Article de synthèse

Rev Neuropsychol

2016 ; 8 (1) : 61-9

La norme en neuropsychologie, un concept à facettes multiples

The norm in neuropsychology, a multifaceted concept

Françoise Colombo[1], Hélène Amieva[2], Thierry Lecerf[3], Vincent Verdon[4]

[1] Hôpital Fribourgeois,
Unité de neuropsychologie,
chemin des Pensionnats 2,
1708 Fribourg, Suisse
<francoise.colombo@h-fr.ch>

[2] Inserm U1219 Bordeaux Population Health Center, Université de Bordeaux,
33076 Bordeaux,
France

[3] Université de Genève et formation universitaire à distance,
Faculté de psychologie et des sciences de l'éducation,
Bd du Pont d'Arve 40,
1211 Genève 4, Suisse

[4] Hôpital Neuchâtelois,
Unité de neuropsychologie et logopédie,
rue de la Maladière 45,
2000 Neuchâtel,
Suisse

Pour citer cet article : Colombo F, Amieva H, Lecerf T, Verdon V. La norme en neuropsychologie, un concept à facettes multiples. *Rev Neuropsychol* 2016 ; 8 (1) : 61-9 doi:10.1684/nrp.2016.0365

Résumé

Afin de faciliter l'interprétation des scores des tests, l'utilisation de valeurs normées est devenue la règle en neuropsychologie clinique, parfois au détriment de l'approche clinique. Au vu de l'abondance des outils d'évaluation à disposition, il devient prioritaire de comprendre le plus précisément possible ce que recouvre le concept de normes afin de les utiliser au plus près de la réalité clinique sans en subir la tyrannie. Dans ce but, nous avons réuni un colloque avec différents spécialistes de la question qui nous livrent ici quelques clés de lecture de ce concept. Tout d'abord est présentée une réflexion autour de la normalité, qui s'oppose à la pathologie, à l'aide de normes comparatives ou diagnostiques, tout en utilisant des normes exprimées sous formes de quantiles ou de moyennes. Il sera rappelé qu'une norme est établie pour une population donnée et pour un objectif spécifique. Concernant les tests eux-mêmes, leurs constructions exigent des critères spécifiques, et la discussion s'articule autour des critères de validité et de fidélité, dont l'importance est parfois mal comprise et souvent négligée. Enfin, l'accent est mis sur la diversité des normes pour des tests très largement utilisés, et une tentative de définir des critères communs au sein d'une communauté de spécialistes est proposée.

Mots clés : norme • normalité • diagnostic • validité/fidélité • pathologie

Abstract

The procedure consisting in comparing the individual's test score to a norm group has become the standard in the field of clinical neuropsychology, sometimes to the detriment of a clinical approach. Given the numerous tools of assessment available for clinicians, it is particularly important to precisely understand what underlies the concept of norms in order to adequately use them in clinical practice, without suffering their tyranny. To this end, we organized a workshop involving different experts on this topic. The present article is the fruit of their thoughts and advances some keys to a better understanding of this concept. First, we present considerations on what are "normal scores", as opposed to "pathological scores", and discuss the issue of comparative and diagnostic norms, as well as the use of quantiles or means. We remind the reader that a norm is established for a given population and has a specific purpose. Regarding the tests themselves, their developments require specific criteria, and the discussion is oriented towards the criteria of validity and fidelity, which importance is sometimes poorly understood and often neglected. Finally, the diversity of norms applied to widely used tests is addressed. We present the outcome of a reflection group constituted of specialists in clinical neuropsychology, attempting to define common criteria for specific tests in order to offer uniform decision criteria for a given pathological status in a given population.

Key words: norm • normality • diagnostic • validity/fidelity • pathology

Correspondance :
F. Colombo

doi: 10.1684/nrp.2016.0365

Introduction

Ces dernières décennies ont vu l'apparition d'un nombre important d'outils d'évaluation en neuropsychologie clinique, basés notamment sur le développement des modèles issus des neurosciences cognitives. La mise à disposition de ces outils, qui se sont ajoutés aux tests plus anciens encore largement utilisés, a créé dans la boîte à outils du neuropsychologue un vaste désordre où se côtoient des tests expérimentaux pour l'évaluation de cas uniques et des tests développés et évalués à partir de larges échantillons de population. Le temps est donc venu d'une véritable réflexion sur la notion de normes, les critères de choix des tests dans l'évaluation, préparant ainsi le grand ménage probable de cette boîte à outils qui devrait en découler.

Dans ce but, le club « Évaluation » de la Société de neuropsychologie de langue française, associé aux collaborateurs des unités de neuropsychologie de l'hôpital fribourgeois (Suisse), a mis en place un workshop de réflexion en juillet 2015. Ce sont les points de vue de différents spécialistes des normes – épidémiologiste, psychométricien et clinicien – que nous souhaitons livrer dans cet article, ce qui, dans un premier temps, permettra de décortiquer le substrat de ces outils afin de mieux en comprendre les forces… et les faiblesses et ainsi de mieux aborder la « tyrannie » des normes, telle que la vivent les jeunes, et moins jeunes, cliniciens.

