Mieux comprendre les scores z pour bien les utiliser

Marc Aguert, Aurélie Capel

PRE PRINT

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

Mieux comprendre les scores *z* pour bien les utiliser

Marc Aguert[a] & Aurélie Capel[b]

[a] Normandie Univ, UNICAEN, LPCN, Caen, France
[b] Service de recherche clinique, Centre François Baclesse, Caen

Correspondance : Marc Aguert, Esplanade de la Paix, CS 14032, 14032 Caen cedex 5, France.
Courriel : marc.aguert@unicaen.fr

Résumé : L'évaluation en orthophonie ou en psychologie passe fréquemment par l'utilisation des scores *z*, un outil statistique permettant d'exprimer de manière standardisée la performance d'une personne à un test. L'utilisation des scores *z* est généralement couplé avec des scores-seuils qui, croisés avec des considérations cliniques, permettent de conclure si la performance de l'individu est « normale » ou « pathologique ». Le présent article vise d'abord à rappeler la manière dont fonctionnent les scores *z*, notamment leurs liens étroits avec la distribution des scores dans la population de référence, normale (i.e. gaussienne) ou non. Il vise ensuite à aider le praticien à réduire ses risques de faire des faux-positifs ou des faux-négatifs en concluant qu'une performance est pathologique en attirant son attention sur les facteurs suivants : utilisation d'un test inadapté au sujet, d'un score seuil inadapté, de normes de comparaison inadaptées au sujet, non normalité des scores dans la population de référence, et enfin, non prise en compte de l'erreur de mesure.

Abstract: Assessment in speech therapy or psychology frequently involves the use of *z*-scores, a statistical tool that allows to standardize the person's performance on a test. The use of *z*-scores is generally coupled with cut-off scores that, when crossed with clinical considerations, allow to conclude whether the individual's performance is "normal" or "abnormal". This article first aims to recall the way *z*-scores work, in particular their close links with the distribution of scores in the reference population, normal (i.e. Gaussian) or not. It then aims to help the practitioner reduce the risk of making false positives or false negatives by concluding that a performance is abnormal by drawing attention to the following factors: use of a test unsuited to the subject, an unsuitable cut-off score, norms unsuited to the subject, scores following a non-normal distribution in the population, and finally, not taking measurement error into account.

Mieux comprendre les scores *z* pour bien les utiliser

L'utilisation de tests pour évaluer de manière objective la performance de patients sur une dimension donnée est une pratique très courante, en orthophonie ou en psychologie. Ces tests utilisent des métriques variées et dans le but de comparer les résultats à deux tests distincts ou dans celui de communiquer avec le patient ou entre professionnels, les praticiens convertissent souvent les scores bruts en scores *z*, une échelle de scores standardisée connue de tous. Au-delà de leur aspect standardisé, les scores *z* intéressent les praticiens parce qu'ils leur permettent de distinguer, en comparant le score du patient à un score seuil (« cut-off score »), les performances « normales » des performances « pathologiques ». Cette opération, si elle est évocatrice sur le plan clinique, expose le praticien au risque de commettre des erreurs : des faux-positifs s'il ou elle classe des performances normales dans la catégorie des performances pathologiques ; des faux-négatifs s'il commet l'erreur inverse. Si l'utilisation des scores *z* est largement répandue, la possibilité de commettre ces erreurs et les facteurs aggravant le risque de commettre ces erreurs semblent moins présents dans l'esprit des praticiens. L'objectif de cet article est de rappeler de façon accessible la logique statistique à la base du fonctionnement des scores *z*, les limites imposées par cette logique statistique et les facteurs à considérer pour éviter de produire des conclusions erronées quand un score individuel est comparé à un score seuil pour identifier une performance « pathologique ». Nous nous focalisons sur la situation la plus simple : quand un unique score *z* est calculé dans une visée évaluative. Les situations où au-delà de l'évaluation d'une dimension donnée, le praticien a une visée diagnostique et les situations où il utilise plusieurs scores *z* en combinaison posent d'autres contraintes et limites en plus de celles évoquées ici et qui dépassent le cadre du présent article.

