Définir ou détecter des pathologies ? Utilisation et interprétation des scores seuils à la lumière du débat dimensions/catégories

Marc Aguert, Aurélie Capel, Arnaud Mortier

PRE PRINT

Aguert, M., Capel, A., Mortier, A. (accepté). *Définir ou détecter des pathologies ? Utilisation et interprétation des scores seuils à la lumière du débat dimensions / catégories*. A paraître dans *Pratiques Psychologiques.*

Définir ou détecter des pathologies ? Utilisation et interprétation des scores seuils à la lumière du débat dimensions / catégories

Defining or detecting pathologies? The use and interpretation of cut-off scores in light of the debate dimensions VS categories

Marc Aguert[1], Aurélie Capel[2] & Arnaud Mortier[1,3]

[1] Normandie Univ, UNICAEN, LPCN, 14000 Caen, France

[2] Normandie Univ, UNICAEN, INSERM, ANTICIPE, 14000 Caen, France

[3] Normandie Univ, UNICAEN, CNRS, LMNO, 14000 Caen, France

Contact :

Marc Aguert

MRSH, Esplanade de la Paix, CS 14032, 14032 Caen cedex 5

E-mail: marc.aguert@unicaen.fr

Définir ou détecter des pathologies ? Utilisation et interprétation des scores seuils à la lumière du débat dimensions / catégories

Résumé : L'objectif de certains tests psychométriques est de positionner la performance du sujet le long d'un continuum allant du normal au pathologique. D'autres tests ont pour objectif de détecter si le sujet est « sain » ou « pathologique » relativement à une pathologie psychologique, psychiatrique ou neurologique, définie indépendamment du test. Les premiers reposent plutôt sur une conception dimensionnelle des psychopathologies ; les seconds impliquent une conception catégorielle. Les premiers supposent de comparer la performance du sujet à celle d'un échantillon normatif représentatif de la population générale ; les seconds supposent de comparer la performance du sujet à celle d'un échantillon normatif constitué uniquement de sujets sains. Les premiers peuvent impliquer l'utilisation d'un score seuil pour signaler qu'un certain niveau de rareté de la performance a été dépassé ; les seconds requièrent obligatoirement l'utilisation d'un score seuil pour contrôler le risque de faire des erreurs de catégorisation (faux positifs et faux négatifs) au moment de conclure si le sujet est dans la catégorie « sain » ou « pathologique ». Avec les premiers, le score seuil définit la psychopathologie ou au moins, il y contribue. Avec les seconds, le score seuil a comme rôle de détecter la psychopathologie. Cet article vise à bien distinguer ces deux familles de tests et à souligner les répercussions aussi bien pratiques que théoriques de leur usage sur la pratique psychologique.

Mots-clés : test psychométrique ; score seuil ; faux positif ; approche dimensionnelle ; approche catégorielle

Defining or detecting pathologies? The use and interpretation of cut-off scores in light of the debate dimensions VS categories

Abstract: A number of psychometric tests are aimed at locating the performance of an individual along a continuous range that goes from normal to worrying. In a different paradigm, there are tests whose goal is to assess whether or not an individual is healthy with respect to a psychological, psychiatric or neurological pathology that is defined independently from the test itself. The former rely on a dimensional concept of psychopathologies; the latter imply a categorical concept. The former require to compare the performance of the individual to a standard sample that is representative of the general population; the latter involve a standard sample representative of the healthy population only. The former might imply the use of a cut-off score to indicate that a certain degree of rarity of the performance has been reached, while the latter must involve a cut-off score to make a decision (whether the individual belongs to the "healthy" or "pathological" category), a score that is most often set so as to control the false-positive rate in making this decision. In the former paradigm, the cut-off score defines or contributes to define the pathology. In the latter, it only helps detecting it. This article aims at making clear the difference between these two realms and highlight their practical as well as theoretical impact on psychological practice.

Keywords: psychometric test; cut-off score; false-positive; dimensional approach; categorical approach.

Définir ou détecter des performances pathologiques ? Utilisation et interprétation des scores seuils à la lumière du débat dimensions / catégories

La pratique de la psychologie suppose fréquemment d'objectiver une plainte du sujet[1] afin de caractériser au mieux la nature et l'ampleur du trouble et d'y répondre par une prise en charge adaptée. Un grand nombre de psychologues appuie en partie cette démarche sur l'utilisation de tests standardisés. Ces tests sont souvent couplés à l'utilisation de scores seuils, score au-delà desquels la performance est jugée « pathologique » ou « déficitaire », dont l'utilisation et l'interprétation peut parfois s'avérer ardue. Dans cet article, nous proposons que l'utilisation de ces scores seuils devrait en particulier être considérée à la lumière de la conception, dimensionnelle ou catégorielle, que la psychologue se fait de la psychopathologie qu'elle est en train d'évaluer.

1. Approches dimensionnelle et catégorielle des psychopathologies

Distinguer une personne avec un fonctionnement psychologique « normal » d'une personne avec un fonctionnement psychologique « pathologique » est un exercice notoirement difficile, que ce soit sur le plan théorique ou sur le plan clinique, et pourtant un enjeu au cœur de la pratique quotidienne des psychologues. En effet, comme on peut le supposer, une personne au fonctionnement « normal » aura moins besoin d'un suivi ou d'une prise en charge qu'une personne au fonctionnement « pathologique ». Les discussions autour de cette question ont été largement structurées par deux positions antagonistes : l'approche dimensionnelle et l'approche catégorielle. Pour les tenants de l'approche dimensionnelle, la pathologie ne constitue ni plus ni moins qu'un extrême du spectre des possibles sur une dimension donnée (e.g., l'intelligence, l'anxiété, etc.). L'ensemble de la population se distribue sur cette dimension, des moins intelligents aux plus intelligents, des moins anxieux aux plus anxieux, etc., et c'est une différence de degrés, quantitative, qui va distinguer les personnes au fonctionnement atypique des personnes au fonctionnement typique comme cela est représenté sur la figure 1A. Le passage du normal au pathologique est progressif bien qu'il puisse être, pour des raisons pratiques, formalisé par un seuil. Cette conception

[1] Dans la suite de cet article, nous désignerons par le mot « sujet » la personne, le patient, le participant testé ; « la psychologue » sera la personne, le clinicien, le praticien, le chercheur qui fait passer le test et l'interprète

dimensionnelle du fonctionnement psychologie atypique est intimement lié au développement de la psychométrie, depuis les premiers travaux d'A. Binet au début du XXe, cette discipline ayant largement recouru à des échelles pour explorer et quantifier le fonctionnement psychologique. Pour les tenants de l'approche catégorielle, davantage calquée sur le modèle médical dominant de la maladie, les personnes avec un fonctionnement pathologique ont un fonctionnement structurellement différent des personnes dites « saines ». En ce sens, elles forment deux catégories qualitativement distinctes, la catégorie des personnes normales et celle des personnes avec une pathologie, comme cela est représenté sur la figure 1B.

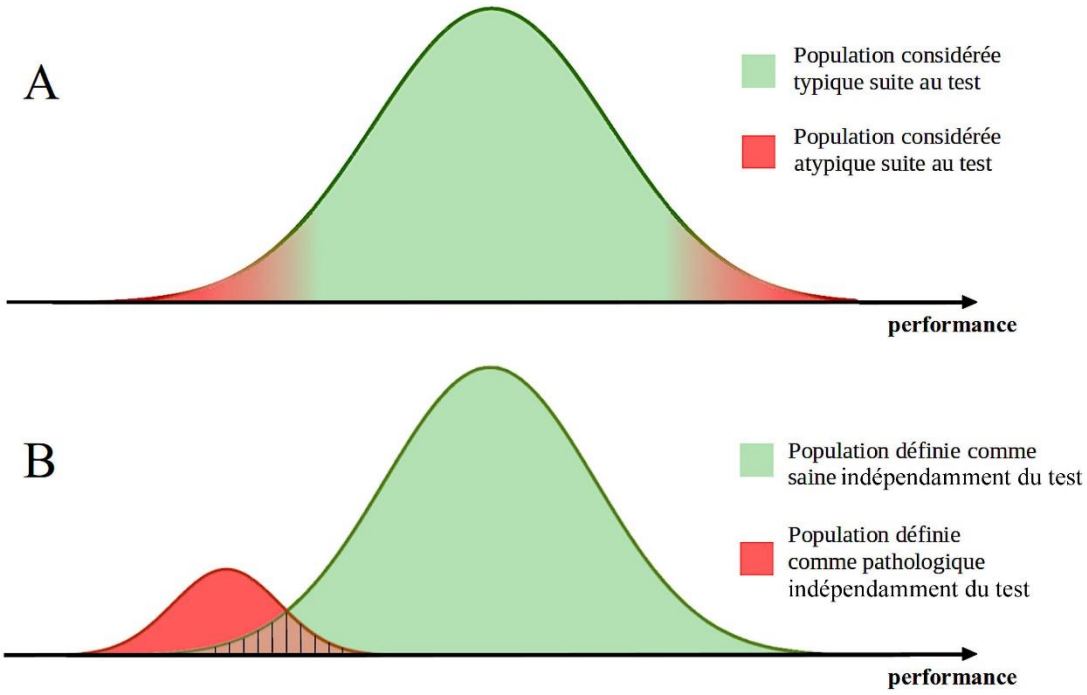

Figure 1. Panel A : approche dimensionnelle. Distribution des scores de l'ensemble d'une population sur une dimension donnée. Le rouge caractérise les sujets dont la performance sur la dimension mesurée est atypique (rare) ; le vert caractérise les sujets dont la performance est typique (fréquente) ; le dégradé entre les deux couleurs symbolise le passage progressif d'une performance typique à une performance atypique.
Panel B : approche catégorielle. En vert est représentée la distribution des scores d'une population saine au regard d'une certaine pathologie, que le test cherche à détecter. En rouge la distribution des scores des sujets présentant la pathologie (ici représentée suivant une distribution normale pour des raisons de simplicité mais souvent inconnue en pratique). Dans la zone hachurée se trouvent à la fois des personnes saines et des personnes présentant la pathologie, mais il est impossible pour un même sujet d'appartenir aux deux catégories à la fois.

Ces deux approches ont suscité plus ou moins d'adhésion à différentes périodes de l'histoire. Les années 1970, 1980, 1990 ont été très largement dominées par la conception catégorielle comme cela est assumé très explicitement dans les quatre versions du DSM (DSM-III, DSM-IV et leurs versions révisées respectives) ayant fait référence ces décennies-là. Mais la conception catégorielle a essuyé des critiques dont les plus accablantes sont (i) l'appartenance aux catégories n'est pas mutuellement exclusive (nombreuses comorbidités) et (ii) l'accord inter-évaluateurs pour les diagnostics est généralement modeste (Adam, 2013; Coghill & Sonuga-Barke, 2012; Hengartner & Lehmann, 2017; Krueger et al., 2005; Sapolsky, 2017). Les années 2000 ont été l'occasion d'un retour en force de l'approche dimensionnelle. Plusieurs méta-analyses ont été publiées défendant une supériorité de l'approche dimensionnelle à la fois sur le plan de la validité et de la fidélité des mesures (Markon et al., 2011) mais aussi sur le plan de l'utilité clinique (Bornstein & Natoli, 2019). Dans ce contexte, la publication du DSM-5 (American Psychiatric Association, 2013) apparaît comme un compromis entre les tenants d'une ligne historique, plutôt catégorielle, et les tenants d'un renouveau qui ont regretté que l'approche dimensionnelle n'ait pas pu être déployée autant que souhaité (Crocq et al., 2015, p. 861; Kupfer & Regier, 2011).

