sciences de l'éducation. De Boeck Supérieur.

Leclef, P., Ponchel, A., Chancenotte, S., Marey, S., & Muneaux, M. (2018). L'interprétation des scores en neuropsychologie : la tour de Babel ? Cahiers de Neuropsychologie Clinique, 5, 42-56.

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. International Review of Social Psychology, 32(1).

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of experimental social psychology, 49(4), 764-766.

Meyer, A.-C. L., Boscardin, W. J., Kwasa, J. K., & Price, R. W. (2013). Is it time to rethink how neuropsychological tests are used to diagnose mild forms of HIV-associated neurocognitive disorders? Impact of false-positive rates on prevalence and power. Neuroepidemiology, 41(3–4), 208–216.

Michael, G., & Amieva, H. (2023). Les normes : utilité et utilisation. In H. Amieva., P. Azouvi, E. Barbeau, & F. Collette, Traité de neuropsychologie clinique de l'adulte Tome 1-Évaluation. Paris, France : De Boeck Supérieur.

Palmer, B. W., Boone, K. B., Lesser, I. M., & Wohl, M. A. (1998). Base rates of "impaired" neuropsychological test performance among healthy older adults. Archives of Clinical Neuropsychology, 13(6), 503-511.

Pedraza, O., & Mungas, D. (2008). Measurement in cross-cultural neuropsychology. Neuropsychology review, 18, 184-193.

Postal, K., Chow, C., Jung, S., Erickson-Moreo, K., Geier, F., & Lanca, M. (2018). The stakeholders' project in neuropsychological report writing: A survey of neuropsychologists' and referral sources' views of neuropsychological reports. The Clinical Neuropsychologist, 32(3), 326-344.

Roussel, M., & Godefroy, O. (2016). Quand faut-il considérer un bilan neuropsychologique comme anormal ? Importance de l'interprétation des performances multiples. Amieva, H., Belin, C. & Maillet, D. (Éds.), L'Évaluation neuropsychologique : De la norme à l'exception, 11-18.

Schoenberg, M. R., & Rum, R. S. (2017). Towards reporting standards for neuropsychological study results: A proposal to minimize communication errors with standardized qualitative descriptors for normalized test scores. Clinical Neurology and Neuropsychology, 162, 72–79.

Schoenberg, M. R., Osborn, K. E., Mahone, E. M., Feigon, M., Roth, R. M., & Pliskin, N. H. (2018). Physician preferences to communicate neuropsychological results: Comparison of qualitative descriptors and a proposal to reduce communication errors. Archives of Clinical Neuropsychology, 31, 631–643.

Schretlen, D. J., Testa, S. M., & Pearlson, G. D. (2010). Calibrated Neuropsychological Normative System professional manual. Lutz, FL: Psychological Assessment Resources.

Stern, R. A., & White, T. (2003). NAB, Neuropsychological Assessment Battery: Administration, scoring, and interpretation manual. Lutz: Psychological Assessment Resources.

Sokal, R. R., & Rohlf, F. J. (1995). Biometry. San Francisco, CA : W.H. Freeman.

Wechsler, D. (1997). Wechsler Adult Intelligence Scale–third edition. San Antonio, TX: Psychological Corporation.

Willems, S., & Seron, X. (2023). Le rapport écrit et la remise de conclusions orales. In H. Amiéva, P. Azzouvi, E. Barbeau, & F. Collette, Traité de neuropsychologie de l'adulte Tome 1 – Évaluation. Paris, France : De Boeck supérieur.

ANNEXE 1 - La sensibilité et la spécificité

Pour quelques épreuves ou questionnaires, seul un score seuil (**cut-off score**) est communiqué. Nous recommandons dans ce cas de communiquer la sensibilité et la spécificité (**encart 1**) associées. Le qualificatif choisi ensuite devrait rester cohérent. Ainsi, si la sensibilité est de 95% et la spécificité de 98%, un score sous le seuil devrait être qualifié de très peu fréquent dans la population tout-venant et fréquent dans la population clinique. Il est toutefois habituel d'être face à un seuil de moindre spécificité (par exemple, 85%). En dessous d'une spécificité de 95%, nous recommandons à la/au neuropsychologue d'indiquer que le score n'est « **pas rare** » dans la population tout-venant.