Introduction : la normalité statistique

L'un des premiers problèmes rencontré par les orthophonistes et les psychologues avant d'accompagner un patient est d'objectiver l'existence d'un trouble. Une rééducation ou une remédiation se justifie parce que le patient a un fonctionnement pathologique ou au moins, infranormal. Or, distinguer le normal du pathologique est un exercice notoirement difficile. Sous l'influence notamment des travaux de Canguilhem (1966/2013), on distingue généralement la normalité sociale, la normalité fonctionnelle et la normalité statistique. La normalité sociale (ou idéale) renvoie à la norme sociale majoritaire, à ce qui est considéré « normal » et à ce qui est considéré « déviant » dans une société donnée. Cette normalité dépend donc des évolutions du corps social. Un exemple est la manière dont on a considéré l'homosexualité dans les sociétés occidentales au cours du XXe siècle, d'un comportement déviant (et donc pathologique) à une simple question de préférence sexuelle. La normalité fonctionnelle se définit par rapport à l'individu lui-même : la personne est dans son état « normal » quand elle est en pleine possession de tous ses moyens, ceux déterminés par ses caractéristiques physiologiques et psychologiques propres. C'est donc une forme de normalité pertinente pour la pratique orthophonique ou neuropsychologique où une partie des personnes consulte en raison d'une diminution de leur efficience neurocognitive, suite à un accident, un traitement, etc. Malheureusement, cette normalité est difficilement quantifiable puisqu'on

ignore toujours quel était l'état « normal » du sujet avant son accident, son traitement, etc. C'est pour évaluer cette normalité fonctionnelle que les praticiens procèdent toujours à une anamnèse méticuleuse. Finalement, la normalité statistique est définie à travers la fréquence d'apparition des comportements : un comportement ou une performance fréquente est jugée « normale » tandis qu'un comportement ou une performance rare est jugée « anormale », et donc possiblement pathologique. A défaut d'être toujours pertinente dans l'évaluation des fonctions cognitives, la normalité statistique est largement utilisée en orthophonie et en psychologie, les praticiens étant souvent séduits par les gages d'objectivité qu'elle offre.

Les statistiques ne permettent pas de dire qu'un comportement en soi, ou une performance, est faible ou fort, normal ou pathologique. Les statistiques se bornent à renseigner, après collecte d'un nombre important de comportements / performances sur la population d'intérêt, la fréquence avec laquelle ces comportements / performances apparaissent. Ce qui permet ensuite de juger de leur fréquence ou de leur rareté. Imaginons qu'on s'intéresse aux résultats nationaux du baccalauréat en 2017 en France. La figure 1 représente la **distribution de fréquence** des performances des lycéens : en abscisse, la performance (exprimée par la note obtenue) et en ordonnée, la fréquence[1] de lycéens entrant dans chacune des catégories définies en abscisse. Un simple coup d'œil à cette distribution de fréquence permet de constater que certaines performances (par exemple, avoir 10 au bac, qui concerne 25,4% des lycéens) sont plus fréquentes que d'autres (par exemple, avoir 19 au bac qui concerne 0,1% des lycéens). Ce premier exemple met tout de suite en évidence la faiblesse principale de l'approche statistique de la normalité : avoir son bac avec une moyenne de 19/20 est très rare. Pour autant, personne n'aurait l'idée de qualifier cette performance de pathologique…

La distribution de fréquence permet de juger une note de manière relative, c'est-à-dire au regard de sa position dans la distribution. Imaginons que Paul ait réalisé 12 erreurs à un test. Cette performance, qui n'a pas de sens en soi, devient interprétable quand on constate sur la distribution de fréquence qu'une grande majorité des individus de la population à laquelle Paul appartient fait plus de 12 erreurs. Faire 12 erreurs est alors plutôt une bonne performance. A l'inverse, si la distribution de fréquence montre qu'une grande majorité des individus de la population fait moins de 12 erreurs, alors Paul a plutôt réalisé une mauvaise performance. En regardant la figure 1, on constate que 9/20 n'est pas une bonne note car une grande majorité de lycéens (88%) ont plus que cette note. Il existe deux techniques pour situer précisément un individu par rapport à sa population de référence : les **quantiles** et les **scores standards**. Ces deux techniques sont basées sur les informations apportées par la distribution de fréquence et ont chacune des avantages et des inconvénients sur lesquels nous reviendrons au fil de cet article.