Les psychologues ne sont bien sûr pas tenus de choisir une approche une fois pour toutes et leur conception, plutôt dimensionnelle ou plutôt catégorielle, va surtout dépendre de la psychopathologie considérée. Un tableau clinique comparable à un âge donné peut même donner lieu à deux interprétations concomitantes. Ainsi traditionnellement en France, un enfant de 3 ans présentant des troubles de l'acquisition du langage oral pourra potentiellement être considéré comme ayant un retard de langage (approche dimensionnelle, l'enfant retardé suit les mêmes étapes de développement qu'un enfant typique mais avec un délai) ou comme ayant une dysphasie, si les difficultés persistent dans le temps (approche catégorielle, l'enfant dysphasique suivant une trajectoire de développement atypique) (Piérart, 2004). Plus récemment, pour ce qui concerne les troubles de l'acquisition du langage oral, la conception catégorielle semble avoir perdu du terrain, au moins dans les pays anglo-saxons, comme l'illustre cette citation de Bishop : « There is no support for a sharp dividing line between language disorder and normality. Children develop language at different rates, and some will have problems severe enough to cause lifelong problems. But the differences between children appear to be quantitative rather than qualitative, and can take many different forms » (Bishop, 2017, p. 673). Pour autant, Bishop défend la nécessité d'un label clair pour identifier la catégorie des enfants ayant un trouble de l'acquisition du langage oral (« Developmental

Language Disorder »), communiquer, que ce soit en recherche ou en clinique, mobiliser les financeurs et prendre des décisions, en distinguant a minima les enfants ayant besoin d'aide et les autres.

Toujours à titre illustratif, notons que le débat est particulièrement vif en ce moment sur la meilleure manière de définir l'autisme. Alors que cette condition a largement été décrite par le prisme catégoriel dans les deux dernières décennies du XXe siècle (étiologie largement génétique, pathologie avec laquelle on naît, on vit et on meurt, persistance des troubles, fonctionnement « différent », en opposition avec le fonctionnement « neuro-typique », etc.), la montée en puissance à partir des années 2000 d'une littérature sur les « traits autistiques » a changé la donne. Des études ont mis en évidence que ces traits caractérisant les autistes (compétences sociales pauvres, difficultés pragmatiques et avec la communication non verbale, rigidité, etc.) se distribuaient de manière continue dans la population générale et non d'une manière dichotomique qui révèlerait l'existence de deux catégories distinctes (Constantino & Todd, 2003 ; Posserud et al., 2006 ; Ruzich et al., 2015). L'existence d'un phénotype autistique élargi (« broad autistic phenotype ») qui caractérise une partie des proches des personnes autistes (parents, fratrie, etc.), phénotype se rapprochant de celui de l'autisme mais pas assez marqué pour être caractérisé ainsi semble également brouiller la frontière entre normalité et pathologie (Bolton et al., 1994 ; Rubenstein & Chawla, 2018). La conception catégorielle n'est pas seulement remise en question par la difficulté à établir une frontière nette entre autisme de haut niveau et fonctionnement typique mais également par les nombreuses comorbidités de l'autisme qui rendent floue la frontière entre l'autisme et d'autres troubles neuro-développementaux comme le trouble de l'attention avec hyperactivité (TDA/H, Septier et al., 2019 ; Stevens et al., 2016). En parallèle, la structuration en cinq catégories de TED (Troubles Envahissant du Développement) du DSM-IV-R (autisme typique, syndrome d'Asperger, trouble envahissant du développement non spécifié, syndrome de Rett, syndrome désintégratif de l'enfant) s'est révélée peu valide cliniquement (Kamp-Becker et al., 2010 ; Esler & Ruble, 2015), tant et si bien que ces catégories ont été fusionnées dans le DSM-5 (à l'exception du Syndrome de Rett qui a une étiologie génétique bien identifiée) au sein d'une entité diagnostique unique définie comme un spectre (le Spectre des Troubles de l'Autisme, TSA) avec différents niveaux de sévérité ; cette entité restant qualitativement différente du fonctionnement typique dans le DSM-5. En réaction à la montée en puissance de l'approche dimensionnelle, les tenants d'une approche catégorielle s'inquiètent de ce que l'étude des traits autistiques plutôt que celle de l'autisme comme

catégorie dilue l'effort de recherche en l'éloignant des personnes prototypiques qui en ont le plus besoin (Mottron, 2021).

L'objectif de cet article n'est pas de discuter de la pertinence d'un positionnement plutôt en dimensions ou plutôt en catégories pour telle ou telle pathologie, le positionnement se faisant au cas par cas (quid des personnes alcooliques ? Des personnes transgenres ? Des enfants avec une déficience intellectuelle ? Ceux avec une déficience visuelle ? etc.). Par contre, nous avons souhaité montrer que ce positionnement, généralement discuté sur un plan théorique, avait également un impact sur la pratique des psychologues, sur les plans diagnostique et thérapeutique, quand bien même il était peu ou pas explicité. Sur le plan diagnostique par exemple, « une conception catégorielle des troubles induit que la prise en charge intervient de manière légitime lorsque le diagnostic est identifié et avéré. Dans une perspective dimensionnelle, en revanche, il apparaît non seulement légitime mais peut-être souhaitable d'intervenir sur une dimension avant qu'elle n'atteigne le seuil critique d'un critère psychopathologique » (Perret & Faure, 2006, p. 327). Parmi les implications concrètes sur le plan diagnostique, la manière d'interpréter les scores seuils (« cut-off » en anglais) suite à la mesure de la performance du sujet sur une habileté donnée sera très différente selon qu'on se situe dans le cadre d'une approche catégorielle ou d'une approche dimensionnelle.

2. Les scores seuils

Rappelons d'abord, si cela était nécessaire, que l'évaluation psychologique ne se réduit pas à une question psychométrique et à l'utilisation d'un ou plusieurs tests : l'anamnèse, l'observation de la personne, les échanges avec l'équipe pluridisciplinaire sont autant de briques élémentaires de la réflexion. Nous nous focalisons dans ce travail sur la contribution de la mesure psychométrique dans cette réflexion et plus particulièrement sur l'interprétation des scores seuils. Notons ensuite un paradoxe apparent : Qu'elle ait une conception dimensionnelle ou catégorielle de la pathologie qu'elle approche, la psychologue aura une démarche étonnamment semblable. Même avec une conception catégorielle, elle commencera probablement d'abord par quantifier le comportement du sujet sur une *dimension* donnée à l'aide d'un score, d'un temps de réponse ou d'un nombre d'erreurs ; ceci parce que la plupart des tests de personnalité et la quasi-totalité des tests de performance sont basés sur une logique de quantification. Ensuite, même avec une conception dimensionnelle, pour répondre très pratiquement à la question de « oui ou non, la prise en charge est-elle justifiée ? », la

psychologue va probablement finir par opérer une *catégorisation* de la performance du sujet au test et c'est là qu'interviennent les scores seuils. Ce paradoxe s'explique si l'on distingue clairement le niveau latent et le niveau observé (Markon et al., 2011). Il est en effet possible de défendre théoriquement une conception catégorielle d'une pathologie (niveau latent) tout en mesurant concrètement l'expression de cette pathologie sur des dimensions (niveau observé) ou, à l'inverse, défendre théoriquement une conception dimensionnelle mais opérer avec des catégories dans la pratique (Sonuga-Barke, 1998).

Comme leur nom l'indique, les scores seuils sont des scores au-delà desquels il devient légitime de considérer que la performance du sujet sur la dimension mesurée doit susciter une réaction particulière de la psychologue. Schématiquement, cela revient à tracer une frontière quelque part sur l'axe des abscisses des deux distributions de scores de la figure 1 de manière à distinguer les sujets typiques des sujets atypiques à l'aune de la performance mesurée. Il existe des dizaines de manières de décider de ces seuils (Cizek, 2012 ; Mueller & Munson, 2015). Si le champ des sciences de l'éducation a plutôt été dominé par des pratiques où les seuils étaient déterminés sur la base de critères externes, indépendants des performances des personnes testées (le niveau de connaissance requis pour obtenir un diplôme par exemple, on parle de « tests en référence à un critère »), le champ de la psychologie est largement dominé par une conception où le seuil dépend directement des performances des personnes testées (« tests en référence à une norme »). Cette pratique consiste à constituer un échantillon normatif (ou « échantillon-étalon »), représentatif de la population d'intérêt à laquelle appartiennent les personnes testées. Ceci est une règle d'or maintes fois répétée dans la littérature : les personnes constituant les normes doivent être aussi proches que possible des personnes que l'on envisage de tester (AERA, APA, NCME, 1999 ; Aguert & Capel, 2018 ; Strauss et al., 2006). Puis on observe la distribution des scores dans l'échantillon normatif et on considérera que le seuil correspond par exemple au score séparant les 5% des individus les moins performants des 95% des individus les plus performants. Le seuil est donc relatif et indépendant des standards de réussite ou d'échec des évaluateurs. Pour bien comprendre cette manière de procéder basée sur la distribution des scores des sujets testés, il va nous être utile de rappeler quelques notions de statistiques et de psychométrie. En effet, chaque test utilisant sa métrique propre, les scores seuils ne sont généralement pas exprimés sous la forme de scores bruts (dans la métrique originale du test) mais sous la forme de scores standards.

2.1. Les scores standard

Classiquement, la passation d'un test par un sujet permet d'établir une performance exprimée en score brut : par exemple, un temps de réalisation de la tâche ou un nombre d'items correctement répondus. Mais ce score brut est souvent difficile à interpréter : à partir de combien d'items incorrectement répondus doit-on considérer que le sujet est en difficulté avec la tâche ? C'est la raison pour laquelle les scores bruts sont très souvent convertis en scores standards, c'est-à-dire en scores exprimés sur une échelle ou dans une métrique connue. Plus précisément une distribution de scores standards a toujours la même moyenne et toujours le même écart-type. Il existe plusieurs types de scores standards dont les plus connus sont les scores T ($\mu = 50$; $\sigma = 10$), les QI ($\mu = 100$; $\sigma = 15$) et les scores *z* ($\mu = 0$; $\sigma = 1$)[2]. La connaissance qu'a la psychologue de la moyenne et de l'écart-type de la distribution des scores standards lui facilite grandement l'interprétation de la performance du sujet. Par exemple, si Tom, un garçon de 10 ans, a un score brut de 30 au subtest « code » de la WISC-V, cette performance est difficilement interprétable sauf pour la psychologue très expérimentée avec ce subtest. Au contraire, dire que Tom a un score $z = -1$ permet immédiatement d'interpréter cette performance comme dans la « moyenne basse » puisqu'on sait la moyenne des scores *z* ($\mu = 0$) et leur écart-type ($\sigma = 1$). Dans cet article, nous nous focalisons sur les scores *z* dans un souci de simplicité mais les différents types de scores standards fonctionnent sur la même logique de paramètres statistiques (moyenne et écart-type) tenus constants. La formule de « standardisation », où m_x et s_x représentent la moyenne et l'écart-type d'une distribution de scores x (scores bruts) est :

$$(1) \qquad z = \frac{x - m_x}{s_x}$$

L'utilisation de scores standards facilite les discussions autour des scores seuils. En effet, la valeur du score seuil sera exprimée indépendamment du test finalement retenu par la psychologue. Les scores seuils les plus couramment rencontrés sont $z = \pm 1{,}65$ et $z = \pm 1{,}96$[3]. La connaissance de ces seuils va aider la psychologue à tirer ses conclusions, confronter les résultats de son test à des informations d'autre nature et rédiger son bilan. Mais d'où viennent ces seuils ? Beaucoup de psychologues savent que le choix de ces scores seuils a à voir avec la distribution normale des scores *z* et au travail de mathématiciens comme C. F. Gauss (1777-

[2] En statistique, on désigne par des lettres grecques les paramètres des populations et par des lettres latines les statistiques des échantillons pour bien les distinguer. Pour la population, on note la moyenne « μ » et l'écart-type « σ ». Pour les échantillons issus de cette population, on note les moyennes « *m* » et les écart-types « *s* ».

[3] Le signe « ± » indique que l'intérêt de la psychologue peut tantôt être lorsque les performances du sujet sont supérieures à $z = 1{,}65$ (quand on comptabilise des erreurs par exemple), tantôt lorsque les performances sont inférieures à $z = -1{,}65$ (quand réussir la tâche fait gagner des points par exemple).

1855). C'est en effet le cas. Dès lors que les scores *z* se distribuent de manière normale (i.e., en suivant une loi normale)[4], il est possible de savoir précisément quelle est la proportion de la population de référence qui se situe entre tel et tel scores *z*. Ces correspondances entre scores *z* et proportions/probabilités sont rapportées dans la « table des probabilités de la loi normale », table que l'on retrouve à la fin de tous les bons manuels de statistiques. La figure 2 ci-dessous est une représentation classique de ces correspondances.