La psychologie expérimentale a, dès la fin du XIX^{e} siècle, cherché à établir des normes de référence pour des fonctions cognitives données, un des exemples célèbres étant les travaux d'Ebbinghaus (1850-1909) sur la mémoire. D'importants travaux de normalisation ont été développés dans le domaine de l'intelligence, et les travaux de Wechsler (1896-1981) n'ont cessé de s'enrichir de multiples adaptations dès la fin des années 1930, avec la publication de l'échelle d'intelligence bien connue sous le nom de WAIS sous sa première forme en 1955, et dont nous utilisons actuellement la quatrième forme.

La neuropsychologie s'est développée avec des outils initialement cliniques ou empruntés à d'autres champs de la neuropsychologie (psychologie de l'enfant, psychologie différentielle, orientation professionnelle) avant de développer ses propres outils, raison pour laquelle nous devons évoquer le caractère à facettes multiples des évaluations et prendre conscience du caractère polymorphe de nos influences. Les travaux présentés dans cet article offrent l'opportunité de saisir cette complexité.

Le point de vue de l'épidémiologiste : à propos des concepts de normalité et de pathologie

Moyennes ou quantiles ? Définitions

Le score à un test neuropsychologique est une estimation quantitative de la capacité cognitive mise en jeu dans le test en question. Le travail du neuropsychologue ne se limite toutefois pas à la mesure des capacités cognitives mais comprend l'évaluation de cette mesure. C'est à cette étape que le neuropsychologue a souvent recours à un échantillon de normalisation. L'échantillon de normalisation fournit des paramètres statistiques, tels que la moyenne, la médiane ou encore la variance, qui permettront de situer le score d'un sujet donné à un test donné par rapport à une distribution de scores observés dans une population donnée. Une précaution impérative à respecter pour le bon choix de ces indicateurs est de s'interroger sur la nature de la distribution des scores au test dont il est question.

Par définition, pour qu'une distribution de scores soit considérée comme « normale », il faut que la distribution des observations soit continue. Or rappelons que ce cas de figure n'est rencontré que pour une petite minorité de tests. Dans certains cas, le degré de déviance de la distribution par rapport à une distribution normale est acceptable, on utilise alors le modèle normal comme approximation pratique des distributions de scores psychométriques. Lorsque la distribution s'éloigne trop d'une distribution normale, alors, appliquer un seuil de détection d'un possible déficit déterminé par la moyenne moins une valeur x d'écart-type est totalement inapproprié. C'est notamment le cas de l'incontournable test de MMSE [1], présentant une distribution particulièrement asymétrique avec un effet plafond majeur, la très grande majorité des observations se situant dans un intervalle très réduit.

Le calcul du z-score (score centré-réduit) peut s'avérer intéressant, le z-score ayant l'avantage de permettre une comparaison des performances obtenues à différents tests qui ne partagent pas la même métrique. Néanmoins, et de nouveau, il n'est utilisable que lorsque les performances suivent une loi normale puisque son calcul est basé sur la moyenne et l'écart-type de la distribution.

Dans ces nombreux cas où la distribution des scores n'est pas normale, utiliser des normes calculées en quantiles s'avère plus approprié. Les quantiles sont des valeurs numériques associées à une fréquence cumulée. Un centile (appelé parfois percentile selon la terminologie anglaise) est chacune des 99 valeurs seuils qui partagent les observations

en 100 parts égales, de sorte que chaque centile représente un centième de l'échantillon. Parmi les valeurs seuils, on distingue notamment la médiane partageant la distribution en deux effectifs égaux, les trois quartiles, bornes respectives des classes des 25 %, 50 % et 75 % des scores les plus bas, ou encore les quatre quintiles, bornes respectives des classes des 20 %, 40 %, 60 % et 80 % des scores les plus bas. Le caractère « haut » ou « bas » d'une performance à un test sera dans ce cas déterminé par le degré de rareté des observations dans la population. Certains quantiles, comme le dixième percentile, sont utilisés comme des valeurs seuils, en deçà desquelles un score sera considéré comme possiblement pathologique, en renfort d'une impression clinique ayant besoin d'être objectivée.

L'importance des caractéristiques de l'échantillon de normalisation

À ce stade, il est nécessaire de rappeler que pour que la comparaison du score d'un sujet donné à ceux observés dans l'échantillon de référence soit informative, il faut que ces normes aient été obtenues sur un échantillon de référence constitué de sujets les plus représentatifs du (ou des) sujet(s) à qui est destiné le test en question. La « représentativité » de l'échantillon de référence, loin d'être une précaution méthodologique superflue, va déterminer en grande partie la qualité de la mesure et sa capacité à nous renseigner sur le degré de déviance du score d'un individu. Par commodité de recrutement pourtant, force est de constater que les échantillons de référence à partir desquels les normes des tests sont établies sont très souvent constitués de sujets volontaires fortement sélectionnés présentant un profil psycho-sociodémographique très particulier les rendant peu représentatifs de la population à laquelle le test s'adressera. La principale conséquence sera une surestimation du niveau de performance attendu.