[1] Les fréquences sont une manière d'exprimer un effectif sous forme de proportion. Fréquences et pourcentages sont des proportions qui se distinguent simplement par le fait que l'échelle des fréquences va de 0 à 1 et l'échelle des pourcentages va de 0 à 100.

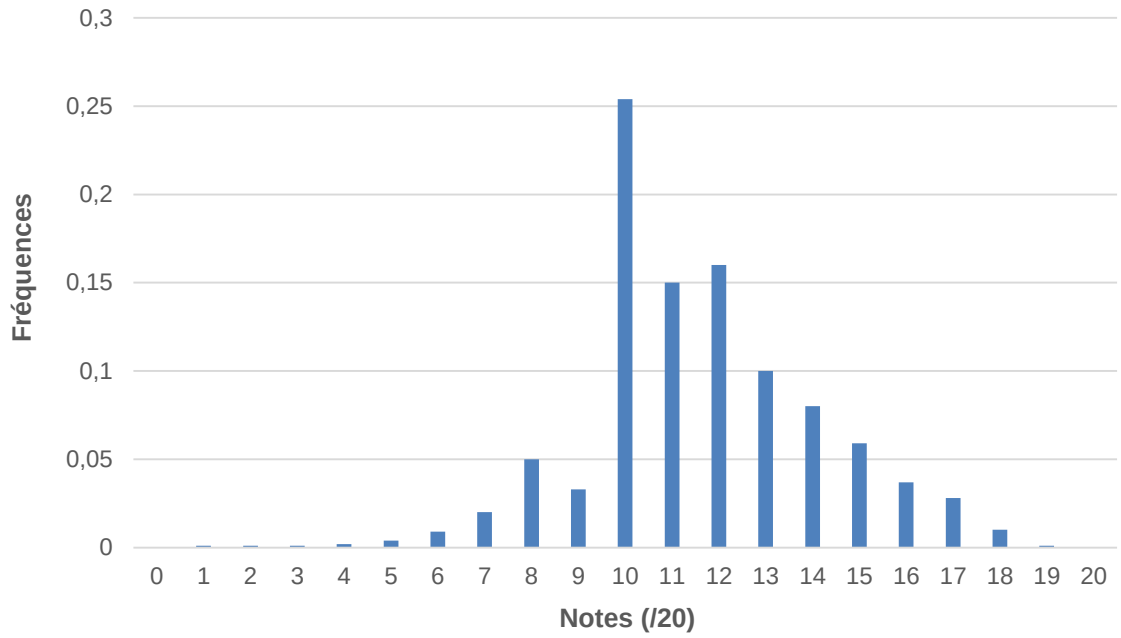

Figure 1. Résultats nationaux au bac 2017 (données fictives)

Les scores *z*

Une distribution centrée réduite

Les scores *z* sont un type de scores standards, c'est-à-dire de scores dont la moyenne et l'écart-type de la distribution sont conventionnels et connus de ceux qui les utilisent, ce qui facilite l'interprétation de ces scores. Parmi les scores standards les plus connus on trouve donc les scores *z* (μ = 0 ; σ = 1), les scores T (μ = 50 ; σ = 10), les QI (μ = 100 ; σ = 15)[2]. Il est possible de transformer n'importe quelle distribution de notes en scores *z* en appliquant la formule (1) ci-dessous où *x* est la note brute d'une personne, m_x la moyenne de la distribution des notes et s_x, l'écart-type de la distribution des notes[3].