Attention, il est important de rappeler au lecteur de ne pas confondre la faible fréquence dans la population « tout-venant » avec la probabilité d'avoir un vrai positif. Pour cela, faut-il encore connaître la prévalence de manière à calculer la valeur prédictive positive (VPP) (voir **encart 2**).

Encart 1. Sensibilité et spécificité, rappel de base

La sensibilité ou puissance d'un test

Le manque de précision d'un test peut conduire à deux types d'erreurs évidentes : 1) ne pas identifier une personne qui est porteuse de la caractéristique que nous tentons de détecter. On parlera, dans ce cas, de rejets incorrects ou faux négatifs. 2) détecter par erreur une personne comme étant porteuse de cette particularité alors qu'en réalité ce n'est pas le cas. On parlera alors d'acceptations erronées ou de faux positifs. Pour diminuer le premier type d'erreurs, il faut qu'un test possède ce que l'on appelle une bonne sensibilité. La sensibilité se réfère à la capacité du test à détecter ou mesurer les plus fines des différences réelles entre les performances ou les caractéristiques d'un même individu ou entre les performances ou les caractéristiques de plusieurs individus. On parle alors du pouvoir discriminant du test. La sensibilité d'un test est mesurée par la proportion de personnes ayant réellement la caractéristique mesurée et qui est identifiée comme telle. Par exemple si 80 % des personnes souffrant d'un problème d'anxiété sont diagnostiquées comme telles par un test (80 % de vrais positifs ; VP), on dira que la sensibilité de ce test est de 0,80. Par déduction, la probabilité qu'un test ayant une sensibilité de 0,80 ne permette pas de détecter un cas, ou la probabilité qu'un cas donné ne soit pas identifié par le test est de 0,20 (20 % de faux négatifs, FN). La sensibilité d'un test peut donc se calculer très simplement. Par exemple, si 200 personnes souffrant de problème d'anxiété sont évaluées, et que 160 sont diagnostiquées comme étant anxieuses tandis que 40 personnes ne sont pas détectées par le test, la sensibilité sera : 160 / (160 + 40) = 0,80.

La spécificité ou validité discriminante d'un test

Pour diminuer le second type d'erreurs qui découlent du manque de précision d'un test (c'est-à-dire, détecter par erreur une personne comme étant porteuse d'une particularité), il faut que celui-ci possède ce que l'on appelle une bonne spécificité. La spécificité représente la proportion de personnes n'étant pas porteuse de la caractéristique mesurée qui est identifiée par le test comme ne possédant effectivement pas cette caractéristique. Un test spécifique est un test qui détecte uniquement la caractéristique recherchée. Il a donc une portée restreinte, ce qui explique le terme « spécifique » ou « discriminant ». Une spécificité de 0,80 signifie donc que la probabilité de repérer un vrai négatif (VN) sera de 80 %. Par déduction, la probabilité qu'un test possédant une spécificité de 0,80 détecte erronément un cas (faux positifs ; VP) sera de 20 %. La spécificité se calcule très simplement : Spécificité=VN/(VN+FP). Par exemple, si 200 personnes ne souffrant pas d'anxiété sont évaluées, que 160 sont correctement rejetées par le test, mais 40 sont erronément détectées comme étant anxieuses, la spécificité du test sera : 160 / (160 + 40) = 0,80.

Encart 2. Quelle est ma probabilité de vrais positifs ?

Supposons que la sensibilité et la spécificité d'un test de screening d'un vieillissement cérébral pathologique soient très élevées : Sensibilité = 0,95 ; spécificité = 0,99. La/Le clinicien.e applique ce test à un sujet tout-venant et obtient un résultat positif. Quelle est la probabilité que le sujet soit réellement atteint d'un trouble neurocognitif ? Pour résoudre ce problème, une information supplémentaire est nécessaire : la fréquence de la

variable mesurée dans la population – la prévalence. La prévalence dans la population est la probabilité a priori (avant connaissance du résultat du test) que la/le patient.e soit malade. Supposons que la prévalence soit de 1/100 pour l'âge du sujet. Il conviendra ensuite de faire de simples calculs :
A) Supposons que j'aie un échantillon de 10 000 individus, on peut s'attendre à y trouver 100 malades et 9900 sains.
B) Étant donné ma sensibilité de 0,95, je peux m'attendre à avoir 95 vrais positifs et donc 5 faux négatifs
C) Étant donné ma spécificité de 0,99, je peux m'attendre à ce que 0,99 de mes 9900 (soir 9801) sains classés comme vrais négatifs. Il n'en demeure pas moins 99 faux positifs.
D) Ma probabilité de positif sur 10 000 individus est donc de 95 vrais positifs + 99 faux positifs, soit 194.
C) La proportion de malades sur les positifs (valeur prédictive positive, VPP) est donc de 95 / 194, soit 48,9 %.
Il est souvent difficile de connaître la prévalence d'un trouble avec précision. Il convient alors de vérifier le test avec les différentes valeurs de prévalence disponibles.