Prenons l'exemple du RL/RI 16 items, un des tests les plus utilisés par les neuropsychologues à l'heure actuelle pour objectiver les troubles de la mémoire épisodique chez des personnes suspectées de présenter une maladie neurodégénérative. Comparons les scores obtenus à ce test dans un large échantillon de personnes âgées sélectionnées en population générale, tel que l'échantillon de l'étude épidémiologique des 3 Cités dans lequel le large spectre de performances qui existe dans la population sera représenté [2] aux scores obtenus en développant l'équation de régression proposée par l'étude du GREMEM, estimée à partir des scores d'échantillon de plus petite taille constitué de personnes âgées saines volontaires [3]. On constate alors que ces derniers sont systématiquement supérieurs au score correspondant à la médiane de la distribution en population générale, et ce quelles que soient les catégories d'âge et de niveau d'étude considérées. Prenons l'exemple d'une femme de 78 ans ayant arrêté sa scolarité en fin d'enseignement secondaire. En développant l'équation de régression du GREMEM, les scores attendus aux trois premiers rappels libres et au rappel libre différé seront de 7,9, 9,3, 10,5 et 10,7 respectivement. En population générale, ces scores se situent entre la médiane et le 75^{e} percentile de la distribution. Cet exemple illustre combien il est important, à l'heure d'interpréter un score, de s'interroger sur les caractéristiques de l'échantillon de référence à partir duquel les normes qui feront office de point de comparaison ont été calculées, et ce afin d'éviter de considérer pathologique un score du fait de scores supposés « normatifs », qui en réalité, sont probablement surestimés.

Normes comparatives et normes diagnostiques

Il est inutile de dire à quel point la performance aux tests neuropsychologiques – la très grande majorité des tests – est influencée par l'âge, le niveau d'étude et dans certains cas, le sexe. Pour cette raison, la plupart des normes publiées sont généralement exprimées pour une catégorie d'âge, et/ou par niveau d'étude, et/ou en fonction du sexe. Pourtant, dans la situation où le neuropsychologue cherche à détecter des pathologies comme les pathologies démentielles, telles que la maladie d'Alzheimer, pathologies dont les facteurs de risque principaux sont précisément l'âge et le niveau d'étude, n'y a-t-il pas une contradiction ? En d'autres termes, en utilisant des normes pondérées sur l'âge et le niveau d'étude, n'est-il pas contre-productif de chercher à contrôler l'effet de ces variables qui à elles seules expliquent une part du risque d'avoir une de ces pathologies ? Cette question, à savoir, des scores standardisés permettent-ils de mieux détecter la présence d'une démence que des scores non standardisés, a fait l'objet de l'étude de O'Conell *et al.* [4]. Leur étude a consisté à comparer, au sein d'une cohorte de 1 252 personnes dont 213 sujets déments ayant complété de nombreuses épreuves neuropsychologiques (échelle de mémoire de Wechsler, empan de chiffres, épreuve des 15 mots de Rey, test de Benton, épreuves de similitudes et de compréhension de la WAIS et test des codes de Wechsler), les performances de classement des sujets déments/non déments au moyen des scores bruts et des scores corrigés pour l'âge à ces différents tests. Ils ont ainsi montré que des scores corrigés pour l'âge classent les déments avec une meilleure spécificité mais que les scores non corrigés pour l'âge classent les déments avec une meilleure sensibilité. De la même manière, Sliwinski *et al.* [5] avaient montré, quelques années plus tôt, que des scores corrigés pour l'âge classaient des sujets déments avec une sensibilité inférieure de 28 %. Ces résultats avaient amené ces auteurs à proposer de distinguer deux types de normes : d'une part, les « normes comparatives » établies sur la base de scores corrigés (pour l'âge et/ou le niveau d'étude) et les « normes diagnostiques », d'autre part, établies au moyen des scores non corrigés. L'intérêt des premières sera d'apporter au psychologue des éléments statistiques pour l'aider à situer la performance d'un sujet par rapport à son groupe de référence, tandis que l'intérêt des normes diagnostiques sera de contribuer à établir une probabilité de présenter ou de développer une démence. Idéalement, un travail de normalisation d'un test donné

devrait être assorti de ces deux types de normes, fournissant aux neuropsychologues des informations tout à fait complémentaires sur la performance d'un sujet par rapport à la performance attendue compte tenu de ses caractéristiques sociodémographiques, mais aussi sur sa probabilité de présenter ou de développer une démence. Malheureusement, si les premières figurent souvent dans les articles rapportant des normes, les secondes sont, à tort, trop souvent négligées dans les publications.

Le point de vue du psychométricien : à propos des concepts de fidélité et de validité

Préambule

En 2014, la dernière version des *Standards for Educational and Psychological Testing* a été publiée [6]. Le but de ces *Standards* est de proposer des critères pour le développement et l'évaluation des tests, et des pratiques de testing. Dès l'introduction, les *Standards* soulignent que seule l'utilisation de tests *valides* et *fidèles* peut conduire à des décisions individuelles adéquates. Par conséquent, un test qui ne présente aucune estimation de la validité et de la fidélité ne devrait pas être interprété de manière quantitative (pour exemple, le test de la Figure de Rey, bien connu de l'ensemble des cliniciens et largement utilisé, ne rapporte pas ces données). De façon générale, rappelons que les estimations de la fidélité et de la validité sont dépendantes des échantillons. Ainsi, les données recueillies à partir d'un échantillon américain ne peuvent pas être appliquées à un échantillon francophone.