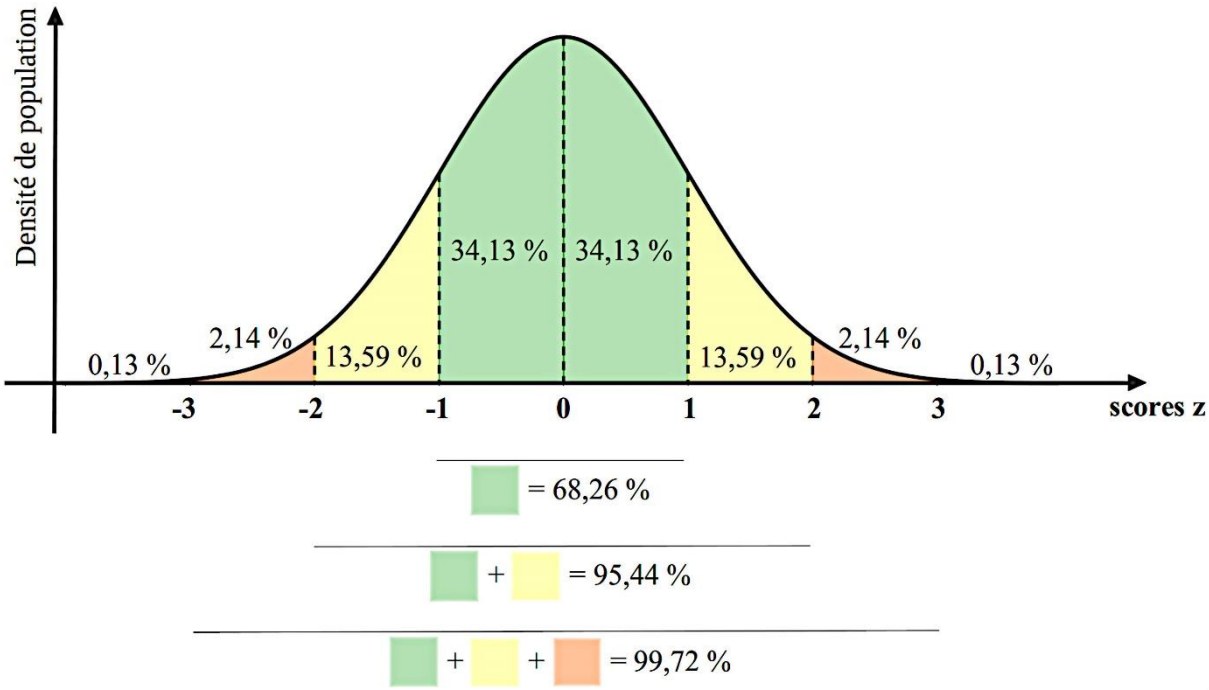

Figure 2. Distribution d'un score *z* qui suivrait une loi normale standard parfaite et probabilité/proportion de sujets se situant à différents endroits de la distribution.

La table de probabilité de la loi normale standard nous indique que 5% de la population de référence aura un score $z \leq -1{,}65$ et 2,5% aura un score $z \leq -1{,}96$. L'utilisation de ces seuils comme frontière entre ce qui est typique et ce qui ne l'est plus a donc à voir avec ces proportions et la notion de rareté. Mais, de manière cruciale, il ne s'agit pas de la même

[4] Ce qui n'est pas systématiquement le cas ! Rappelons que l'application de la formule de conversion des scores bruts en scores *z* n'affecte pas la forme de la distribution et si, pour un score donné, la distribution des scores bruts n'est pas normale, la distribution des scores *z* ne le sera pas non plus. Néanmoins, l'hypothèse que les scores *z* se distribuent en suivant une loi normale est raisonnable quand on s'intéresse à des variables « naturelles », physiques ou psychologiques comme la taille, l'intelligence, l'extraversion, etc. Cela reste une hypothèse qui est loin de se vérifier systématiquement et qu'il convient de tester, en particulier quand on mesure sur une échelle des construits que l'on a théorisés dans une perspective catégorielle.

rareté selon que l'on se place dans l'approche dimensionnelle ou dans l'approche catégorielle. Avant d'explorer plus loin cette idée, arrêtons-nous sur une dernière notion importante à la compréhension du propos : l'erreur de mesure.

2.2. Le problème de l'erreur de mesure

La mesure de la performance du sujet sur un test particulier à un moment t peut ne pas refléter sa compétence, ne pas mesurer son score « vrai » avec suffisamment de précision. En effet, d'une situation à l'autre, sur un même test, la performance du sujet peut varier, à cause de facteurs liés à l'environnement de test (bruit, température, etc.), de facteurs liés au testeur (sa présence, ses feedbacks, etc.), au testé (sa motivation, son stress, son état émotionnel, sa compréhension des consignes, etc.) et à l'interaction entre ces deux protagonistes. Ces différents facteurs génèrent de l'erreur de mesure et expliquent qu'une performance n'est jamais parfaitement reproductible à l'identique. Les tests qui génèrent le moins d'erreur de mesure sont les tests qui sont dits les plus *fidèles*. La fidélité d'un test peut être évaluée avec un simple coefficient de corrélation entre deux mesures au même test (ou des versions parallèles du même test) chez le même sujet. Plus un test est fidèle, plus il produit des mesures stables, reproductibles. La fidélité des tests est donc un indicateur important qui est rapportée dans les manuels de tous les tests bien conçus.

Prendre une décision quant au caractère typique ou non de la performance du sujet en comparant son score observé standardisé à un score seuil peut conduire la psychologue à faire un *faux positif* (conclure que le sujet est atypique alors que ce n'est pas le cas) car le score observé peut être inférieur au score seuil mais pas le score vrai ! A l'inverse, le score observé peut être supérieur au score seuil mais pas le score vrai, ce qui conduirait la psychologue à faire un *faux négatif* (conclure que le sujet n'est pas atypique alors que c'est le cas). La valeur du score vrai ne peut pas être connue avec précision mais elle peut être approchée avec le calcul d'un intervalle de confiance à 95%. Pour calculer un intervalle de confiance à 95% de score *z*, on utilise la formule (2) ci-dessous où *r* est le coefficient de fidélité du test utilisé [5].

(2) $IC_{95} = z \pm 1{,}96\sqrt{1-r}$

Revenons à l'exemple de Tom qui a z = -1 au subtest « code » de la WISC-V. Le manuel d'interprétation précise qu'à 10 ans, le coefficient de fidélité de ce subtest est r = 0,79. En

[5] Dans cette formule, le signe « ± » suppose qu'on fasse une addition pour calculer la borne supérieure de l'IC et une soustraction pour calculer la borne inférieure de l'IC.

appliquant la formule ci-dessus, le psychologue obtient l'intervalle de confiance à 95% suivant : [-1,88 ; -0,12] pour le score vrai de Tom. L'interprétation de cet intervalle non négligeable est délicate. Ce n'est pas, comme on le pense assez intuitivement, l'intervalle dans lequel le score vrai de Tom a 95% de chances de se trouver. L'interprétation alambiquée mais correcte est la suivante : si l'on construisait 100 intervalles de confiance à 95% pour le score vrai de Tom sur la base de 100 scores observés, on attendrait que 95 de ces intervalles contiennent en effet le score vrai de Tom. Nous pouvons donc être confiant (à 95% précisément) dans le fait que le score vrai de Tom se situe quelque part dans l'intervalle [-1,88 ; -0,12] mais il est impossible d'affirmer où.

3. Approche dimensionnelle : un score seuil pour définir la pathologie

A quoi correspond le score seuil (e.g. z = -1,65) dans la conception dimensionnelle ? La logique est basée sur la notion de normalité statistique qui remonte probablement au mathématicien belge Adolphe Quetelet (1796-1874) et à son concept « d'homme moyen ». La normalité « statistique » implique que les scores les plus rares, les plus éloignés de la norme, au sens de la moyenne statistique, sont les scores atypiques. Pour une dimension (e.g. l'intelligence) qui se distribue de manière normale dans la population générale, les scores les plus rares se retrouvent aux deux extrémités de la distribution. Si l'on se focalise sur la partie gauche de la distribution (les scores les plus faibles), plus on s'éloigne de la moyenne, plus on s'approche d'un profil de sujet avec un handicap intellectuel. S'il est entendu que le passage d'un profil typique à un profil atypique est progressif, il est néanmoins souvent utile de définir un seuil en-deçà duquel le manque d'intelligence est si important qu'il nécessite une prise en charge, un suivi dans une classe spécialisée, une notification MDPH, etc. Ce seuil est déterminé par le niveau de rareté au-delà duquel la communauté des psychologues considère que la performance est atypique, généralement 5%. Il faut insister sur ce que cette proportion de 5% a d'arbitraire. Probablement ancrée dans l'histoire de la statistique (dans les années 1920, R. Fisher produit des tables statistiques avec trois niveaux de rareté, 1%, 2% et 5%, Field et al., 2012), elle a depuis été ré-affirmée et légitimée par des conférences de consensus dans les communautés francophone (Colombo et al., 2016) et anglophone (Guilmette et al., 2020) mais elle n'en demeure pas moins susceptible de varier, d'une époque à l'autre et d'une dimension à l'autre. Si l'on considère cette proportion de 5% des personnes de manière bilatérale (i.e., en considérant la rareté des deux côtés de la distribution normale), alors cela correspond à un score seuil $z = \pm 1{,}96$ (cf. figure 3B) tandis que si l'on considère cette

proportion de personnes de manière unilatérale (i.e., en considérant la rareté d'un seul côté de la distribution normale), alors cela correspond soit à un score seuil z = -1,65, soit à un score seuil z = 1,65 (cf. figure 3A).

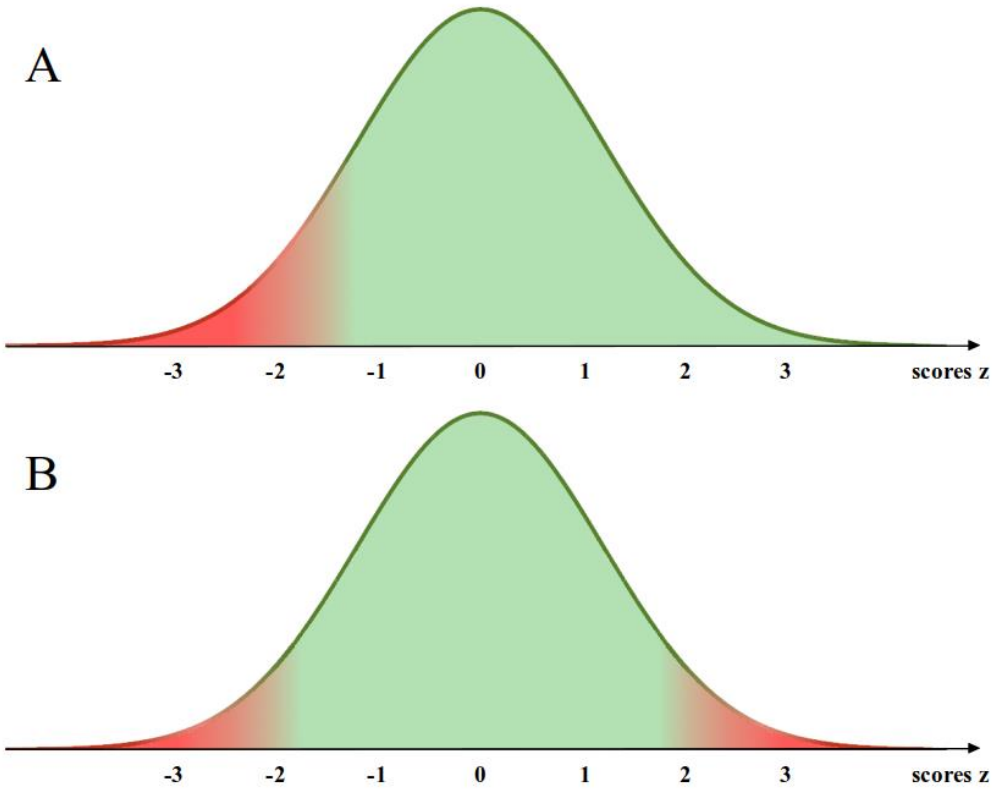

Figure 3. Distribution des scores *z* typiques (en vert) et atypiques (en rouge) respectivement dans une perspective unilatérale (panel A) et bilatérale (panel B). Dans la perspective bilatérale, les scores atypiques sont les scores les plus rares de la distribution. Dans la perspective unilatérale, les scores atypiques sont les scores qui sont à la fois les plus rares *et* les plus faibles (ou, selon le contexte, les plus rares *et* les plus forts).