ANNEXE 2 – Effet de l'asymétrie sur le taux de faux positifs associé au score z

Le taux de faux positifs associé à l'utilisation du score z augmente en fonction de la « sévérité » de l'asymétrie (Crawford & Garthwaite, 2005). Par exemple, lorsque l'échantillon comporte 10 individus et que la distribution s'écarte fortement d'une distribution normale tant en ce qui concerne la symétrie que l'aplatissement, ce taux peut atteindre 10,65%. Ce taux est de 8,59% lorsque l'asymétrie est modérée, de 9,64% lorsqu'elle est sévère et de 10,23% lorsqu'elle est très sévère ou extrême (Crawford & Garthwaite, 2005). Le degré d'aplatissement de la courbe de distribution est une autre variable à considérer. Pour un même effectif de l'échantillon de référence, par rapport à une distribution normale, les taux de faux positifs ne sont pas influencés par l'aplatissement de la distribution (Crawford & Garthwaite, 2005). En revanche, lorsque le degré d'aplatissement est combiné au degré d'asymétrie, le taux de faux positifs augmente quel que soit l'effectif. Par exemple, pour un effectif de 10 et une distribution modérément leptokurtique (c'est-à-dire, dont la cloche est plus pointue que celle de la loi gaussienne), il est de 8,99% lorsque l'asymétrie est modérée, de 9,98% lorsqu'elle est sévère, de 10,47% lorsqu'elle est très sévère et de 10,65% lorsqu'elle est extrême. Même lorsque l'effectif est égal à 100, l'asymétrie modérée et la leptokurtie modérée, le taux de faux positifs est égal à 6,19% (Crawford & Garthwaite, 2005). Notons que les mêmes conclusions peuvent être tirées si la distribution est modérément ou sévèrement platykurtique (c'est-à-dire, dont la cloche est plus plate que celle de la loi gaussienne).

ANNEXE 3 – Calcul du percentile

Il y a plusieurs méthodes de calcul du percentile et toutes n'ont pas la même pertinence (Crawford, Garthwaite & Slick, 2009). Lorsque les données normatives individuelles sont disponibles, alors la méthode recommandée par Crawford et al. (2009) est la suivante (Crawford, Garthwaite & Slick, 2009) :

$$\text{Nbre C's<score P} + (\text{Nbre C's=score P} \times 0{,}5)\text{N C's} \times 100$$

Où :

Nbre C's <score P : nombre d'individus de l'échantillon de référence dont le score est inférieur à celui de la personne évaluée

Nbre C's = score P : nombre d'individus de l'échantillon de référence dont le score est égal à celui de la personne évaluée

N C's : effectif de l'échantillon de référence

ANNEXE 4 – Effet de l'asymétrie sur le taux de faux positifs associé au t de Crawford

En ce qui concerne l'asymétrie, avec un échantillon de 10 personnes, le taux de faux positifs est de 6,04% si l'asymétrie est modérée, 7,14% si elle est sévère, 7,80% si elle est très sévère et 7,94% si elle est extrême. Ces taux sont d'autant plus élevés que l'effectif de l'échantillon est petit (Crawford et Garthwaite, 2005). Bien que les taux de faux positifs atteints en utilisant le t modifié soient inférieurs à ceux atteints en utilisant le score z, ils sont au-dessus du seuil acceptable. Afin de résoudre ce problème, lorsque la distribution est asymétrique, il est indiqué d'abaisser le seuil de significativité à 0,025 (unilatéral). En effet, avec un tel seuil, le taux de faux positifs varie entre 3,23% et 4,99% quel que soit l'effectif et la valeur d'asymétrie (sauf quand l'asymétrie est extrême et que l'effectif est égal à 5 (5,21% de faux positifs) ou 10 (5,18% de faux positifs)) (Crawford et Garthwaite, 2005).