Concept de validité

« *Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests* » [6]. Selon les *Standards*, la validité est un concept unitaire qui fait référence à l'interprétation des scores, et non au test lui-même. Dans le but d'établir la validité de l'interprétation des scores d'un test (*i.e.* hypothèses sur l'attribut psychologique), les constructeurs doivent s'appuyer sur un modèle théorique et sur des données empiriques. Si les arguments empiriques en faveur de la validité de l'interprétation d'un score peuvent être apportés *via* de multiples sources (structure interne, contenu, conséquences, et processus de réponses), nous souhaitons détailler ici la stratégie qui consiste à s'appuyer sur l'étude des (co-)relations entre les scores à un test et d'autres variables externes (appelées *evidence based on relations to other variables* dans les *Standards*).

Dans cette procédure, appelée auparavant validité critérielle, il s'agit non seulement d'administrer le test à des individus, mais également de recueillir le score sur une ou plusieurs variables externes, appelées alors critères. Par exemple, on peut calculer la corrélation entre le QI (le test) et le salaire annuel (le critère), entre le QI et le milieu socio-économique, entre le QI et les matrices de Raven, etc. Chaque coefficient de corrélation (ex. QI – salaire annuel) représente alors un coefficient de validité. Aujourd'hui encore, cette corrélation Test-Critère est souvent considérée comme la preuve ultime de la validité de l'interprétation du score d'un test. On fait l'hypothèse qu'une corrélation Test-Critère significative s'explique par le fait que le test et le critère évaluent le même attribut psychologique (notion de validité convergente). À l'inverse, si la corrélation n'est pas significative, on fait l'hypothèse que le test et le critère évaluent des attributs distincts (notion de validité discriminante). De *simples* corrélations Test-Critère sont donc censées apporter des preuves de la validité de l'interprétation du score du test. En d'autres termes, si le test corrèle avec le critère, qui est considéré comme une approximation d'un attribut donné, alors cela signifie que l'on a identifié l'attribut évalué par le test. Cette logique des corrélations Test-Critère est remise en question par Borsboom *et al.* [7, 8]. Selon ces auteurs, la validité d'un test ne peut pas être évaluée à partir de *simples* corrélations, car il s'agit d'un concept causal. Dans le cas de la validité convergente, la *logique* sous-jacente est que les différences interindividuelles observées sur le test et le critère sont déterminées par un seul et même attribut psychologique. On s'inscrit bien dans une perspective causale, puisque l'attribut est responsable des différences interindividuelles à la fois sur le test et le critère. C'est seulement ce cas de figure qui permettrait d'apporter des preuves sur la validité de l'interprétation du score du test. Si nous revenons maintenant à l'exemple de la corrélation entre le QI et le revenu annuel, on peut se demander quel attribut explique à la fois les différences interindividuelles sur le QI et sur le revenu annuel ; quelle est la cause commune ? Au passage, puisque le revenu annuel sert ici de critère, cela devrait signifier que l'on a identifié l'attribut psychologique sous-tendant les différences sur le revenu annuel. Parmi les exemples ci-dessus, la seule preuve de la validité de l'interprétation du QI pourrait être établie à partir de la corrélation avec les matrices de Raven. Toutefois, si on considère le QI comme une variable formative (et non comme une variable réflective), alors il est difficile d'interpréter la corrélation QI-Raven comme la preuve d'un attribut psychologique commun. En conséquence, la validité critérielle n'est pas une bonne méthode pour établir la validité de l'interprétation du score d'un test. Ce sont là les conclusions de Markus et Borsboom [7] : « *criterion validity was truly one of the most serious mistakes ever made in the theory of psychological measurement* » ; ou encore « *The idea that validity consists in the correlation between a test and a criterion has obstructed a great deal of understanding and continues to do so* ». La crédibilité des corrélations Test-Critère dépend de la pertinence, de la fidélité et de la validité des critères utilisés. Or, le plus souvent cette évaluation du (des) critère(s) n'est pas effectuée. En résumé, la question de la validité (de l'interprétation des scores) reste souvent problématique, ce qui fait dire à Brennan [9] : « *Unfortunately, far too many tests (even*

well-known, large-scale tests) lack adequate, timely documentation with respect to validation, even validation of the most fundamental claims ». On peut espérer que la récente conception de la validité avancée par Borsboom *et al.* [7, 8] conduise à de nouveaux développements. Selon ces auteurs, un test est valide (et non son interprétation), si l'attribut évalué par le test existe, et si les différences interindividuelles sur l'attribut sont à l'origine des différences interindividuelles sur les scores du test (perspective causale). Selon Borsboom *et al.*, les problèmes liés à la validité sont avant tout des problèmes PSYCHOmétriques et non psychoMÉTRIQUES.

Concept de fidélité

Le second chapitre des *Standards* est consacré à la notion de fidélité/précision. Si la validité est le critère le plus important pour évaluer l'utilité d'un test, la fidélité des scores représente le complément indispensable. Les *Standards* définissent ce concept de la manière suivante : « *. . .we use the term reliability/precision to denote the more general notion of consistency of the scores across instances of the testing procedure* ». Si les performances obtenues au travers des passations sont cohérentes, alors la fidélité/précision est élevée. Il faut rappeler que selon la Théorie Classique des Tests, sur laquelle s'appuient les *Standards*, le score observé à un test comprend l'*erreur de mesure* et le *score vrai*. L'erreur de mesure traduit des fluctuations aléatoires, tandis que le score vrai traduit le niveau de l'individu sur l'attribut. Si l'erreur de mesure d'un test est trop importante, alors le clinicien ne pourra pas utiliser les scores observés. Il serait délicat d'émettre des hypothèses sur le fonctionnement d'un individu à partir de scores qui fluctuent aléatoirement. Toutes les procédures d'estimation de la fidélité/précision visent à *quantifier* l'erreur de mesure (pour une discussion détaillée de la quantification dans les tests, se référer au livre de J. Michell [10]).