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

Cette simple opération arithmétique aura pour effet de *centrer* la moyenne de la distribution sur 0 et de *réduire* l'écart-type sur 1, quelle que soit la forme de la distribution originelle des notes. On obtient ainsi une distribution des scores dite « centrée réduite » où chaque score *z* exprime directement la « distance » du patient à la moyenne en nombre d'écart-type. Un patient ayant

[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* ». Ainsi, la distribution des scores *z* dans la population a une moyenne μ et un écart-type σ.

[3] Ici, nous utilisons les notations « *m* » et « *s* » puisqu'en pratique, nous connaissons rarement la moyenne et l'écart-type de notre population d'intérêt. A défaut, nous utilisons la moyenne et l'écart-type de l'échantillon ayant servi à étalonner / normer le test. Les notations « SD » (pour l'anglais Standard Deviation) et « ET » (pour Ecart-Type) sont à éviter (« ET » étant employé pour désigner l'erreur-type), de même que l'anglicisme « deviation standard » pour désigner les écart-types.

un score *z* = +2 a une performance se situant au-dessus de la moyenne et plus précisément, une performance deux fois supérieure à la variation moyenne autour de la moyenne dans sa population de référence. Cependant, il est très important de noter que connaître la distance d'un individu à sa moyenne en nombre d'écart-types (que cette information soit « brute » ou standardisée) n'apporte aucune information sur la rareté ou la fréquence de son score si on ignore quelle est la forme de la distribution de fréquence des scores dans la population de référence. Si on regarde la distribution des scores *z* de la figure 2, on constate qu'avoir un score à -1,8 écart-types de la moyenne (*z* = -1,8) est plus fréquent qu'avoir un score *z* = -1,2 ou *z* = 1,5 écart-type de la moyenne !

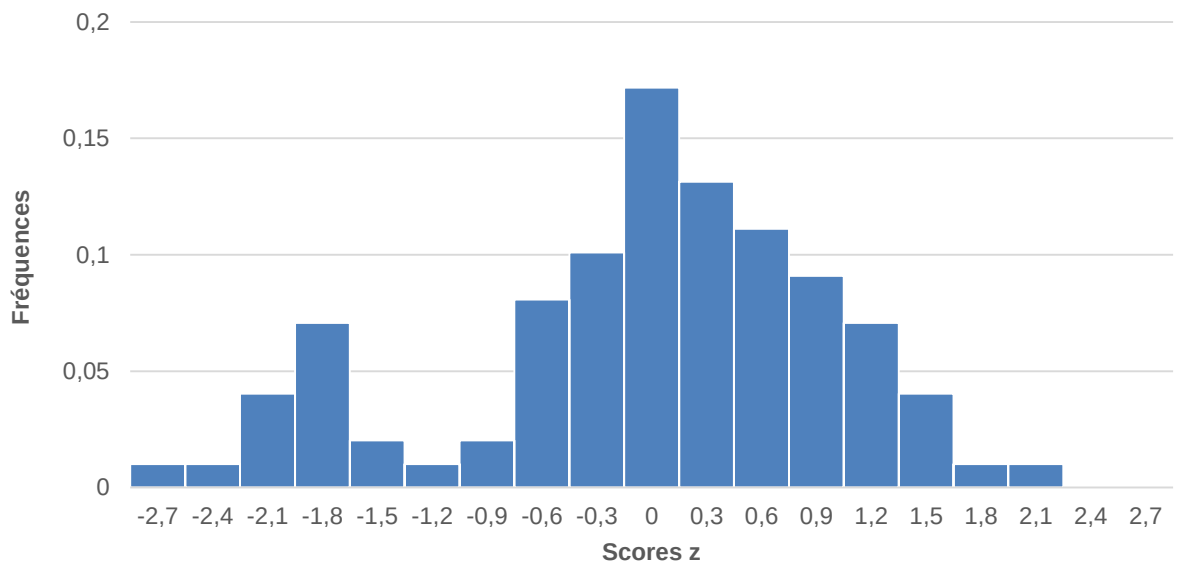

Figure 2. Distribution de fréquence des scores *z* d'un test fictif (*m* = 0 ; *s* = 1 ; N = 99)

C'est là l'inconvénient majeur de la technique des scores standards : l'interprétation correcte d'un score suppose non seulement de connaître la moyenne et l'écart-type de la distribution des scores mais aussi la forme de cette distribution, étant entendu que la conversion en score *z* ne change pas la forme de la distribution originale. En pratique, l'utilisation des scores standards et donc des scores *z* se limitent aux cas où la distribution de fréquence des scores au test dans la population suit une **loi normale**.