En ce qui concerne l'aplatissement de la distribution, les taux de faux positifs sont peu affectés par le degré de leptokurtie de la distribution. Ils restent inférieurs à 5% ou très proches de 5% lorsque l'effectif est supérieur ou égal à 20, que la distribution soit modérément ou sévèrement leptokurtique (Crawford, Garthwaite, Azzalini, Howell & Laws, 2006). En revanche, lorsque la distribution est à la fois asymétrique et leptokurtique, le taux de faux positifs dépasse les 5% quel que soit l'effectif (Crawford, Garthwaite, Azzalini, Howell & Laws, 2006). Par exemple, pour un effectif de 10 et une distribution modérément leptokurtique, il est de 7,42% lorsque l'asymétrie est modérée, de 7,85% lorsqu'elle est sévère, de 8,37% lorsqu'elle est très sévère et de 8,56% lorsqu'elle est extrême (Crawford, Garthwaite, Azzalini, Howell & Laws, 2006). Toutefois, en abaissant le seuil à 2%, le taux de faux positifs diminue dans toutes les conditions (mais reste à 6,4% pour un effectif de 5 lorsque la distribution est extrêmement asymétrique et sévèrement leptokurtique). En bref, ce seuil (correspondant à une valeur p égale à 0,02 en unilatéral) maintient le taux de faux positifs sous les 5% pour les distributions modérément ou sévèrement asymétriques (quel que soit le degré de leptokurtie et l'effectif). Pour les distributions très sévèrement ou extrêmement asymétriques, le taux de faux positifs est inférieur ou égal à 5% uniquement en l'absence de leptokurtie quel que soit l'effectif, en présence d'une leptokurtie modérée lorsque l'effectif est supérieur ou égal à 20 et en présence d'une leptokurtie sévère lorsque l'effectif est supérieur ou égal à 50. Dans les autres conditions, le taux de faux positifs varie entre 5,29% et 6,4%.

ANNEXE 5 - La prise en compte des valeurs extrêmes dans l'échantillon de référence

Lors de la récolte de données sur un échantillon « sain » afin d'obtenir des normes, des scores extrêmes peuvent être présents et ce même au sein d'échantillons de référence stratifiés en fonction de l'âge, de la scolarité et du genre. Par « scores extrêmes » (*outliers* en anglais), on désigne des scores qui sont très distants des autres scores de l'échantillon (Leys et al., 2019). Il s'agit donc de scores peu fréquents qui sont soit très bas, soit très élevés. Comme ces scores vont affecter l'aplatissement (kurtosis) de la distribution, la moyenne et l'écart-type, il est parfois tentant de les retirer de l'échantillon. Toutefois, en matière de construction de données normatives, cette pratique a été remise en cause (Crawford et al., 2006). Un premier argument avancé est que si une valeur extrême ne résulte ni d'une erreur de codage (erreur dans la transcription d'un score par exemple) ni de l'inclusion par erreur d'une personne appartenant à un groupe clinique, alors elle est une caractéristique du phénomène étudié et doit dès lors être maintenue dans l'échantillon. Un second argument est que retirer les valeurs extrêmes peut conduire à des erreurs dans la décision de considérer un score comme étant hors seuil ou non. Crawford et al. (2006) prennent l'exemple suivant : si les scores extrêmes de bas niveau d'un échantillon de référence sont retirés, alors la comparaison du score bas (et peut-être égal aux scores retirés) d'une personne évaluée à cet échantillon de référence conduira à un risque de faux positifs (puisque, en quelque sorte, la moyenne est surestimée et l'écart standard est réduit). On peut ajouter que de la même manière, retirer les scores extrêmes de haut niveau conduira à un risque de faux négatifs. Un troisième argument qui peut être avancé est qu'il est difficile d'estimer la rareté d'un score extrême dans la population lorsque l'effectif de l'échantillon de référence est petit. Or, les échantillons de référence associés aux tests cognitifs sont souvent de petite taille.