Le score seuil permet donc de délimiter une certaine proportion de la population générale (généralement 5%) dont les scores sont potentiellement préoccupants au regard de leur rareté, de leur éloignement de la norme statistique. Historiquement, la rareté du score a longtemps suffi à le rendre préoccupant voire carrément à le qualifier de « déficitaire » ou « pathologique », en particulier lorsqu'il était faible. De nombreuses pathologies, à commencer par la déficience intellectuelle, étaient purement et simplement définies par le fait de se situer en dessous de tel ou tel seuil sur la distribution des scores. Ainsi Goddard (1920) définissait les « idiots » comme les personnes ayant un âge mental inférieur à 3 ans, les

« imbéciles » comme les personnes ayant un âge mental entre 3 et 7 ans et les « faibles d'esprits » (« morons » en anglais) comme les personnes avec un âge mental entre 8 et 12 ans. Un siècle plus tard, les psychologues inscrits dans une perspective dimensionnelle prennent garde à ce que le fait d'avoir une performance inférieure à un seuil arbitraire ne définisse pas en soi un fonctionnement troublé ou déficitaire. Récemment, des auteurs restituant les réflexions d'une conférence de consensus sur l'étiquetage des performances à des tests précisaient « the labels do not convey impairment or other evaluative judgments; scores in isolation cannot be impaired or deficient » (Guilmette et al., 2020, p. 445). Dans cette perspective, ils proposaient d'étiqueter les performances situées au-delà de 2 écart-types de la moyenne de « scores exceptionnellement bas / hauts ». Il demeure que la performance à des tests, rapportée à un seuil, continue à être utilisée pour définir, au moins en partie, des dysfonctionnements et des pathologies. C'est ainsi que le « handicap intellectuel » est défini dans le DSM-5 (American Psychiatric Association, 2013) par trois critères diagnostiques dont deux (les « déficits dans les fonctions intellectuelles » et les « limitations significatives du comportement adaptatif en général ») sont objectivés par une performance inférieure à la moyenne de la population générale « d'environ deux écarts-types » à un test standardisé.

En somme, dans l'approche dimensionnelle, si le sujet a une performance inférieure à un certain score seuil, défini comme un niveau de rareté, cela peut conduire à qualifier son fonctionnement de « pathologique » ou « déficitaire »… dans la mesure où la pathologie en question est notamment *définie* par le fait d'avoir une performance inférieure à un certain score seuil ! Dans les autres cas, on se contentera de constater que son score est « exceptionnellement haut » ou « exceptionnellement bas ». Dit autrement, et c'est là le principal point de vigilance à retenir pour cette approche dimensionnelle basée sur la notion de normalité statistique : tout ce qui est rare n'est pas préoccupant ou *a fortiori* déficitaire. Courir le 100 mètres en moins de 12 secondes ou devenir président de la République sont des performances relativement rares et qui, pour autant, ne nécessitent pas une prise en charge psychologique.

Adopter cette approche suppose d'être vigilant sur au moins trois autres points :

1. Les performances ne se distribuent pas toujours selon une loi normale. Le cas échéant, les scores les plus rares ne sont plus automatiquement les plus faibles (et/ou les plus forts). Quand bien même ce serait le cas, l'utilisation de la table de probabilités de la loi normale serait compromise. La psychologue doit alors se rabattre vers la méthode des quantiles qui permet également de déterminer

si le sujet se situe dans les 2,5% (ou autre proportion au choix) des personnes avec les scores les plus faibles.

2. La performance observée n'est jamais qu'une estimation du « score vrai » du sujet. Un score observé proche du score seuil doit toujours être interprété avec une grande prudence car non seulement le seuil n'est qu'une frontière tracée en pointillés sur un continuum pour aider à penser la prise en charge des personnes, mais en plus car le score « vrai » du sujet peut très bien se situer de l'autre côté du seuil ! Dès lors que l'intervalle de confiance du score « vrai » du sujet est à cheval sur le score seuil, il devient problématique de tirer une conclusion définitive de la simple passation du test.
3. Identifier les 5% (ou autre proportion) de la population ayant les scores les plus faibles dans l'objectif de leur offrir un accompagnement spécifique n'est effectif que si le test utilisé est *sensible*. Au sens psychométrique classique du terme, la sensibilité est définie comme la capacité d'un test à correctement discriminer des sujets dont la performance est proche sur une échelle. Un test sensible est donc un test proposant un continuum de performances possibles le long duquel les personnes testées se répartissent harmonieusement. Si le test utilisé n'est pas du tout sensible, qu'il est par exemple très difficile et que 90% de la population se retrouve avec une performance très basse (effet « plancher »), identifier les 5% des personnes avec les performances les plus basses n'est pas très pertinent. Ainsi, le fait que le score seuil est déterminé en référence aux performances des sujets eux-mêmes, indépendamment des standards des évaluateurs (cf. supra), ne dispense pas d'une réflexion sur le contenu de la tâche proposée aux sujets (Cizek, 2012).

Une dernière implication de l'approche dimensionnelle doit être soulignée qui concerne la manière dont sont constitués les échantillons normatifs permettant de calculer les scores *z* du sujet que l'on souhaite évaluer. Dans l'approche dimensionnelle, l'échantillon normatif implique des participants représentatifs de la population générale, tout-venante. Si le test a vocation à tester n'importe quel adulte de la population générale, conformément à la règle d'or de constitution de l'échantillon normatif énoncée plus haut (cf. section 2), alors ce dernier doit être au maximum représentatif de la population générale. Attention, composer un échantillon normatif de manière à ce qu'il soit représentatif de la population générale n'est pas synonyme de recruter n'importe quel participant de manière aveugle. Des critères

d'exclusion légitimes existent dès lors que la personne va probablement échouer le test pour d'autres raisons qu'un déficit dans l'habileté mesurée. Inclure des personnes qui ne maîtrisent pas la langue utilisée dans le test ou qui ont des troubles sensoriels non corrigés biaiserait la validité de la mesure et donnerait une image déformée du niveau de performance dans la population générale pour l'habileté d'intérêt. Mais si l'on s'intéresse par exemple à l'intelligence, on inclura dans l'échantillon des personnes de tous QI, y compris potentiellement des personnes avec de très faibles QI. C'est une condition *sine qua non* avant d'établir que, par exemple, les 2,5% des QI les plus bas (QI < 70 ; $z < -2$) sont ceux des personnes ayant une intelligence atypiquement faible (toujours du point de vue de la normalité statistique). Le manuel d'interprétation de la WISC-V précise par exemple « Pour chaque groupe d'âge, ont été inclus dans l'échantillon d'étalonnage deux enfants souffrant d'Handicap intellectuel et deux enfants identifiés comme ayant un Haut potentiel afin de refléter la population des enfants français dans son ensemble » (Wechsler, 2016, p. 38). Les choses sont tout à fait différentes dans l'approche catégorielle.

4. Approche catégorielle : un score seuil pour détecter la pathologie

Dans l'approche catégorielle, il n'est pas question de sujets plus ou moins (a)typiques au regard d'une norme statistique comme dans l'approche dimensionnelle, mais de sujets issus de deux catégories distinctes (cf. Figure 1B). Il peut s'agir d'être porteur d'une maladie vs. non porteur, porteur d'une lésion cérébrale vs. non porteur, porteur d'une anomalie génétique vs. non porteur, etc. Reprenant la terminologie médicale d'où vient cette approche, nous parlerons ici de la catégorie des sujets « sains » (« healthy » en anglais) vs. la catégorie des sujets « pathologiques ». Le caractère sain ou pathologique d'un sujet est indépendant du test lui-même, test dont la fonction se borne à *détecter* la catégorie à laquelle appartient le sujet, en évitant si possible de faire trop d'erreurs de catégorisation (faux positifs et faux négatifs). La démarche est donc diagnostique au sens classique de la reconnaissance d'une maladie ou d'une condition d'après ses manifestations observables. Le rôle clé du score seuil dans ce cadre est précisément de distinguer les sujets sains des sujets pathologiques sur la base de la performance mesurée. Dans la figure 4A par exemple, les choses sont simples : $z = -4$ est le score z tel que tous les sujets pathologiques sont en-dessous et tous les sujets sains sont au-dessus. Ce score $z = -4$ est donc le candidat idéal pour servir de seuil. Mais dans la réalité, les choses sont moins nettes et il est fort probable que les performances des sujets

pathologiques et celles des sujets sains se superposent au moins en partie pour la tâche utilisée (cf. panels B, C, D et E de la figure 4). Où alors placer le seuil ?

Une connaissance précise des performances au test des deux populations permet à la psychologue de choisir le seuil qu'elle juge le plus pertinent, en fonction de ses priorités : diminuer le risque de faux positifs ou celui de faux négatifs (puisque généralement réduire l'un augmente l'autre, cf. figure 4). Dans le champ de la santé, plus un test limite le nombre de faux positifs, plus on dit de lui qu'il est *spécifique*. La spécificité renvoie à la capacité du test à indiquer que vous avez la Covid-19 uniquement quand c'est bien le cas, i.e. sans inclure des personnes négatives. Mais votre test doit également limiter au maximum le nombre de faux négatifs et correctement indiquer que vous êtes positif à la Covid-19 quand c'est le cas, i.e. sans exclure des personnes positives ! Si le test a une bonne capacité à limiter les faux négatifs, on dit de lui qu'il est *sensible*. Il est crucial de noter, pour éviter toute ambiguïté, que le terme « sensibilité » recouvre deux définitions différentes : la sensibilité « psychométrique » (cf. section 3) est la capacité à discriminer des sujets dont la performance est proche sur une échelle, capacité pertinente dans le cadre d'une approche dimensionnelle ; la sensibilité qu'on appellera ici « diagnostique » est la capacité du test à ne pas conclure qu'une personne pathologique est saine, capacité pertinente dans le cadre d'une approche catégorielle. Dans les deux cas, il est question du pouvoir discriminant du test mais la perspective est différente.

Dans l'approche catégorielle, plus la sensibilité augmente, plus la spécificité diminue et vice-versa. Une technique statistique permettant de trouver le meilleur compromis entre sensibilité et spécificité est la courbe ROC. Cette courbe quantifie le risque relatif de faire des erreurs de catégorisation (faux positifs et faux négatifs) pour différentes valeurs du score seuil (voir Delacour et al., 2005, pour un article introductif à cette technique en français). Choisi sur la base d'une courbe ROC, le score seuil permet donc de conclure quant à la catégorie d'appartenance du sujet testé, sain ou pathologique, en contrôlant précisément les risques de faire une erreur de catégorisation en produisant cette conclusion. Par exemple, sur la figure 4B où le score seuil est fixé à z = -1,65, la psychologue peut affirmer qu'une personne ayant un score z < -1,65 est dans la catégorie des personnes avec pathologie et ce, avec un risque de faire un faux positif de 5 chance sur 100 (puisque 5% des personnes saines sont sous le seuil) et un risque de faire un faux négatif de 0 chance sur 100 (puisque 0% des personnes avec la pathologie sont au-dessus du seuil). Abaisser le score seuil à z = -2,65 pourrait être judicieux

puisque cela réduirait de beaucoup le risque de faire un faux positif, quitte à augmenter un peu le risque de faire un faux négatif.

Le recours aux courbes ROC n'est pas toujours possible faute d'information sur la performance au test (moyenne, écart-type, etc.) des personnes appartenant à la catégorie « pathologie ». En pratique, il est toutefois toujours possible de ranger le sujet testé dans une catégorie en comparant son score au score seuil. Pour cela, il suffit d'avoir une bonne connaissance des performances à la tâche utilisée par les personnes saines, i.e., sans pathologie. L'échantillon normatif doit ainsi être constitué de personnes pour lesquelles on a des garanties qu'elles n'appartiennent pas à la catégorie des personnes pathologiques. Le recrutement de ces personnes implique tout un éventail de critères d'exclusion de manière à garantir que les normes sont représentatives de la population saine, évitant les personnes dont les performances pourraient être dégradées, en particulier dans le domaine d'intérêt. C'est ainsi par exemple que pour les normes du GREFEX (Roussel & Godefroy, 2008), les personnes souffrant d'un trouble de l'usage de l'alcool, trouble connu pour détériorer les fonctions exécutives, sont a priori exclues de l'échantillon normatif. Autre exemple : un test visant à détecter des démences chez les personnes âgées (Raoux et al., 2014) exclut de son échantillon normatif toutes les personnes âgées institutionnalisées, une manière résolue de garantir le recrutement de personnes « saines » au regard de la pathologie en question ici.