L'indicateur le plus souvent rapporté de la fidélité est sans aucun doute le coefficient alpha de Cronbach. Bien que l'usage de ce coefficient soit très répandu, il conduit néanmoins à de nombreux mésusages et erreurs. Tout d'abord, le coefficient alpha n'est pas réellement un indicateur de la fidélité, puisqu'il n'évalue pas la cohérence des scores obtenus par des individus au travers d'un grand nombre de passations indépendantes. En estimant l'amplitude moyenne des co-variations entre tous les items, ce coefficient est donc avant tout un indicateur de la consistance interne. La deuxième erreur est de considérer que le coefficient alpha apporte des preuves sur l'unidimensionnalité des items ; ce n'est pas le cas. Le coefficient alpha ne doit être calculé que pour des items dont l'unidimensionnalité a préalablement été démontrée à partir d'analyses factorielles, par exemple. Le coefficient alpha pose donc de nombreux problèmes, ce qui fait dire à Kane [11] : « *To simply estimate coefficient alpha or some other measure of internal consistency and assume that the question of reliability/generalizability has been addressed is to beg the question of generalizability* ». Enfin, les données du coefficient alpha sont très souvent à l'origine de la détermination des intervalles de confiance qui vont être utilisés dans une perspective clinique.

Une autre méthode, qui permet d'évaluer la cohérence des scores, est la procédure Test-Retest. Elle consiste à administrer le test une première fois et à l'administrer une deuxième fois après un intervalle de temps variant de quelques jours à plusieurs années. On calcule alors la corrélation entre les scores de la première et de la deuxième passation. Cette procédure repose sur le postulat selon lequel la première passation n'a aucune influence sur la deuxième (notion d'indépendance). On peut rappeler que la stabilité des différences interindividuelles au cours du temps représente un argument essentiel pour considérer qu'on évalue bien un attribut, un trait, et non un état. Plus important, si la procédure Test-Retest évalue la stabilité des différences interindividuelles, elle n'apporte aucune information sur la stabilité intra-individuelle, c'est-à-dire sur la stabilité des performances d'un individu. Plusieurs études ont montré que des tests qui présentent une stabilité interindividuelle satisfaisante ($r > 0{,}80$), ne présentent pas pour autant une stabilité intra-individuelle adéquate. En d'autres termes, l'hypothèse de la stabilité des performances d'un individu à partir des données interindividuelles risque d'être non vérifiée, ce qui pourrait conduire à des résultats inexacts.

Proposition

Si nous voulons estimer la validité et la fidélité des scores des tests, il est alors impératif de conduire des études longitudinales et de faire des analyses intra-individuelles. Il faut par exemple déterminer combien d'individus présentent des performances stables ; c'est ce que les *Standards* appellent *Decision consistency*. Nous obtiendrons des données de fidélité/précision vraiment utiles sur le plan clinique. En l'état, les coefficients de fidélité reposent sur des postulats discutables et ne traduisent que la stabilité des différences interindividuelles. En ce qui concerne la validité, nous devons proposer des modèles sur les processus impliqués dans la résolution d'une épreuve et tenir compte des différences interindividuelles qui peuvent survenir. Compte tenu de toutes ces limites, les hypothèses diagnostiques doivent reposer sur l'intégration non seulement des données quantitatives mais aussi qualitatives. À ce jour, la seule utilité des tests réside peut-être dans leur validité prédictive, comme le suggéraient Cronbach et Gleser en 1957 [12] : « *The usefulness of tests in predictive or decision-making contexts is a issue independent of whether or not test measure psychological attributes. . .* ».

Le point de vue du clinicien : à propos du choix des critères normatifs

Préambule

En Suisse, l'Association suisse des neuropsychologues (ASNP), affiliée à la Fédération suisse des

psychologues (FSP), est l'organe officiel regroupant les psychologues travaillant dans le domaine de la neuropsychologie. L'ASNP cherche à actualiser, optimiser et uniformiser les pratiques des cliniciens suisses, actions qui passent notamment par l'édiction de lignes directrices pour les expertises neuropsychologiques en 2011 et la rédaction des rapports neuropsychologiques en 2013 (www.neuropsy.ch/fr/professionnels/assurance-de-qualite). Au plan clinique, un constat réalisé ces dernières années était l'absence de consensus sur la manière de catégoriser les résultats aux tests psychométriques, que ce soit pour le nombre de catégories pour caractériser les résultats (*i.e.* nombre « d'étiquettes » assignables aux scores mesurés), pour la symétrie de ces catégories (*i.e.* catégories en dessous/au-dessus du centile 50) ou encore pour le seuil (centile) définissant un score comme « limite » (*cut-off* entre une performance déficitaire et une performance dans la norme). Partant de ce constat, l'ASNP a mis en place, en 2014, un groupe de travail pour définir des lignes directrices pour la classification et l'interprétation des résultats aux tests neuropsychologiques. Parallèlement à une proposition de consensus (lignes directrices), le document issu des réflexions du groupe de travail précise également les étapes indispensables allant de la « simple » catégorisation des résultats à l'interprétation clinique de l'ensemble de l'examen neuropsychologique [13].