Si les scores extrêmes de l'échantillon de référence ne doivent pas être écartés, il reste à évaluer si leur maintien affecte ou non le risque d'erreur quant à la décision que le score observé soit hors seuil ou non. Afin de répondre à cette question, il convient de pouvoir identifier la présence ou non de scores extrêmes dans un échantillon et pour cela choisir une méthode permettant leur identification. Une méthode assez classique consiste à considérer un score comme extrême s'il s'écarte de la moyenne d'au moins 3 écarts-types. Le critère ici est donc la valeur du score *z* (effectivement ces scores sont rares puisqu'ils ne sont présents que dans 0,13% de la population). Cependant, d'une part, utiliser le score *z* suppose que la distribution des scores soit normale, et d'autre part, son calcul repose sur la moyenne et l'écart-type lesquels sont affectés par la valeur des scores extrêmes s'ils sont présents. Considérons par exemple cette série fictive de 10 scores correspondant à des nombres de réponses correctes (classés par ordre décroissant) : 19, 19, 19, 18, 18, 17, 17, 16, 15 et 1. Alors que 9/10 personnes ont obtenu un score situé entre 15 et 19, une personne a obtenu 1 (*de visu*, un score extrême). La moyenne de cette série est égale à 15,9 et la déviation standard à 5,4. Appliquer le critère fondé sur le score *z* conduit à considérer le score 1 comme n'étant pas extrême, car il est supérieur à -0,3 (résultat du calcul suivant : 15,9 – (3 x 5,4)). Notons que si le score le plus bas était 5 dans cette même série, la moyenne étant égale à 16,2 et l'écart-

type à 4,19, ce score ne serait pas non plus identifié comme extrême (5 étant supérieur à 16,2 – (3 x 4,19) = 3,63). Cette méthode peut donc ne pas détecter des scores extrêmes, car les indicateurs sur lesquels elle repose sont affectés par la valeur des scores extrêmes (Leys et al., 2013).

Une méthode alternative fondée sur des indicateurs plus robustes (qui ne dépendent pas de la valeur des scores dans l'échantillon) a été proposée par Leys et al. (2013). Elle repose sur la médiane et la médiane des écarts absolus de chaque score par rapport à la médiane. Dans la série fictive de 10 scores reprise ci-dessus, la médiane est égale à 17,5. La médiane des écarts absolus par rapport à la médiane s'obtient en calculant la différence absolue entre chaque score et leur médiane et en déterminant la médiane de cette série de différences. Dans l'exemple, elle est égale à 1,5. Le critère recommandé par Leys et al. (2013) pour identifier les scores extrêmes est le suivant : la médiane ± (2,5 x la médiane des écarts à la médiane x 1,4826) (voir Leys et al. (2013) pour une explication sur la nécessité de cette pondération). Appliqué à la série fictive, on obtient 17,5 – (2,5 x 1,5 x 1,4826) = 11,94. Dans ce cas, tous les scores inférieurs à 11,94 sont des scores extrêmes et le score 1 est identifié comme extrême. Notons qu'avec un score le plus bas égal à 5, le critère demeure inchangé (11,94) puisque ce qui le définit est la position des scores dans la série et la position des écarts absolus entre les scores et leur médiane. La valeur des scores extrêmes n'affecte donc pas le critère.

Maintenant que nous disposons d'une méthode permettant l'identification de scores extrêmes, nous pouvons évaluer leur éventuel impact sur la décision de considérer le score d'une personne évaluée comme hors seuil ou non. Nous allons nous limiter à l'évaluation de cet impact lorsque le t modifié (Crawford & Howell, 1998) est utilisé comme outil statistique de décision. Notons d'abord que dans la série fictive que nous utilisons, la distribution des scores n'est pas symétrique (score *z* de l'asymétrie = 4,09) et n'est pas mésokurtique (score *z* de l'aplatissement = 6,25). La valeur p est donc fixée au seuil ≤0,02 (unilatéral) pour déterminer si un score observé s'écarte significativement de la moyenne des scores de l'échantillon de référence. Le tableau 1 montre la valeur du t modifié et la valeur p obtenue pour des scores observés variant de 0 à 20 selon que la seule valeur extrême est extrêmement éloignée de l'ensemble des autres scores (1), très éloignée (5) ou assez éloignée (10). Lorsque la seule valeur extrême est égale à 1, seuls les scores inférieurs ou égaux à 2 sont hors seuil et tous les scores compris entre 3 et 20 ne le sont pas. Lorsque la seule valeur extrême est égale à 5, seuls les scores inférieurs ou égaux à 5 sont hors seuil. Lorsqu'elle est égale à 10, seuls les scores inférieurs aux égaux à 9 sont hors seuil. On ne peut donc exclure que le maintien d'une seule valeur extrême puisse conduire à un risque de faux négatifs (et aussi de faux positifs si les valeurs extrêmes sont de haut niveau plutôt que de bas niveau) et ce probablement d'autant plus que l'écart entre cette valeur et les autres valeurs de l'échantillon est grand.