Une fois obtenue une estimation fiable de ce que sont les performances des personnes saines, la psychologue va poser le score seuil permettant de distinguer les personnes saines des personnes pathologiques, relativement à leur performance au test. Ignorant les performances des personnes pathologiques, l'exercice est délicat. Partant du principe vraisemblable que les personnes pathologiques auront en moyenne une performance inférieure aux personnes saines de l'échantillon normatif, la psychologue va opter pour une performance basse, mais laquelle ? La logique suivie est identique à celle des tests d'hypothèses utilisés par les statisticiens fréquentistes : il va s'agir de rejeter l'hypothèse (dite « nulle ») selon laquelle le sujet testé est sain, pour accepter l'hypothèse alternative qu'il est pathologique. Et ce, en limitant la probabilité de commettre un faux positif à un niveau raisonnable, la probabilité de 5% étant le standard actuel en psychologie (cf. supra). La probabilité de 5% correspondant à un score $z = -1,65$ (cf. table de probabilité de la loi normale), celui-ci sera retenu comme score seuil. Comme on le voit sur les figures 4B, 4C ou 4D, 95% de la population saine aura un score $z > -1,65$ et 5% aura un score $z < -1,65$. Ainsi, conclure que toutes les personnes ayant un score $z < -1,65$ sont dans la catégorie pathologique

entrainera 5% de faux positifs, i.e. classera dans la catégorie « pathologique » 5% des personnes saines (une personne saine toutes les 20 personnes saines testées). En résumé, la psychologue aura atteint son objectif d'identification des personnes « pathologiques » tout en limitant à 5% le risque de faire un faux positif. Opter pour une performance plus basse comme seuil réduira le risque de faux positif et opter pour une performance plus haute va l'augmenter.

Bien que courant, cet emploi des scores seuils dans une logique catégorielle, basé uniquement sur la connaissance que l'on a des performances des personnes saines présente plusieurs limites. D'abord, parce que la psychologue ignore les performances des sujets pathologiques, elle ne contrôle que le risque de faire des faux positifs et pas le risque de faire des faux négatifs. Celui-ci est inexistant sur les figures 4A et 4B, très faible sur le panel C, mais considérable (50% !) sur le panel D. Face à ce problème posé par la prévalence des faux négatifs, il est essentiel de ne rien conclure si le sujet testé a un score *z* supérieur au seuil retenu, surtout pas qu'il est sain, car on n'aurait alors aucun contrôle sur le risque de faire un faux négatif induit par cette affirmation.

Ensuite, on observe des cas où non seulement les performances de la population pathologique ne sont pas bien connues mais où la pathologie que le test est censé détecter n'est elle-même pas clairement définie ou identifiée par les utilisateurs du test. L'objectif semble alors être de recourir au test (et à sa logique de rejet de l'hypothèse nulle que le sujet testé est sain) pour identifier des personnes « non saines », plutôt que clairement « pathologiques », personnes susceptibles d'avoir besoin d'une prise en charge ou d'un accompagnement. Cet usage des scores seuils conduit à la constitution d'une pathologie ad hoc, contrepartie logique au fait d'avoir construit une catégorie des personnes « saines », sans qu'on sache précisément ce qui réunit ces personnes d'un point de vue nosologique, hormis le fait d'avoir échoué à un test ou une série de tests. Parmi ces personnes, 5% sont des faux positifs mais sans qu'on sache positif à quoi ! En fin de compte, la catégorie des personnes « non saines » n'existe pas indépendamment du test puisqu'elle est *définie* par le test lui-même (par le fait d'avoir une performance inférieure au score seuil). Cela dévoie la visée diagnostique originale de l'approche catégorielle puisque le test définit la catégorie « pathologie » en même temps qu'il la détecte, de manière circulaire.

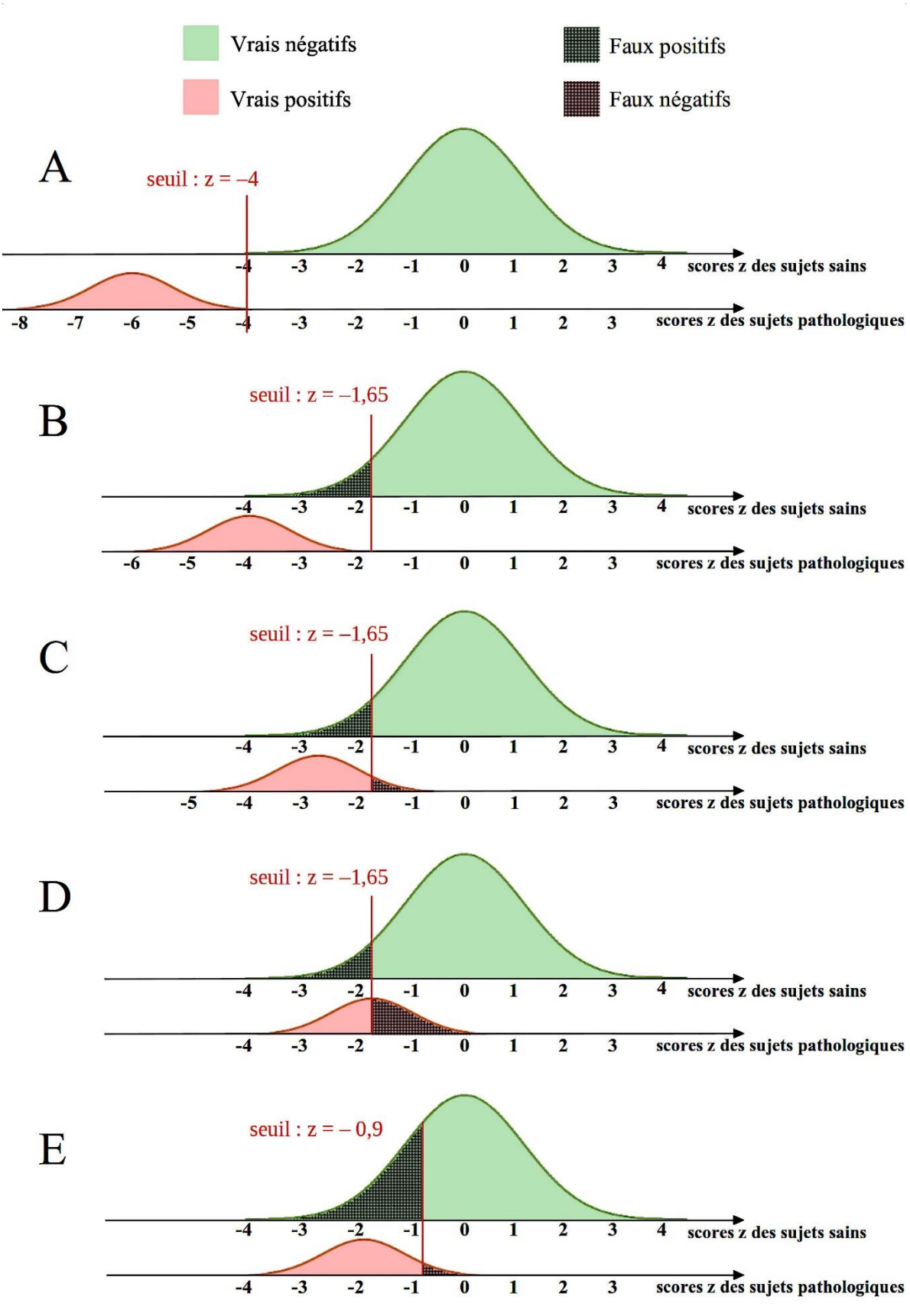

Figure 4. Distribution des scores *z* de la population saine (en vert) et de la population pathologique (en rouge)

selon 5 configurations différentes (panels A à E). Les hachures signalent les personnes qui ont été l'objet d'une erreur de catégorisation (car leur performance tombe du « mauvais côté » du seuil).

Note : Tous les scores *z*, ceux des sujets sains comme ceux des sujets pathologiques, sont calibrés sur la population saine et non la population totale. Les scores *z* des sujets pathologiques ont été représentés suivant une distribution normale pour des raisons pratiques mais d'autres types de distribution sont très plausibles.

L'approche catégorielle implique plusieurs points de vigilance :

1. Comme souligné ci-dessus, cette approche est d'autant plus pertinente que les deux catégories de population sont bien définies indépendamment du test lui-même, en particulier la population « pathologique » (via des critères nosologiques bien établis par une classification diagnostique comme la CIM-11 par exemple). Non seulement cela permet d'établir la sensibilité diagnostique du test et d'établir le score seuil de manière informée avec une courbe ROC mais cela évite de créer une catégorie de personnes aux contours mal définis, les personnes « non saines ». Soulignons qu'en tant que tel, le test (la tâche et les normes pour cette tâche dans la population saine) ne nous apprend rien sur la population pathologique au-delà de l'appartenance d'un sujet donné. Ni la proportion de la population totale qui peut être catégorisée comme pathologique (peut-être 0,5%, peut-être 1% ou peut-être 10% !) ni la performance moyenne des sujets pathologiques, ni la forme de la distribution de leurs scores[6] (cf. légende figure 4).
2. Le test ne nous apprend rien non plus sur la performance individuelle de la personne testée si celle-ci s'avère faire partie de la catégorie des personnes pathologiques. Pour rappel (cf. section 2), si l'échantillon normatif est composé de *sujets sains*, le test ne permet que de caractériser la performance de *sujets sains*. Le rejet de l'hypothèse nulle selon laquelle le sujet testé est sain implique immédiatement que la comparaison de ce sujet à un échantillon normatif composé de sujets sains n'est pas pertinente. Formellement, que la performance d'un sujet pathologique se situe à -2 ou -3 écart-types de la

[6] Si la distribution des scores chez les sujets atypiques est généralement inconnue sans que cela n'impacte le raisonnement, il est par contre primordial, comme dans l'approche dimensionnelle, que la distribution des scores dans l'échantillon normatif (distribution des scores chez les sujets sains) suive une loi normale. Si ce n'est pas le cas, l'utilisation de la table de probabilités de la loi normale pour déterminer des scores *z* est compromise. Il reste possible de déterminer un score seuil avec la méthode des quantiles.

moyenne des sujets sains n'est pas informatif et ne devrait pas participer de l'interprétation de la performance puisque cela revient à comparer une personne dans une catégorie (les personnes pathologiques, les choux) à des personnes d'une autre catégorie (les personnes saines, les carottes).

3. Dès lors qu'on adopte une conception catégorielle, la question des erreurs de catégorisation devient centrale. Or, le risque de commettre des erreurs de catégorisations augmente avec la multiplication des tests puisque l'on reprend un risque à chaque fois. Un sujet sain qui passerait 10 tests (non corrélés entre eux) aurait non pas 5% de chances d'être catégorisé au moins une fois comme pathologique mais près de 40% (Ingraham & Aiken, 1996). De quoi semer le doute dans la tête du psychologue sur l'intégrité psychologique d'un sujet pourtant bien sain. De nombreux articles ont été publiés sur cette question (e.g., Binder et al., 2009; Decker et al., 2012; Roussel & Godefroy, 2016) sans qu'une solution ne s'impose à ce jour.
4. Nous avons vu que la décision de rejeter l'hypothèse selon laquelle le sujet était sain se basait sur son score observé à, au minimum, une tâche. Or, comme souligné dans la section 2.2., ce score observé peut ne pas refléter précisément son score « vrai ». Dans l'approche catégorielle, le recours à un raisonnement de type « rejet d'hypothèse » conduit à mettre le focus sur le risque de faire des faux positifs, risque que l'on souhaite contrôler précisément, en le maintenant à 5% par exemple. Or, il n'est pas possible de définir précisément le niveau de risque en raisonnant avec le score vrai du sujet, score inconnu dont on sait juste qu'il est « quelque part » dans un intervalle de valeurs (l'intervalle de confiance). On raisonne donc avec le score observé, qui est une observation précise, mais on ferme ainsi les yeux sur le fait que le testing implique de l'erreur de mesure (cf. supra).