Concept de norme

Les résultats aux tests neuropsychologiques sont généralement obtenus sous forme de scores bruts ne permettant pas une comparaison directe entre les tests, et devant être intégrés dans un système de référence (valeurs standardisées) permettant de classer les performances aux tests pour un individu. C'est pourquoi les scores bruts sont généralement transformés en scores normalisés (ex. score Z) pour permettre, d'une part, la comparaison avec la population de référence et, d'autre part, la comparaison avec d'autres tests. Selon Fisseni [14] et ci-dessus, pages 62-63, les échelles standard les plus courantes sont les normes qui prennent en compte les variables sociodémographiques (telles que le sexe, l'âge ou le niveau socioculturel), les normes d'équivalence et les rangs percentiles. Si le premier type de normes est pertinent pour l'évaluation neuropsychologique chez l'adulte, il est fréquent d'utiliser des normes d'équivalence chez l'enfant et l'adolescent pour déterminer l'âge du développement par rapport à l'âge réel. De plus, chez les patients adultes, les normes sociales atteignent parfois leur limite, par exemple lorsqu'un patient âgé doit être évalué quant à son aptitude à la conduite automobile ou quant à une suspicion de déclin cognitif pathologique, où il devient alors plus pertinent d'appliquer une approche basée sur des exigences minimales à remplir. Dans le cadre de sa pratique clinique, le neuropsychologue doit être conscient de cet aspect et choisir le système de normes le mieux adapté pour répondre à la question posée. Bien que les lignes directrices ASNP pour la classification et l'interprétation des résultats aux tests neuropsychologiques ne traitent pas de la sélection des outils (tests), il est important de rappeler la recommandation d'utiliser autant que possible des tests standardisés et bien normés qui répondent à certaines exigences méthodologiques (avec les réserves mentionnées ci-dessus, pages 64-65). La performance aux tests neuropsychologiques dépend aussi souvent de variables démographiques telles que l'âge, le sexe ou le niveau socioculturel, et il faut alors choisir dans la mesure du possible les normes du groupe de référence adapté à ces variables. Toutefois, pour des questions spécifiques comme par exemple l'intelligence à l'âge adulte, il est important que la comparaison à la population de référence ne prenne plus en compte certaines variables (comme le sexe ou le niveau socioculturel). Par ailleurs, les différents résultats d'un seul test ou d'une batterie de tests n'ont pas tous le même degré de différentiation (ex. effet plafond chez des sujets sains), et parfois seules des valeurs seuils sont disponibles (au-delà desquelles les performances sont jugées déficitaires). Pour les tests avec un échantillon de normalisation plus large, il existe des normes plus fines et l'examinateur bénéficie alors d'informations sur le degré de différentiation par le biais des valeurs standardisées (ex. valeurs Z, valeurs T, *percentile rank*).

Constats et état des lieux en pratique clinique

Dans un premier temps, le groupe de travail s'est penché sur les différences de pratiques. Au niveau des recommandations faites par les auteurs de tests, une grande diversité est observable tant pour le nombre de catégories que pour la définition du seuil limite. Strauss *et al.* [15] proposent une classification en sept catégories avec une valeur seuil (*borderline*) située entre le centile 2 et le centile 8, alors que Fisseni [14] suggère une classification en cinq catégories avec une valeur limite autour du centile 15. Cette dernière proposition est proche de celle faite pour la batterie NAB par Stern et White [16] qui proposent un seuil limite au centile 15 mais un nombre élevé de catégories (dix). Il n'est ainsi pas possible d'utiliser de manière systématique les procédures de classification proposées par les auteurs, de par leur manque d'uniformité et de par le fait que ces classifications sont en général proposées spécifiquement à un test ou à une batterie. En Suisse, on constate une pratique clinique fortement influencée par les particularités linguistiques et hormis quelques exceptions, la tendance majoritaire est l'utilisation du centile 15 comme valeur limite dans la partie germanophone, et des centiles 5-6 comme valeur limite dans la partie francophone. Les germanophones utilisent par ailleurs une classification en cinq catégories, alors que les francophones sont plus proches d'une classification en sept catégories, sachant qu'en principe, le nombre de catégories est lié au nombre de participants (il faut en pratique 5 à 10 fois plus de participants que de catégories). On retrouve ici l'absence d'uniformité déjà relevée dans la littérature psychométrique, motivant d'autant plus la recherche d'un consensus au plan national.

Tableau 1. Classification et interprétation des résultats aux tests neuropsychologiques selon les lignes directrices proposées par l'Association suisse des neuropsychologues (ASNP, 2014).