Table 1 : valeur du t modifié et valeur p associée (unilatéral) pour chaque score de 0 à 20 comparé à un échantillon où un seul score extrême est présent (variant de 1 à 10).

	Score extrême le plus bas					
	1		5		10	
score	t modifié	p	t modifié	p	t modifié	p
0	-2,8050	0,0103	-3,7081	0,0024	-5,8447	0,0001
1	-2,6285	0,0137	-3,4806	0,0035	-5,4968	0,0002
2	-2,4521	0,0183	-3,2531	0,0050	-5,1489	0,0003
3	-2,2757	0,0245	-3,0256	0,0072	-4,8010	0,0005
4	-2,0993	0,0326	-2,7981	0,0104	-4,4531	0,0008
5	-1,9229	0,0433	-2,5706	0,0151	-4,1052	0,0013
6	-1,7465	0,0573	-2,3431	0,0219	-3,7573	0,0023
7	-1,5701	0,0754	-2,1156	0,0317	-3,4094	0,0039
8	-1,3937	0,0984	-1,8882	0,0458	-3,0615	0,0068
9	-1,2172	0,1272	-1,6607	0,0656	-2,7136	0,0119
10	-1,0408	0,1626	-1,4332	0,0928	-2,3657	0,0211
11	-0,8644	0,2049	-1,2057	0,1293	-2,0178	0,0372
12	-0,6880	0,2544	-0,9782	0,1768	-1,6699	0,0646
13	-0,5116	0,3106	-0,7507	0,2360	-1,3220	0,1094
14	-0,3352	0,3726	-0,5232	0,3067	-0,9741	0,1777
15	-0,1588	0,4387	-0,2957	0,3871	-0,6262	0,2734
16	0,0176	0,5068	-0,0682	0,4735	-0,2783	0,3935
17	0,1941	0,5748	0,1592	0,5615	0,0696	0,5270
18	0,3705	0,6402	0,3867	0,6460	0,4175	0,6569
19	0,5469	0,7011	0,6142	0,7229	0,7654	0,7682
20	0,7233	0,7561	0,8417	0,7891	1,1133	0,8528

Comment dès lors concilier le maintien des scores extrêmes dans les échantillons de référence et la limitation du risque d'erreur dans la décision ? La connaissance de la présence de scores extrêmes et de leur fréquence d'occurrence dans l'échantillon devrait pouvoir guider le processus de décision. Si l'échantillon ne comporte aucune valeur extrême, alors la décision se fait uniquement sur base de la valeur p puisque le risque d'erreur lié aux scores extrêmes

ne se pose pas. Si l'échantillon comporte des valeurs extrêmes de haut niveau, il y a un risque de faux positif si le score observé est de bas niveau mais pas s'il est de haut niveau. En effet, ce sont les scores inférieurs à la moyenne qui risquent d'être considérés comme hors seuil, pas ceux qui sont supérieurs à la moyenne. De même, s'il comporte des valeurs extrêmes de bas niveau, il y a un risque de faux négatif si le score observé est de haut niveau mais pas s'il est de bas niveau. En effet, ce sont les scores supérieurs à la moyenne qui risquent d'être considérés comme non hors seuil, pas les scores inférieurs à la moyenne. Savoir si l'échantillon comporte des valeurs extrêmes qui peuvent engendrer un risque de faux positif ou de faux négatif est donc indispensable. Malheureusement, rares sont les études normatives qui renseignent sur les valeurs extrêmes et leur fréquence (voir Fery et al. pour une exception). Toutefois, cette connaissance ne suffit pas, encore faut-il un critère qui permette de déterminer si la décision basée sur le t modifié est assortie ou non d'un risque de faux positif ou négatif. À nouveau, il est possible d'avoir recours à la médiane et la médiane des écarts absolus des scores par rapport à la médiane en utilisant le critère utilisé pour identifier les scores extrêmes. Dans la série fictive, ce critère est égal à 11,94. Dès lors, tout score inférieur à 11,94 peut être considéré comme hors seuil.