Pour résumer, dans le cadre de l'approche catégorielle, le score seuil identifie bien une forme de rareté, mais pas la rareté avec laquelle une performance faible peut apparaître dans la population générale (i.e. l'atypicité) comme dans l'approche dimensionnelle. Il s'agit ici de la probabilité avec laquelle un sujet sain peut être catégorisé comme un sujet pathologique, i.e. le risque de faire un faux positif, ce risque étant généralement intentionnellement maintenu à un niveau bas de manière à ce que les faux positifs soient rares. Insistons sur le fait que remonter le score seuil ne va pas permettre d'identifier des sujets « moins

pathologiques » (ceci est un raisonnement bon pour une approche dimensionnelle !) Cela va simplement augmenter le risque de faire des faux positifs, cf. figure 4E, et concomitamment, réduire le risque de faire des faux négatifs (si celui-ci n'était pas déjà nul).

5. Discussion

De nombreuses psychologues ont régulièrement besoin, dans leur pratique, d'objectiver les performances des sujets qu'elles rencontrent dans un domaine donné. Cette démarche passe généralement par l'utilisation de tests standardisés et par la comparaison des performances du sujet à une norme. Dans cet article, nous avons souhaité mettre en lumière que deux logiques assez distinctes pouvaient organiser et donner du sens à cette démarche, logiques qui ne sont pas toujours bien identifiées ou dont les conséquences pratiques sont sous-estimées. L'approche dimensionnelle semble plus prégnante dans le champ de la psychologie du développement où les passages d'un fonctionnement adapté à un fonctionnement inadapté (et vice-versa) peuvent se produire de manière souple au gré des dynamiques développementales (Perret & Faure, 2006). On la retrouve notamment dans les échelles de Wechsler, WAIS et WISC. L'approche catégorielle semble plus présente dans le champ de la neuropsychologie où les dysfonctionnements sont traditionnellement pensés comme le résultat d'atteintes cérébrales. On la retrouve par exemple dans les travaux du GREFEX (Roussel & Godefroy, 2008).

La question au cœur de cette discussion est la suivante : dans quelle mesure est-ce important de se positionner clairement dans le cadre de l'une de ces deux approches ? Après tout, dimensions ou catégories, différence de degrés ou de structure, il s'agit de positionnements très théoriques, presque philosophiques, et pour de nombreuses psychopathologies, nous n'aurons jamais aucune certitude. A cet égard, Pickles et Angold (2003) ont argué que de la même manière que la lumière pouvait être tantôt plutôt pensée comme une onde et tantôt plutôt pensée comme un flux de particules, les psychopathologies pouvaient, selon les circonstances et les objectifs, tantôt être pensées comme des catégories, et tantôt comme des dimensions. Sur le plan clinique par ailleurs, les conclusions que la psychologue tirera de son test sont assez proches : un sujet chez qui l'on mesure une performance inférieure au score seuil sera jugé préoccupant.

Nous défendons l'idée qu'il est important pour la psychologue de se positionner ou au moins de comprendre quelle position est tacitement véhiculée par les tests qu'elle utilise. Au-

delà même de l'identité professionnelle et des représentations que chacun se fait des psychopathologies, les implications pratiques sont importantes. L'interprétation des scores seuils est différente : frontière en pointillés pour guider l'interprétation dans l'approche dimensionnelle, garde-fou limitant le risque de faux-positif dans l'approche catégorielle. L'interprétation des scores observés des sujets est différente : positionnement de la performance sur un continuum dans une visée comparative dans un cas ; classification du sujet dans une catégorie et visée diagnostique dans l'autre.

Le fait que les échantillons normatifs ne soient pas de même nature change l'interprétation même de l'échelle des scores. Dans l'approche dimensionnelle, l'échantillon normatif étant composé de sujets tout-venants, 99,72% de la population générale aura un score *z* compris entre -3 et +3 (cf. figure 2). Dans ce cadre, mesurer chez un sujet un score $z = -6$ par exemple devra absolument éveiller les soupçons de la psychologue car la probabilité de tomber sur un individu avec un score si improbable est infime : approximativement une chance sur un milliard ! Il existe sur Terre environ 7 personnes avec un score si bas et vous seriez tombé dessus… Il sera plus pertinent de remettre en question la mesure elle-même, soit que le sujet n'aura pas compris les consignes, été dérangé ou que les normes utilisées n'étaient pas celles de son groupe de référence (e.g., une performance d'enfant comparée à des performances d'adultes). Dans l'approche dimensionnelle, l'échelle des scores *z* est une échelle avec un plancher (approximativement $z = -4$) et un plafond (approximativement $z = 4$) qu'il est théoriquement très peu probable de dépasser, y compris pour les sujets les plus déficitaires. Au contraire, dans l'approche catégorielle, l'échelle des scores *z* que manipule la psychologue réunit dans l'intervalle [-3 ; 3] 99,72% de la population saine. Et rien n'est dit sur les scores *z* des personnes pathologiques (cf. section 4, point de vigilance N°1) ! Dans cette perspective, mesurer chez un sujet un score $z = -6$ n'est pas aussi obligatoirement improbable que dans l'approche dimensionnelle (cf. Figure 4). Mais comme souligné plus haut, il n'est pas pertinent de donner un sens à cette valeur numérique, mesurée chez une personne de la catégorie « pathologie » alors que l'échelle de mesure a été calibrée en se basant uniquement sur la distribution des scores des personnes « saines ». Dans l'approche catégorielle, si la performance du sujet testé est inférieure au score seuil, que ce soit $z = -6$ ou $z = -2$, la conclusion du test est identique et se limite à : rejet de l'hypothèse nulle que la personne est saine, acceptation de l'hypothèse alternative que la personne est pathologique.

Les deux logiques sont différentes et appliquer un raisonnement dimensionnel à un test conçu dans une perspective catégorielle (i.e. avec un échantillon normatif constitué de sujets

sains) conduit en particulier à des erreurs d'interprétation. Par exemple, proposant des données normatives pour le Stroop Victoria collectées dans une population de personnes âgées saines et non-institutionnalisées, Bayard et al. (2011) indiquent dans l'appendice de leur article que les personnes testées dont le score *z* se situerait en deçà de -1,65 peuvent avoir leur performance qualifiée de « déficitaire », et « borderline » quand elle se situe entre *z* = -1,64 et *z* = -0,9. Deux aspects nous semblent problématiques ici. D'abord, dans une perspective catégorielle, la conclusion porte sur la catégorie d'appartenance des *personnes* et non pas sur leur *score*. Formellement, le raisonnement consiste à rejeter l'hypothèse que le sujet testé appartient à la population normative (ici, les personnes saines) pour accepter l'hypothèse alternative qu'il appartient à l'autre catégorie de *personnes* (les personnes pathologiques ou « non saines », a minima). Une interprétation correcte des performances serait : les *personnes* dont le score *z* se situerait en deçà de -1,65 peuvent être qualifiées de « déficitaires » (ou « pathologiques » ou, a minima, « non saines »), avec 5% de risque de commettre un faux positif en opérant cette catégorisation (des personnes). Ensuite, que dire des personnes dont le score se situerait entre *z* = -1,64 et *z* = -0,9 ? Leur score étant supérieur au seuil, et les auteurs ne présentant aucune information sur le risque de commettre un faux négatif en les catégorisant comme « saines », il n'est pas possible de conclure. En aucun cas les *personnes* ne peuvent être « borderline » dans un raisonnement catégoriel à deux catégories : soit on a la Covid-19, soit on ne l'a pas. Quant aux *scores*, au mieux, on peut acter qu'ils sont dans la moyenne basse des performances des sujets sains (Guilmette et al., 2020), ce qui ne remet pas en question le fait que ces personnes avec ces scores soient saines. Il nous semble que si l'intérêt est pour les scores et l'objectif, de situer la performance des sujets testés par rapport à la norme, un test comparatif dans une approche dimensionnelle serait plus pertinent.

Finalement, pour les concepteurs de tests, l'enjeu est important puisque leur choix d'un échantillon normatif (population générale vs. population saine) pour interpréter la performance isolée d'un sujet, va les engager théoriquement sur le construit qu'ils testent ! La constitution d'un échantillon normatif représentatif de la population saine et le recours à un test d'hypothèses pour rejeter avec un certain degré de confiance (selon le niveau de risque assumé de faire un faux positif) l'hypothèse que le sujet testé appartient à cette population saine, cette démarche produit inévitablement des catégories et donc accrédite une conception catégorielle des psychopathologies. A l'inverse, la constitution d'un échantillon normatif représentatif de la population générale, où tous les individus sont représentés sur un axe unique et où les performances individuelles se distinguent simplement par une plus ou moins

grande distance à une moyenne générale commune, cette démarche accrédite plutôt une conception dimensionnelle des psychopathologies.

6. Illustration pratique

Une psychologue reçoit un homme de 60 ans, avec un niveau d'étude CAP, qui a une plainte concernant sa mémoire. Elle souhaite compléter son anamnèse et ses observations du sujet avec des tests standardisés. Elle opte pour le subtest « Mémoire des chiffres » de la WAIS-IV (Wechsler, 2011) qui teste la mémoire à court terme et la mémoire de travail, et pour le RL/RI-16 qui teste la mémoire épisodique verbale en contrôlant les processus d'encodage et de récupération, et qui est reconnu pour son utilité dans le dépistage des démences comme la maladie d'Alzheimer (Amieva et al., 2007 ; Van der Linden et al., 2004).

La WAIS-IV est commercialisée par un éditeur professionnel, Pearson, ce qui offre, en contrepartie d'un prix d'achat non négligeable, un certain confort d'utilisation : le matériel, un manuel d'administration et un manuel d'interprétation incluant les normes et une documentation sur les qualités psychométriques des subtests, toute l'information nécessaire pour le testing est à portée de main. En consultant le manuel de la WAIS-IV, la psychologue constate que l'échantillon normatif (N = 876) est constitué de personnes tout-venantes, i.e. typiques et possiblement atypiques (quoique dans une proportion bien moindre évidemment) sur la dimension mesurée. Il est notamment précisé, p. 26 du manuel d'interprétation, que « l'échantillon de sujets âgés de 16 ans à 79 ans 11 mois [est] considéré comme représentatif de la population française ». Dans la perspective dimensionnelle historiquement associée aux échelles de Wechsler, l'objectif ne va pas être de tester l'hypothèse que le sujet appartient à la catégorie des personnes saines puisqu'il n'y a pas, stricto sensu, deux catégories de sujets. L'objectif va être d'estimer la performance « vraie » du sujet au subtest et d'évaluer dans quelle mesure celle-ci est atypique au regard des scores dans la population générale. Imaginons que le sujet ait un score observé $z = -1{,}58$. Sachant qu'à l'âge du sujet, le coefficient de fidélité du subtest est 0,92, la psychologue obtient l'intervalle de confiance à 95% suivant pour son score vrai : [-2,06 ; -1,10]. Elle interprète cet intervalle de confiance avec en tête, en guise de boussole, la connaissance du fait que seuls 5% des sujets ont un score $z < -1{,}65$ et seuls 2,5% des sujets un score $z < -1{,}96$. Le sujet affiche donc des performances entre « faibles » et « moyenne faibles ».