Lorsque la distribution des notes brutes suit une loi normale, les scores *z* peuvent être mis en lien avec des probabilités de se situer à tel ou tel endroit de la distribution de fréquence grâce à la table de probabilité de la loi normale. Cette table, que l'on retrouve à la fin de tous les bons manuels d'introduction aux statistiques en sciences humaines indique la probabilité pour un sujet d'avoir un score supérieur ou inférieur à une valeur de *z* donnée. Par exemple, la table indique que la probabilité d'avoir un score supérieur à *z* = 0 est 0,5 soit 50% de chance. La probabilité d'avoir un score inférieur à *z* = -3 est 0,0013 soit 0,13% de chance. La figure 3 montre une distribution normale prototypique et la probabilité de se situer à différents endroits

de la distribution. On y voit que 95,44% de la population doivent avoir un score *z* compris entre -2 et +2.

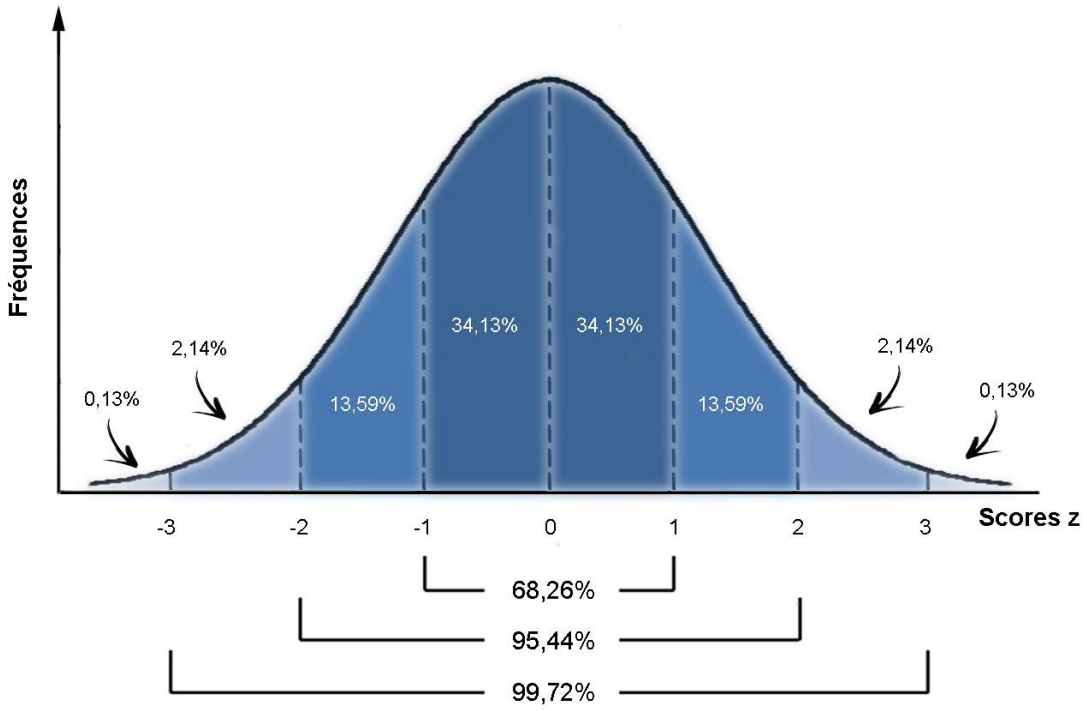

Figure 3. Distribution normale et probabilité de se trouver à différents endroits de la distribution

Tableau 1. Quelques scores *z* et probabilité associée d'avoir un score inférieur à ces scores *z*

Scores *z*	-4	-3	-2	-1,96	-1,65	-1,28	-1	-0,5	0
Pourcentage de la population ayant un score *z* inférieur	0,003%	0,13%	2,3%	2,5%	5%	10%	15%	30%	50%

Le tableau 1 synthétise les probabilités (exprimées en pourcentages) associées à quelques scores *z* intéressants ou communément utilisés en clinique.