Appliquons la procédure proposée à la série fictive dont le score le plus bas est 1. Si le score observé est égal à 5, la valeur p associée au t modifié est égale à 0,0433. Dès lors, sur base du seuil fixé à 0,02, ce score n'est pas hors seuil. Toutefois, la présence d'un score extrême (1) présent chez un individu de l'échantillon de référence diminue la moyenne et augmente l'écart-type, faisant courir un risque de faux négatif. Dans ce cas, en utilisant le seuil fondé sur la médiane et la médiane des écarts des scores par rapport à la médiane (11,94), le score étant inférieur à 11,94, il peut être décidé qu'il soit hors seuil. Si le score observé était égal à 14, la valeur p associée au t modifié serait égale à 0,3726. À nouveau, cela conduit à considérer le score comme n'étant pas hors seuil. Cette fois-ci, 14 étant supérieur au score seuil (11,94), le score est toujours considéré comme n'étant pas hors seuil.

ANNEXE 6 - Quand tou·te·s les participant·e·s du groupe de référence ont le même score

Lorsque le niveau de difficulté d'une tâche est bas, il arrive que tou·te·s les participant·e·s du groupe de référence aient le même score maximum (ou minimum s'il s'agit d'un nombre d'erreurs). Dans ce cas, l'écart type est égal à 0, rendant impossible le calcul du score *z* et du t modifié. Le percentile ne peut pas non plus être utilisé puisque tout score inférieur au score identique chez tou·te·s les participant·e·s sera situé au percentile 0 s'il reflète un bas niveau de fonctionnement. Par exemple, si tou·te·s les participant·e·s ont 0 comme score d'erreur et le score de la personne évaluée est 1, ce score sera situé au percentile 0. Selon la définition du seuil de décision pour un percentile, ce score sera considéré comme hors seuil.

Dans ce cas de figure, on peut se demander si l'application stricte du critère de décision ne risque pas de conduire à un faux positif, surtout lorsque l'échantillon est de petite taille. Si aucun cut-off n'est communiqué par les auteurs, il appartient donc au/à la psychologue clinicien·ne de prendre en considération cette situation particulière dans l'interprétation qui sera donnée au score et de conclure (ou non) qu'il ne reflète pas une altération sous-jacente.

ANNEXE 7 - Scores hors-seuil dans des ensembles de plusieurs tests

Ingraham et Aiken avaient déjà montré en 1996 qu'une batterie de seulement six tests conduit à classer plus de 20 % de l'échantillon normatif comme ayant des performances « anormales » (seuil de -1 SD) sur au moins deux mesures. Godefroy et al. (2014) ont également observé que, dans une batterie pouvant contenir jusqu'à 19 scores, les taux de faux positifs (i.e., la proportion d'individus ayant un score inférieur au percentile 5, P5) augmentent en moyenne de 2,61 % chaque fois qu'un nouveau score cognitif est considéré, et atteint 21 % pour une batterie de 8 scores. Par la suite, d'autres études ont illustré la fréquence à laquelle des individus âgés sains sont susceptibles d'obtenir des scores hors-seuil lorsqu'ils passent une série de tests neuropsychologiques (Brooks et al. 2007 ; 2008 ; voir aussi Brooks, Iverson, & White, 2007 ; Palmer et al., 1998). Brooks et al. (2008) ont par exemple constaté que 25,7 % d'un échantillon de 550 adultes âgés sains ont au moins un score <P5 lorsque les huit scores de la WMS-III (ajustés en fonction de l'âge, Wechsler, 1997) sont examinés simultanément — un pourcentage qui passe à 39 % lorsque les scores sont ajustés en fonction des caractéristiques démographiques. Les résultats obtenus en examinant simultanément les 10 scores du subtest « mémoire » de la NAB sont assez similaires (Brooks et al., 2007) : 30,8 % d'un échantillon de 742 adultes âgés sains ont au moins un score <P5 et 16,4 % d'entre eux obtiennent au moins un score <P2. Dans le contexte du bilan neuropsychologique, ces résultats soulignent les problèmes que peuvent poser l'interprétation des scores isolés et la nécessité de tenir compte du nombre de tests lors de l'analyse des performances pour éviter de diagnostiquer à tort la présence d'un trouble ou d'une altération cognitive. Cela explique pourquoi certain.e.s patient.e.s pour lesquel/le.s a été posé un diagnostic de troubles neurocognitifs légers (« Mild Cognitive Impairment », MCI) sur la base du critère psychométrique d'au moins un score hors-seuil en mémoire (score Z <1,5) n'évoluent pas voire s'améliorent (de Rotrou et al., 2005).