Le RL/RI-16 ne fait pas l'objet d'une commercialisation par un éditeur ce qui va confronter la psychologue à plusieurs difficultés. Une première difficulté est d'obtenir des normes de qualité (récentes, en langue française, collectées auprès d'un échantillon suffisamment important de personnes). A notre connaissance, trois publications ont proposé des normes françaises pour le RL/RI-16 (Amieva et al., 2007 ; Godefroy et al., 2016 ; Van der Linden et al., 2004). A la lecture de ces publications, il n'est pas aisé pour le lecteur peu expert de déterminer si les échantillons normatifs sont composés de sujets sains ou de sujets tout-venants. Van der Linden et al. (2004) ne spécifient aucun critère d'inclusion ou d'exclusion pour leurs 483 adultes. Ces derniers sont plus probablement des sujets sains, de manière cohérente avec l'ancrage en neuropsychologie des auteurs qui notent : « le centile 5 est classiquement considéré comme le "seuil pathologique" puisque la probabilité d'obtenir un tel score ou un score inférieur dans la population normale est égale ou inférieure à 5% » (p. 35, c'est nous qui soulignons). Amieva et al. (2007) se montrent un peu ambigus également en présentant leurs données (N = 1458) comme des normes « en population générale ». Il s'agit néanmoins bien de sujets sains comme en témoignent les critères d'exclusion rapportés qui visent à écarter les personnes potentiellement pathologiques (diagnostic de démence, vie en institution pour personnes âgées) et la citation suivante : « L'intérêt majeur de ce travail est d'avoir été réalisé sur un large échantillon de sujets sélectionnés en population générale susceptible de refléter la grande variabilité des performances cognitives de sujets âgés dits normaux » (p. 216, c'est nous qui soulignons). Godefroy et al. (2016) ont également recruté des sujets sains (N = 767), appliquant les critères d'inclusion et d'exclusion du GREFEX.

Une deuxième difficulté pour la psychologue est que le RL/RI-16 permet la collecte de nombreux scores - jusqu'à une douzaine - et que chacune des études susmentionnées ne propose pas des normes pour les mêmes scores (ou combinaison de scores). Ces différents scores ne se distribuant pas tous suivant une loi normale dans les échantillons normatifs, les auteurs ont la plupart du temps opté pour des scores seuils exprimés en centiles, mais pas toujours le même ! Alors que Van der Linden et al. (2004) et Godefroy et al., 2016) assument un risque de faux positifs de 5% (correspondant au centile 5), Amieva et al. (2007) ont opté pour un risque de 10% (correspondant au centile 10). Pour éviter dans le cadre de cette illustration d'entrer dans la complexité d'un test impliquant des scores multiples (cf. section 4, point de vigilance N°3), considérons que la psychologue ne s'intéresse qu'au score de rappel libre différé. Intéressée par ce score de rappel libre différé et un risque de faux positifs raisonnable de 5%, la psychologue opte pour les normes de Van der Linden et al. (2004),

quand bien même ce sont les plus anciennes et celles dont la taille d'échantillon est la plus petite (N = 13 seulement dans la catégorie qui l'intéresse des hommes de 60 ans avec moins de 12 années d'étude).

Au rappel libre différé, le sujet obtient un score brut de 7 mots rappelés (sur 16) ce qui, avec les normes proposées par Van der Linden et al. (2004), correspond à un score *z* = -2,18 (pour ce sexe, cet âge et ce niveau d'étude). Ce score observé étant inférieur au score seuil (*z* = -1,65, ou centile 5), la psychologue va pouvoir, au seuil de 5% (i.e. avec 5 chance sur 100 de faire un faux positif), rejeter l'hypothèse nulle que le sujet est sain et conclure qu'il est dans la catégorie des personnes « non saines ». Est-ce à dire que ce sujet a la maladie d'Alzheimer ? Non, car bien que chacune des trois études citées plus haut souligne l'intérêt du RL/RI-16 pour identifier les personnes démentes, aucune d'entre elles n'envisage directement le RL/RI-16 comme un outil de dépistage de la maladie d'Alzheimer. Amieva et al. (2007) rappellent même « que les scores normatifs pour le test RL/RI-16 ne sont qu'une aide au diagnostic d'une démence de type Alzheimer et que de faibles performances ne peuvent permettre à elles seules de conclure à l'existence d'une démence » (p. 219). Sur le principe pourtant, un test qui aurait une spécificité de 100% et une sensibilité de 100% serait un parfait outil diagnostique ! C'est une équipe québécoise (Drolet et al., 2014) qui a tenté de déterminer la spécificité et la sensibilité diagnostique du RL/RI-16 pour détecter la maladie d'Alzheimer. Avec un échantillon malheureusement très réduit (20 sujets sains, 20 sujets Alzheimer), ils ont établi qu'utiliser la performance brute de 5 mots rappelés comme un score seuil au rappel libre différé permettait de classer les participants avec une spécificité de 100 % (100% des personnes rappelant 5 mots ou moins étaient des sujets Alzheimer) et une sensibilité de 90 % (10% des sujets Alzheimer [N = 2] ont eu un score supérieur à 5, et ont donc été catégorisés à tort comme des sujets sains). En optant pour ce score seuil de 5 mots rappelés (sur 16), score établi avec une courbe ROC de manière à minimiser le nombre d'erreurs de catégorisation, la conclusion de notre psychologue est différente : avec 7 mots rappelés sur 16, son sujet sera plus justement catégorisé comme un sujet sain (avec 0% de risque de faire un faux positif mais 10% de risque de faire un faux négatif). Ces conclusions contradictoires sur la catégorie d'appartenance du sujet rappellent l'importance d'avoir des échantillons normatifs de très grande taille et qui caractérisent non seulement les performances des sujets sains mais également les performances des sujets dans différents groupes cliniques « pathologiques ». Sinon, le test ne sera pas en mesure de détecter des pathologies particulières, ce qui constitue pourtant l'intérêt principal de l'approche catégorielle.

7. Conclusions et perspectives

Les deux approches pour interpréter une performance observée chez un sujet après un test présentées ici ont des logiques très différentes. Nous avons essayé de récapituler les principaux points distinguant les deux approches dans le tableau 1 et dans la figure 5.

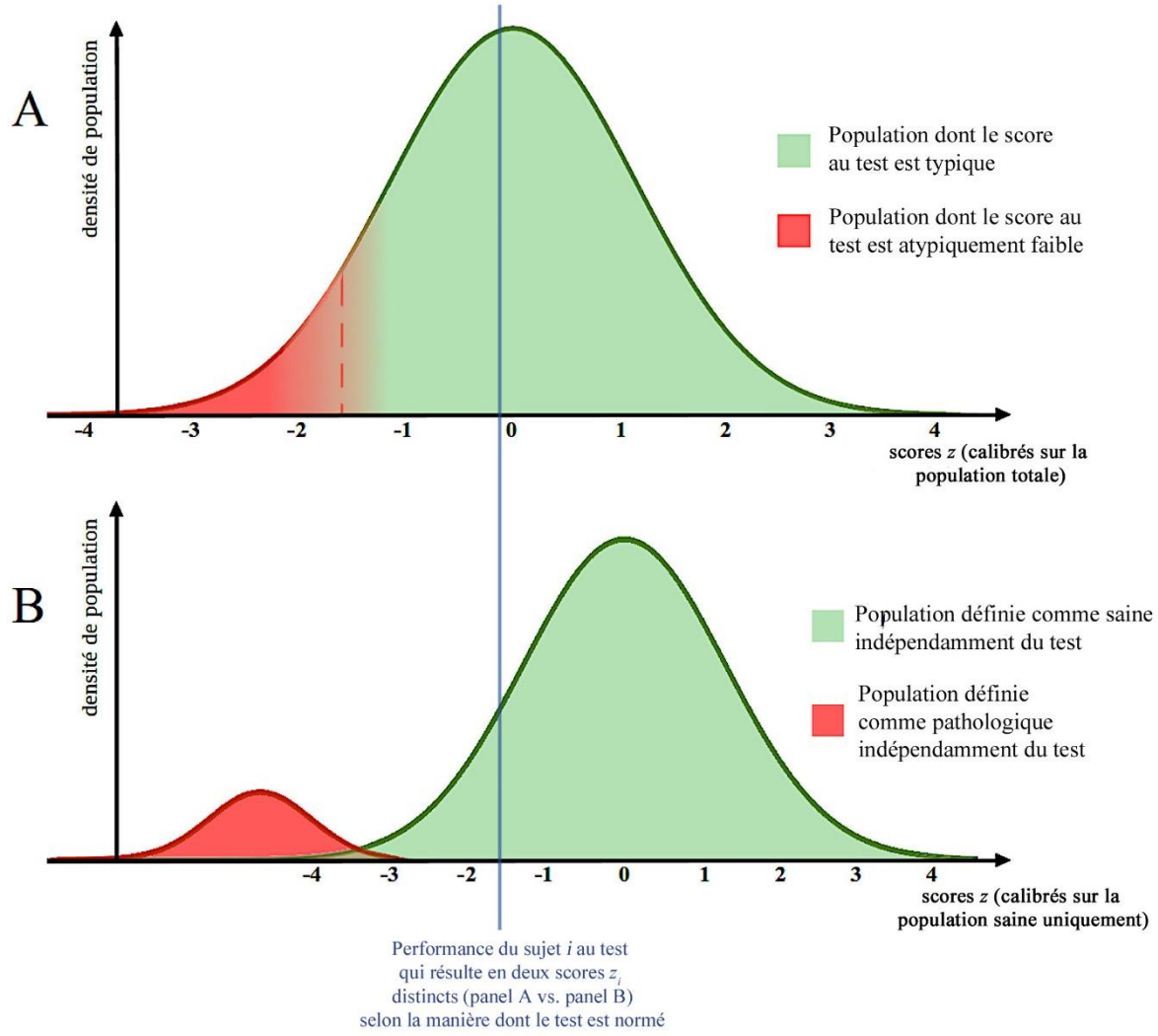

Figure 5. Distribution des scores *z* selon les deux conceptions des psychopathologies, dimensionnelle (panel A) ou catégorielle (panel B) à un même test pour une même population de personnes. Dans le panel A, les scores *z* sont calibrés sur la population totale, 99,97% de la population totale a un score entre -4 et 4, la moyenne générale est 0, l'écart-type est 1. Dans le panel B, les scores *z* sont calibrés uniquement sur le groupe de sujets identifiés comme « sains » en amont du test (en vert). Si par définition la moyenne des scores *z* est toujours 0, cela correspond concrètement à une performance brute plus élevée que la moyenne du panel A puisque le groupe des sujets sains exclut les personnes avec la pathologie (en rouge). Pour la même raison, l'écart-type des scores *z* du

panel B correspond à une variation plus faible que celle qu'on trouve dans la population totale (panel A). Dans le panel A, la proportion de personnes dont le score est atypique dépend directement du score seuil posé (trait pointillé rouge, ici z = -1,65 dans une perspective unilatérale, soit 5%). Au contraire dans le panel B, la proportion de personnes pathologiques dans la population totale est indépendante du seuil choisi (le seuil aura par contre une influence sur la manière dont le test va correctement identifier ces personnes « pathologiques », i.e. sans erreur de catégorisation, cf. Figure 4). Enfin, le trait vertical bleu représente la performance brute d'un sujet lambda au test.

Dans l'approche dimensionnelle, le recours à un score seuil est accessoire : il peut aider la psychologue à prendre une décision quant à la suite à donner à la prise en soin mais le plus souvent, il constitue une simplification binaire de l'information quantitative fournie par le test. Dans l'approche catégorielle, le recours à un score seuil est un passage obligé dans la mesure où la psychologue prend une décision sur la catégorie dans laquelle se trouve le sujet. Ce faisant, elle doit avoir un contrôle minimum sur le risque de faire une erreur de catégorisation. Si les seuils n'ont pas la même fonction, les échantillons normatifs ne sont pas non plus de même nature : composés de personnes tout-venantes dans l'approche dimensionnelle et de personnes « saines » dans l'approche catégorielle. Cette différence est cruciale : imaginons qu'on dispose à la fois, pour un unique test, de ces deux types de normes. Un sujet passe le test, obtient un score brut que la psychologue convertit en deux scores z, l'un avec les normes des personnes tout-venantes et l'autre avec les normes des personnes « saines ». Elle obtiendra deux scores z différents, le premier probablement supérieur au second (cf. figure 5). Donc l'interprétation du seuil ne peut pas et ne doit pas être la même. Ce qui réunit ces deux approches, c'est que dans l'approche dimensionnelle, le consensus sur ce qu'est un score « rare » s'est formé, au moins en France, autour de la proportion 5% (Colombo et al., 2016). Et dans l'approche catégorielle, le niveau de risque de faire un faux-positif consensuellement jugé tolérable en sciences humaines est également de 5%. Cette similitude permet d'utiliser le même score seuil la plupart du temps (z = -1,65) sans se poser la question du rôle joué par ce score seuil. Pourtant, comme on l'a vu, l'interprétation du seuil s'avère très différente. D'autre part, ces deux proportions peuvent être amenées à évoluer, possiblement dans des directions opposées, et il sera alors d'autant plus important de ne pas confondre les logiques sous-jacentes, comme en témoignent les difficultés rencontrées ces dernières années dans le domaine d'étude des troubles cognitifs dits « de basse intensité » (MCI en anglais, voir Godefroy et al., 2014).