Si les notes de notre test se distribuent en suivant une loi normale, il est possible d'utiliser les scores *z* pour identifier quels scores sont pathologiques à la condition d'avoir au préalable posé un score seuil (« cut-off score ») en deçà duquel, la performance est jugée si rare / faible qu'elle en est pathologique. A noter que la forme de la distribution normale suppose que les scores les plus rares sont forcément les scores les plus éloignés de la moyenne. Il devient possible d'assimiler la rareté d'un score à la faiblesse d'un score (les scores rares élevés, à droite de la distribution, étant beaucoup plus rarement considérés comme posant problème, cf. plus loin la section sur les scores seuils). Posons pour l'exemple que des notes sont rares / faibles si elles sont obtenues par moins de 5% des individus. Un coup d'œil au tableau 1 (ou à la table de probabilité de la loi normale) nous permet de constater que ces notes faibles, exprimées en score *z*, sont les scores $z < -1,65$. Un score *z* inférieur à ce seuil $z = -1,65$ permettra d'objectiver

facilement une performance pathologique, ou infranormale pour un patient sur une dimension donnée. Notons pour conclure cette section que l'expression des performances des patients avec des chiffres à virgule, qui plus est négatifs, peut rendre plus difficile la communication de ses résultats au patient. C'est principalement pour cette raison que des auteurs ont préconisé l'usage d'une autre forme de scores standards, les scores T ($\mu = 50$; $\sigma = 10$)[4] plutôt que des scores *z* (Crawford, 2004).

Une distribution avec un plancher et un plafond

Sur n'importe quel type de distribution de scores suivant une loi normale, 99,72% de la population aura un score *z* compris entre -3 et +3 (cf. figure 3) et la quasi-totalité de la population (99,94%) aura un score *z* compris entre -4 et +4. Que conclure si suite à une évaluation, un patient obtient un score $z = -8$? Nous pensons qu'il est problématique de conclure sans plus réfléchir que cette personne est « très déficitaire » sur la dimension évaluée. Il est intéressant de faire une analogie avec la manière dont fonctionnent les échelles de Wechsler pour la mesure de l'intelligence (WAIS et WISC) qui sont parmi les outils psychométriques les plus sophistiqués, leur rayonnement mondial et leur valeur commerciale permettant aux éditeurs de consacrer beaucoup de moyens à leur conception et leur modernisation. La moyenne des QI est 100, l'écart-type des QI est 15. Le score le plus bas que vous pouvez obtenir à une échelle de QI, si vous ne donnez aucune bonne réponse, est 40 ; soit la moyenne moins 4 écart-types ce qui vous donne un score $z = -4$. Le score le plus haut que vous pouvez obtenir avec une échelle de QI est 160 soit la moyenne plus 4 écart-types, soit score $z = 4$. Est-ce à dire qu'il n'est pas possible d'avoir une intelligence inférieure à QI = 40 ou supérieur à QI = 160 ? C'est très peu probable mais cela reste bien sûr possible. Les échelles de QI ne sont juste pas *conçues* pour mesurer ces niveaux de performance « extrêmes »[5]. Les items les plus difficiles sont encore trop faciles pour une personne ultra-intelligente et ne sont donc plus suffisamment discriminants. A l'image des échelles de QI, tous les instruments de mesure ont un plancher et un plafond contre lesquels les performances de quelques personnes « exceptionnelles »[6] vont venir buter. On retrouve également ces effets plancher et plafond avec les instruments de mesure « physiques » : peser un éléphant sur un pèse-personne va forcément conduire à un effet plafond. Bien comprendre cet outil que sont les scores *z*, ce n'est pas seulement comprendre que la moyenne de la distribution est 0 et l'écart-type est 1. Il faut aussi avoir conscience que l'échelle des scores *z* a un plancher ($z = -4$) et un plafond ($z = 4$) entre lesquels se distribuent les scores de la quasi-totalité de la population. Que conclure donc si un patient se trouve au-delà du seuil plafond avec par exemple un score $z = 8$? Imaginons un exemple simple : la performance à une tâche lambda se mesure en comptant le nombre d'erreurs du patient. La moyenne des erreurs dans la population est 8 erreurs avec un écart-