Percentiles	Valeurs T	Valeurs Z	Description des valeurs statistiques (classification)	Interprétation clinique
100 ↑	80 ↑	3,0 ↑	Très supérieur à la norme	
95 ↑	66,4 ↑	1,6 ↑	Supérieur à la norme	
84 ↑ 16	60 ↑ 40	1,0 ↑ -1,0	Dans la norme	Sans particularité
↑ 5	↑ 33,6	↑ -1,6	Inférieur à la norme	Possiblement déficitaire
↑ 0	↑ 20	↑ -3,0	Très inférieur à la norme	Déficitaire

Le groupe de travail s'est ensuite approché de collègues helvétiques et européens (sondage non représentatif mais illustratif des pratiques) auxquels la question suivante a été posée : « *The Swiss Neuropsychology Society is working on a best harmonization of the neuropsychology evaluation in Switzerland. We would like to know at what threshold (i.e. percentile rank) do you consider that the behavioural measure (i.e. score) is abnormal (i.e. outside the range of normal subjects/general population?* ». Le principal constat dans les réponses était l'absence de consensus également dans les autres pays. Dans les pays germano- et anglophones, la limite de la norme semble généralement être établie au centile 16 (un écart-type en dessous de la moyenne de la population générale). Cette limite des normes est probablement due à la large diffusion de la batterie Halstead-Reitan et de ses normes par Heaton *et al.* [17]. Dans les pays francophones, la limite est plus généralement établie au centile 5. De plus, les centiles 25 ou 10 sont parfois utilisés comme limites par d'autres cliniciens. L'absence de consensus sur ces valeurs limites implique que le lecteur d'un rapport neuropsychologique soit familiarisé à cette problématique de classification des résultats et qu'il connaisse les différences régionales, voire parfois les conventions internes à certaines institutions.

Démarche réflexive et consensus

Après administration des tests, les scores doivent être analysés et interprétés en fonction des normes et standards récents et des connaissances professionnelles. Partant des différentes pratiques suisses et européennes, le groupe de travail s'est alors penché sur la proposition d'un système de classification (catégories et valeurs seuils) généralisable et utilisable par tous les cliniciens et pour tous les tests individuels (la plupart des batteries proposant déjà leur propre système de classification) lors de l'étape initiale de cette démarche clinique. De nombreuses propositions existent pour classifier les résultats dans diverses catégories telles que « *inférieur à la moyenne* », « *dans la moyenne* » ou « *supérieur à la moyenne* » proposées par Brooks *et al.* [18]. Ces propositions peuvent varier d'un faible niveau de différenciation (trois catégories) à un haut niveau de différentiation (neuf catégories). Les lignes directrices ASNP proposent un consensus pragmatique en cinq catégories (*tableau 1*). À noter que les termes « *inférieur à la norme* » et « *très inférieur à la norme* » ne reflètent pas forcément un résultat pathologique ou une dysfonction cérébrale, mais visent à catégoriser ces résultats d'une manière commune.

Lors de l'interprétation des résultats qui suit leur classification, il faut garder à l'esprit quelques principes, comme le fait que plus le nombre de tests utilisés augmente, plus la probabilité d'obtenir des résultats faibles est élevée tel que rappelé par Binder *et al.* [19]. L'interprétation de l'examen dans son ensemble implique que les informations obtenues soient pondérées et intégrées selon leur importance relative. Le DSM-5 [20] intègre le diagnostic de « *trouble cognitif mineur/majeur* ». Les troubles cognitifs majeurs (légers, modérés, sévères) impliquent des performances aux tests inférieures à deux écart-types, alors que pour un trouble mineur, ces performances se situent entre un et deux écart-types en dessous de la moyenne. Le DSM-5 souligne à plusieurs reprises que l'interprétation doit être faite dans le contexte du niveau de performance pré-morbide. Il va de soi que ce postulat est moins pertinent pour les interprétations en neuropsychologie développementale (utilisation privilégiée de normes d'équivalence) ou lorsque plusieurs événements cliniques se succèdent. Il est enfin important de

rappeler que l'interprétation des résultats de patients avec des niveaux d'éducation et/ou d'intelligence très élevés ou très faibles représente une difficulté particulière, de même que pour des patients issus d'un autre milieu culturel et/ou linguistique.

Proposition

Les lignes directrices proposées par l'ASNP constituent une recommandation afin de parvenir à une plus grande cohérence interindividuelle dans la classification et l'interprétation des résultats aux tests. Issues d'un consensus pragmatique et largement soutenu, ces lignes directrices ont pour but d'améliorer la qualité dans ce domaine ainsi que la communication entre spécialistes médicaux et paramédicaux.

Conclusion

Au terme de ce survol méthodologique, il pourrait ressortir l'impression d'une grande faiblesse des outils d'évaluation qui sont traditionnellement utilisés en neuropsychologie clinique. Néanmoins, la perspective de dépasser cette impression par la confrontation de données quantitatives, de mauvaise qualité, selon les différents aspects évoqués (normes comparatives, manque de données concernant la validité et la fiabilité) à des données qualitatives, à l'heure actuelle largement développées grâce aux abondantes recherches en neurosciences, fournit un espoir qui devrait permettre de combler les lacunes durant quelques années. En effet, la création d'outils d'évaluation tenant compte des critères maintenant bien élaborés au niveau méthodologique va certainement occuper les chercheurs durant un certain temps, alors que la nécessité de poursuivre les évaluations, de la façon aussi spécifique que possible, durant tout ce temps, reste de première importance. Les critères de construction de tests sont maintenant à disposition, et les nouveaux tests devront, à terme, éliminer de la fameuse boîte à outils des tests historiquement importants (entre autres les tests d'André Rey, comme la figure complexe ou l'apprentissage d'une liste de mots) mais ne correspondant plus aux critères de qualité actuels. La distinction entre normes comparatives et normes diagnostiques appelle que nous portions une attention particulière à l'objectif de l'évaluation, comme dans les exemples de l'évaluation de suspicion d'une démence ou de l'aptitude à la conduite automobile chez la personne âgée. Les critères de construction de tests valides et fidèles sont maintenant clairement établis, mais il semble particulièrement difficile d'arriver à un consensus aussi important que celui qui est à la base des travaux de Wechsler et qui a mobilisé un nombre très élevé de collaborations. Il n'empêche que l'exemple de la démarche du groupe de travail de l'ASNP pour harmoniser les critères d'évaluation offre un exemple pour des collaborations internationales qui devront, sans aucun doute, se mettre progressivement en place pour parvenir à la meilleure compréhension possible de processus neurodégénératifs ou à l'opposé neurodéveloppementaux, dans le respect de la diversité des comportements humains. ■