Tableau 1. Récapitulatif des principaux points de comparaison entre l'approche dimensionnelle et l'approche catégorielle.

Points de comparaison	Approche dimensionnelle	Approche catégorielle
Conception théorique de la pathologie	- La pathologie se définit comme un extrême du spectre des possibles sur une dimension donnée - Différence de degré, quantitative	- La pathologie se définit comme un mode de fonctionnement différent du fonctionnement normal - Différence de structure, qualitative
Objectif du test	- Positionner la performance du sujet le long d'un continuum allant du typique à l'atypique - Visée comparative	- Placer le sujet dans une catégorie : les personnes « saines » (acceptation de H0) vs. les personnes « pathologiques » (acceptation de H1) - Visée diagnostique
La sensibilité du test	- Sa capacité à discriminer des sujets dont la performance est proche sur l'échelle de mesure (sensibilité psychométrique)	- Sa capacité à limiter les faux négatifs, i.e. à détecter seulement les personnes avec la pathologie et à rejeter les personnes « saines » (sensibilité diagnostique)
Interprétation du score *z* du sujet	- Traduit la performance du sujet en la situant sur le continuum des performances dans la population générale - Interprétable tant qu'il est dans l'étendue des scores probables (entre -4 et 4)	- Interprétation binaire en termes de rejet ou non de l'hypothèse selon laquelle le sujet est « sain » (H0) - Ne permet pas une interprétation en termes de plus ou moins grande sévérité de la pathologie
Le score seuil	- Informe sur le niveau de rareté de la performance du sujet - Contribue à définir la pathologie dans la nosologie - Changer le seuil permet d'identifier des personnes plus ou moins sévèrement déficitaires	- Détermine les risques de faire des erreurs de catégorisation (faux positifs et faux négatifs) - Permet de prendre une décision sur la catégorie d'appartenance du sujet (catégories définies indépendamment du test) - Changer le seuil modifie les risques de faire des erreurs de catégorisation (faux positifs et faux négatifs)
Population constituant l'échantillon normatif	- Tout venante (avec quelques critères d'exclusion possibles pour garantir que la mesure de la performance est valide)	- Saine (avec des critères d'exclusion permettant de garantir qu'aucune personne avec la pathologie n'a été incluse)
Points de vigilance	- Le seuil doit garder une valeur purement descriptive et ne pas servir à fonder des catégories - Toutes les performances rares ne sont pas obligatoirement préoccupantes du point de vue psychologique - Raisonner quand c'est possible, non pas avec le score observé du sujet, mais avec l'intervalle de confiance de son score « vrai »	- Le test ne dit rien des caractéristiques des sujets pathologiques, ni comme groupe, ni au niveau individuel (cf. ci-dessus, ne permet pas une interprétation en termes de « degré de pathologie ») - Le risque de faux-négatif est impossible à contrôler si les performances de la population pathologique ne sont pas connues - La multiplication des tests augmente le risque de produire des erreurs de catégorisation (faux positifs et faux négatifs)

Déclaration de liens d'intérêts

Les auteurs déclarent n'avoir aucun lien d'intérêt en relation avec cet article

Remerciements

Les auteurs souhaitent remercier Céline Becquet et Renaud Coppalle pour avoir relu et commenté une version initiale de cet article.

Bibliographie

Adam, D. (2013). Mental health: On the spectrum. *Nature News*, *496*(7446), 416.

Aguert, M., & Capel, A. (2018). Mieux comprendre les scores z pour bien les utiliser. *Rééducation orthophonique*, *274*, 61-85.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing. Washington*, DC: American Psychological Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders DSM-5* (5e éd.). American Psychiatric Publishing.

Amieva, H., Carcaillon, L., L'Alzit-Schuermans, P. R., Millet, X., Dartigues, J. F., & Fabrigoule, C. (2007). Test de rappel libre/rappel indicé à 16 items : Normes en population générale chez des sujets âgés issues de l'étude des 3 Cités. *Revue Neurologique*, *163*(2), 205–221.

Bayard, S., Erkes, J., & Moroni, C. (2011). Victoria Stroop Test: normative data in a sample group of older people and the study of their clinical applications in the assessment of inhibition in Alzheimer's disease. *Archives of Clinical Neuropsychology*, *26*(7), 653-661.

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*(1), 31-46.

Bishop, D. V. (2017). Why is it so hard to reach agreement on terminology? The case of developmental language disorder (DLD). *International Journal of Language & Communication Disorders*, *52*(6), 671–680.

Bolton, P., Macdonald, H., Pickles, A., Rios, P. al, Goode, S., Crowson, M., Bailey, A., & Rutter, M. (1994). A case-control family history study of autism. *Journal of Child Psychology and Psychiatry*, *35*(5), 877–900.

Bornstein, R. F., & Natoli, A. P. (2019). Clinical utility of categorical and dimensional perspectives on personality pathology : A meta-analytic review. *Personality Disorders: Theory, Research, and Treatment*, *10*(6), 479-490.

Cizek, G. J. (2012). *Setting performance standards: Foundations, methods, and innovations*. (2d éd). Routledge.

Coghill, D., & Sonuga-Barke, E. J. S. (2012). Categories versus dimensions in the classification and conceptualisation of child and adolescent mental disorders—Implications of recent empirical study. *Journal of Child Psychology and Psychiatry*, *53*(5), 469-489.

Colombo, F., Amieva, H., Lecerf, T., & Verdon, V. (2016). La norme en neuropsychologie, un concept à facettes multiples. *Revue de neuropsychologie, Volume 8*(1), 61-69.

Constantino, J. N., & Todd, R. D. (2003). Autistic traits in the general population: A twin study. *Archives of General Psychiatry*, *60*(5), 524–530.

Crocq, M.-A., Guelfi, J.-D., Boyer, P., Pull, C.-B., & Pull, M.-C. (2015). *DSM-5—Manuel diagnostique et statistique des troubles mentaux* (5e édition). Elsevier Masson.

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.

Delacour, H., Servonnet, A., Perrot, A., Vigezzi, J. F., & Ramirez, J. M. (2005). La courbe ROC (receiver operating characteristic) : Principes et principales applications en biologie clinique. *Annales de biologie clinique*, *63*(2), 145-154.

Drolet, V., Vallet, G. T., Imbeault, H., Lecomte, S., Limoges, F., Joubert, S., & Rouleau, I. (2014). Comparaison des performances à l'épreuve des 15 mots de Rey et au RL/RI 16 dans le vieillissement normal et la démence de type Alzheimer. *Gériatrie et Psychologie Neuropsychiatrie Du Vieillissement*, *12*(2), 218–226.

Esler, A. N., & Ruble, L. A. (2015). DSM-5 diagnostic criteria for autism spectrum disorder with implications for school psychologists. *International Journal of School & Educational Psychology*, *3*(1), 1–15. https://doi.org/10.1080/21683603.2014.890148

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE Publications Ltd.

Goddard, H. H. (1920). *Human efficiency and levels of intelligence*. Princeton University Press.

Godefroy, O., Diouf, M., Bigand, C., & Roussel, M. (2014). Troubles neurocognitifs d'intensité légère ou performances normales basses ? *Revue de Neuropsychologie*, *6*(3), 159-162.

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

Guilmette, T. J., Sweet, J. J., Hebben, N., Koltai, D., Mahone, E. M., Spiegler, B. J., Stucky, K., & Westerveld, M. (2020). American Academy of Clinical Neuropsychology consensus conference statement on uniform labeling of performance test scores. *The Clinical Neuropsychologist*, *34*(3), 437-453. https://doi.org/10.1080/13854046.2020.1722244

Hengartner, M. P., & Lehmann, S. N. (2017). Why psychiatric research must abandon traditional diagnostic classification and adopt a fully dimensional scope: Two solutions to a persistent problem. *Frontiers in Psychiatry*, *8*, 101

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.

Kamp-Becker, I., Smidt, J., Ghahreman, M., Heinzel-Gutenbrunner, M., Becker, K., & Remschmidt, H. (2010). Categorical and dimensional structure of autism spectrum disorders: The nosologic validity of Asperger Syndrome. *Journal of Autism and Developmental Disorders*, *40*(8), 921–929. https://doi.org/10.1007/s10803-010-0939-5

Krueger, R. F., Markon, K. E., Patrick, C. J., & Iacomo, W. G. (2005). Toward a dimensionally based taxonomy of psychopathology: Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for DSM-V. *Journal of abnormal psychology*, *114*(4), 537-550.

Kupfer, D. J., & Regier, D. A. (2011). Neuroscience, clinical evidence, and the future of psychiatric classification in DSM-5. *American Journal of Psychiatry*, *168*(7), 672-674.

Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological bulletin*, *137*(5), 856-879.

Mottron, L. (2021). Les « traits autistiques » ne sont pas autistiques. *Enfance*, *3*, 293–311.

Mueller, L., Munson, L. (2015). Setting cut scores. In: Hanvey, C., Sady, K. (Éds.) *Practitioner's guide to legal issues in organizations*. Springer. https://doi.org/10.1007/978-3-319-11143-8_6

Perret, P., & Faure, S. (2006). Les fondements de la psychopathologie développementale. *Enfance*, *58*(4), 317-333.

Pickles, A., & Angold, A. (2003). Natural categories or fundamental dimensions: On carving nature at the joints and the rearticulation of psychopathology. *Development and psychopathology*, *15*(3), 529-551.

Piérart, B. (2004). Introduction : Les dysphasies chez l'enfant : un développement en délai ou une construction langagière différente? *Enfance*, *561*(1), 5–19.

Posserud, M.-B., Lundervold, A. J., & Gillberg, C. (2006). Autistic features in a total population of 7–9-year-old children assessed by the ASSQ (Autism Spectrum Screening Questionnaire). *Journal of Child Psychology and Psychiatry*, *47*(2), 167–175.

Raoux, N., Le Carret, N., Meillon, C., Blanchard, C., Bergua, V., Dartigues, J.-F., & Amieva, H. (2014). Validation d'un test court de génération de concepts à partir des données issues de la cohorte 3C de sujets âgés en population générale. *Revue de neuropsychologie*, *6*(2), 129-137.

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

Roussel, M., & Godefroy, O. (2016). Faut-il considérer un bilan neuropsychologique comme anormal ? Importance de l'intégration des performances multiples. In C. Belin, D. Maillet, & H. Amieva (Éds.), *L'évaluation neuropsychologique : De la norme à l'exception* (p. 11-18). De Boeck Université.

Rubenstein, E., & Chawla, D. (2018). Broader autism phenotype in parents of children with autism: A systematic review of percentage estimates. *Journal of Child and Family Studies*, *27*(6), 1705–1720.

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & Baron-Cohen, S. (2015). Measuring autistic traits in the general population: A systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. *Molecular Autism*, *6*(1), 1–12.

Sapolsky, R. M. (2017). *Behave: The biology of humans at our best and worst*. Penguin.

Septier, M., Peyre, H., Amsellem, F., Beggiato, A., Maruani, A., Poumeyreau, M., Amestoy, A., Scheid, I., Gaman, A., & Bolognani, F. (2019). Increased risk of ADHD in families with ASD. *European Child & Adolescent Psychiatry*, *28*(2), 281–288.

Sonuga-Barke, E. J. (1998). Categorical models of childhood disorder: A conceptual and empirical analysis. *The journal of child psychology and psychiatry and allied disciplines*, *39*(1), 115–133.

Stevens, T., Peng, L., & Barnard-Brak, L. (2016). The comorbidity of ADHD in children diagnosed with autism spectrum disorder. *Research in Autism Spectrum Disorders*, *31*, 11–18.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3e édition). Oxford University Press.

Van der Linden, M., Coyette, F., Poitrenaud, J., Kalafat, M., Calicis, F., Wyns, C., & Adam, S. (2004). L'épreuve de rappel libre / rappel indice à 16 items (RL/RI-16). In M. Van der Linden & GREMEN (Éds.), *L'évaluation des troubles de la mémoire : Présentation de quatre tests de mémoire épisodique (avec leur étalonnage)* (p. 25-47). Solal.

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

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