[4] La formule pour transformer un score brut en score T est : $T = \frac{10}{s_x} \times (x - m_x) + 50$

[5] Est-ce bien utile d'ailleurs, pour la prise en charge, de savoir exactement si un enfant à un QI de 40 ou de 36 ? La mesure doit rester un moyen et non une fin.

[6] Ou plus simplement de personnes à qui le test n'est pas destiné. Par exemple des adultes vont plafonner sur un test destiné à des enfants.

type de 2 erreurs. Votre patient a fait 24 erreurs soit $z = 8$[7]. Cette performance est tellement improbable (au regard de la distribution de fréquence des performances dans la population de référence), pratiquement impossible, qu'il convient de s'interroger sérieusement sur les causes qui ont pu y mener. Trois possibilités se présentent. La première est que le patient est effectivement très très déficitaire sur la dimension mesurée : vous êtes tombé sur un spécimen *extrêmement* rare. Dans ce cas de figure, il ne faut pas perdre de vue que le test n'a pas été originellement pensé pour des personnes aussi déficitaires et il est très probable que la tâche demandée n'était pas adaptée, manifestement, aux possibilités du patient. Cela revient à faire passer une WISC à un enfant avec un handicap mental profond ou à peser un éléphant sur une balance de cuisine. Il faut également garder à l'esprit que par définition, au-delà des seuils plancher et plafond, le test utilisé n'a plus aucun pouvoir de discrimination entre les sujets (qui « butent » sur le plafond ou le plancher). Il serait donc peu judicieux de considérer qu'un patient avec un score $z = 8$ à cette tâche est « plus déficitaire » qu'un patient avec un score $z = 6$. Bref, la mesure n'a pas grand sens et il serait bon de considérer la possibilité de mesurer les performances du patient avec un test plus adapté à son profil. La seconde possibilité, plus probable en réalité, est que le patient a échoué pour d'autres raisons qu'un déficit important sur la dimension mesurée. Les explications possibles sont nombreuses : le patient n'a pas compris la consigne, le patient a un trouble sensoriel, le patient a été dérangé au cours de la tâche, etc. Là encore, la mesure n'a pas grand sens. Enfin, troisième possibilité, le patient a été comparé aux normes d'une population qui n'est pas sa population de référence, par exemple, un enfant dont on calculerait le score *z* sur la base de normes établies chez l'adulte. En d'autres termes, le patient n'a pas échoué, c'est l'opération de conversion en scores *z* qui est fallacieuse. Nous revenons sur l'importance de bien cibler la population de référence dans la suite de cet article. Prendre conscience de l'existence de ces seuils plancher et plafond, ce n'est pas seulement mieux utiliser les scores *z*, c'est aussi la possibilité pour le clinicien de constater que son outil de mesure n'était sans doute pas approprié pour mesurer telle dimension avec tel patient.