Remerciements

Le contenu de cet article est issu d'une double démarche : d'une part, le projet d'un groupe de travail issu de l'Association suisse des neuropsychologues (ASNP) constitué de cliniciens confirmés issus de plusieurs cantons romands et alémaniques : Christian Balzer (Argovie), Adrian Frei (Zürich), Ruth Möckli (Zürich), Andreas Noser (Berne), Arnaud Saj (Genève), Vincent Verdon (Neuchâtel), Roland Vocat (Valais) ; d'autre part, le résultat de la réflexion proposée dans le cadre d'un workshop organisé sous les auspices de la SNLF (Françoise Colombo et Nathalie Rosenau) et de l'HFR (Françoise Colombo et Anne-Isabelle Gasser) en juillet 2015 à Fribourg.

Liens d'intérêts

les auteurs déclarent ne pas avoir de lien d'intérêt en rapport avec ce texte.

Références

1. Folstein M, Folstein S, McHugh P. 'Mini-Mental State'. A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 1975 ; 12 : 189-98.
2. Amieva H, Carcaillon L, Rouze L'Alzit-Schuermans P, *et al*. Cued and uncued memory tests: norms in elderly adults from the 3C epidemiological study. *Rev Neurol (Paris)* 2007 ; 163 : 205-21.
3. Van der Linden M, Adam S, Agniel A, *et al*. *L'évaluation des troubles de la mémoire*. Marseille : Éditions Solal, 2004.
4. O'Connell ME, Tuokko H, Kadlec H. Demographic corrections appear to compromise classification accuracy for severely skewed cognitive tests. *J Clin Exp Neuropsychol* 2011 ; 33 : 422-31.
5. Sliwinski M, Buschke H, Stewart WF, *et al*. The effect of dementia risk factors on comparative and diagnostic selective reminding norms. *J Int Neuropsychol Soc* 1997 ; 3 : 317-26.
6. AERA, APA, NCME. *The Standards for educational and psychological testing*. (AER Association, AP Association, NCoMi. Education, Eds.). Washington : American Educational Research Association, 2014.
7. Markus KA, Borsboom D. *Frontiers of test validity theory*. New York : Routledge, 2013.
8. Borsboom D, Mellenberg GJ, van Heerden J. The concept of validity. *Psychol Rev* 2004 ; 111 : 1061-71.
9. Brennan RL. Commentary on "validating the interpretations and uses of test scores". *J Educ Meas* 2013 ; 50 : 74-83.
10. Michell J. *Measurement in psychology. A critical history of a methodological concept*. Cambridge (UK) : Cambridge University Press, 1999.
11. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas* 2013 ; 50 : 1-73.
12. Cronbach LJ, Gleser GC. *Psychological tests and personal decisions*. University of Illinois Press : Urbaba (USA), 1957.
13. ASNP. *Lignes directrices pour la classification et l'interprétation des résultats aux tests neuropsychologiques*. 2014. www.neuropsy.ch/fr/professionnels/assurance-de-qualite.

14. Fisseni HJ. *Lehrbuch der psychologischen Diagnostik. Mit Hinweisen zur Intervention. Ausgabe 3*. Göttingen : Hogrefe Verlag GmbH, 2004.
15. Strauss E, Sherman EMS, Spreen O. *A Compendium of neuropsychological tests: administration, norms, and commentary*, 3rd ed. New York : Oxford University Press, 2006.
16. Stern RA, White T. *Neuropsychological assessment battery: administration, scoring, and interpretation manual*. Lutz : Psychological Assessment Resources, 2003.
17. Heaton RK, Miller SW, Taylor MJ, *et al*. *Revised comprehensive norms for an expanded Halstead–Reitan Battery: demographically adjusted neuropsychological norms for African American and Caucasian adults professional manual*. Lutz : Psychological Assessment Resources, 2004.
18. Brooks LB, Sherman EMS, Iverson GL, *et al*. Psychometric foundations for the interpretation of neuropsychological test results. In : Schoenberg MR, Scott JG, éds. *The little Black Book of Neuropsychology*. New York : Springer, 2011, p. 892-922.
19. Binder LM, Iverson GL, Brooks BL. To err is human: "abnormal" neuropsychological scores and variability are common in healthy adults. *Arch Clin Neuropsychol* 2009 ; 24 : 31-46.
20. APA. *Diagnostic and Statistical Manual of Mental Disorder*, 5th ed. Washington : American Psychiatric Association, 2013.