ANNEXE 8 - Proposition de tableau pour le compte-rendu écrit

Comme illustré dans le tableau ci-dessous, les qualificatifs peuvent être présentés dans une table avec les percentiles et scores standards équivalents. Pour faciliter la lecture, il nous apparaît utile d'associer à cette table une brève explication du principe de la comparaison normative (Willems & Seron, 2023). Par facilité, nous proposons deux tables avec les scores seuils adaptés aux raisonnements unilatéral et bilatéral.

Comparaison normative (unilatéral)

En ce qui concerne l'interprétation des résultats aux épreuves, les scores obtenus sont comparés aux performances des personnes du groupe normatif (si possible de même âge, sexe et niveau d'éducation). Les résultats sont indiqués en référence à ces données normatives. Elles sont présentées sous forme de fréquence estimée de la population ayant ce score ou moins (la correspondance avec différents scores standardisés utilisés est également notée). Une fréquence de 5% (score *z* de -1,65, QI de 75, note standard de 5 ou score T de 34) correspond à une valeur en dessous de laquelle se trouve 5 % des performances des personnes du groupe normatif. Le tableau ci-dessous présente la grille d'interprétation des résultats quantitatifs et les scores seuils utilisés pour juger un score comme s'écartant des scores attendus.

Fréquence estimée* de la population ayant ce score ou moins	**Score *z***	**QI**	**Note standard (n)**	**Note T**	**Description**	**Interprétation statistique**
≧98%	≧+2,05	≧131	≧16	≧71	Score **très peu fréquent** de haut niveau	
Entre ≧95% et <98%	≧+1,65 à <+2,05	≧125 à <131	≧15 à <16	≧67 à <71	Score **peu fréquent** de haut niveau	
Entre >5% et <95%	>-1,65 à <+1,65	>75 à <125	>5 à <15	>34 à <67	Score **fréquent**	
Entre >2% et ≦5%	>-2,05 à ≦-1,65	>69 à ≦75	>4 à ≦5	>30 à ≦34	Score **peu fréquent** de bas niveau	hors norme (**seuil libéral)**
≦2%	≦-2,05	≦69	≦4	≦30	Score **très peu fréquent** de bas niveau	hors norme **(seuil conservateur)**

Note. Valeur p associée au test *t* modifié, percentile et probabilité associée au score *standardisé*.

Comparaison normative (bilatéral)

En ce qui concerne l'interprétation des résultats aux épreuves, les scores obtenus sont comparés aux performances des personnes du groupe normatif (si possible de même âge, sexe et niveau d'éducation). Les résultats sont indiqués en référence à ces données normatives. Elles sont présentées sous forme de fréquence estimée de la population ayant ce score ou moins (la correspondance avec différents scores standardisés utilisés est également notée). (score *z* de -1,96, QI de 71, note standard de 4 ou score T de 30) correspond à une valeur en dessous de laquelle se trouve 5 % des performances des personnes du groupe normatif. Le tableau ci-dessous

présente la grille d'interprétation des résultats quantitatifs et les scores seuils utilisés pour juger un score comme s'écartant des scores attendus.

Fréquence estimée* de la population ayant ce score ou moins	Score *z*	QI	Note standard (n)	Note T	Description	Interprétation statistique
≧98%	≧+2,33	≧135	≧17	≧73	Score **très peu fréquent** de haut niveau	hors norme **(seuil conservateur)**
Entre ≧95% et <98%	≧+1,96 à < +2,33	≧129 à <135	≧16 à <17	≧67 à <73	Score **peu fréquent** de haut niveau	hors norme (**seuil libéral)**
Entre >5% et <95%	>-1,96 à <+1,96	>71 à <129	>4 à <16	>30 à <67	Score **fréquent**	
Entre >2% et ≦5%	>-2,33 à ≦-1,96	>65 à ≦71	>3 à ≦4	>27 à ≦30	Score **peu fréquent** de bas niveau	hors norme (**seuil libéral)**
≦2%	≦-2,33	≦65	≦3	≦27	Score **très peu fréquent** de bas niveau	hors norme **(seuil conservateur)**

Note. Valeur p associée au test *t* modifié, percentile et probabilité associée au score *standardisé*.