Les scores seuils : dichotomiser la distribution en deux catégories de scores

Les scores *z* permettent d'exprimer avec un seul chiffre la performance du patient et sa place relative par rapport à sa population de référence. Souvent, le praticien voudra aller plus loin dans son interprétation de la performance en la classant de manière binaire : performance normale vs. performance pathologique ou déficitaire. Cette opération constitue un appauvrissement de l'information quant à la performance du sujet et expose le praticien au risque de faire des erreurs (faux positifs et faux négatifs, cf. section suivante) mais dans des

[7] Nous insistons sur le fait que si certaines variables communément mesurées dans les tests orthophoniques ou neuropsychologiques (compter les erreurs ou les tentatives, faire tourner un chronomètre) semblent en apparence ne pas avoir de plafond, ceci est un leurre, conforté par le fait que la conversion arithmétique des chiffres bruts en scores *z* aboutit toujours à « un chiffre ». Un effet plafond (ou un effet plancher) est classiquement le résultat d'une inadéquation entre le niveau de difficulté de la tâche et le niveau d'efficience du sujet. On peut affirmer qu'un sujet qui réalise une tâche en commettant 22 erreurs là où 99,94% de la population ne dépasse pas 16 erreurs (soit 8, le nombre d'erreurs moyen + 4 écart-types) « plafonne », au sens d'une telle inadéquation, quand bien même le compteur des erreurs ou le chronomètre continue de tourner.

contextes de prise de décision, elle peut s'avérer utile pour décider de la suite à donner à la relation de soin.

Nous avons vu que la conception statistique de la normalité, qui va de pair avec l'utilisation des scores *z*, repose sur l'idée que la rareté de certaines performances peut permettre de les considérer comme pathologiques. A partir de quel niveau de rareté une performance peut-elle être jugée pathologique ? Le seuil de $z < -1,65$ (correspondant au 5% des performances les plus rares / faibles) est souvent considéré comme le seuil numérique définissant le passage de la normalité à l'anormalité (Amieva, Michael, & Allain, 2011). Il est probable que ce seuil de 5% nous vienne du seuil de la rareté en statistiques tel qu'il a été défini de manière tout à fait arbitraire par Fisher dans les années 1920 (Field, 2009; Schwartz, 2012). La logique d'un test statistique est de calculer la probabilité (valeur *p*) avec laquelle nous observerions un effet X dans un échantillon si cet effet X n'existait pas dans la population. Si cette probabilité est très faible, alors le test est significatif et on conclut à l'existence de l'effet X dans la population. A l'époque de Fisher, les ordinateurs ne sont pas là pour calculer une valeur *p* exacte pour chaque test statistique effectué comme c'est le cas aujourd'hui. Il faut donc construire des tables de probabilités à la main ce qui implique une logique de seuils. Fisher produit donc des tables pour des probabilités précises : 5%, 2% et 1% de chance. L'usage fera le reste : 5% devient en sciences humaines et sociales, le seuil de la rareté statistique. Néanmoins, l'idylle s'arrête là. Aujourd'hui, les statisticiens plaident en faveur d'une diminution de ce seuil de rareté en sciences humaines et sociales[8], de 5% à 5‰ (Benjamin et al., 2017; Button et al., 2013) tandis que de leur côté, les psychologues de plus en plus portés sur l'étude des troubles cognitifs légers (MCI en anglais, Mild Cognitive Impairment), s'intéressent à la zone des performances « limites », comprise entre $z < -1,65$ et $z < -1,28$. Une étude de Antinori et al. (2007, citée par Amieva et al., 2011) a même pris le parti de considérer comme non normaux tous les scores *z* inférieurs à -1. L'étude des MCI soulève immanquablement la question de la normalité : sont-ils des troubles neurocognitifs d'intensité légère ou des performances normales basses (Godefroy, Diouf, Bigand, & Roussel, 2014) ? Si l'on suit le critère d'Antinori et al (2007), un coup d'œil à la figure 3 permet de constater que 32% de la population se rend coupable de performances « rares » (avec des performances $z < -1$ et par symétrie, $z > 1$) ! Pas étonnant dans ce contexte que des psychologues s'alarment et lancent un appel pour « sauver la normalité » (Frances, 2013). Un changement de domaine permet d'éclairer l'aspect très arbitraire de ces choix de seuils. Considérons que la taille moyenne des femmes françaises adultes est 163 cm (écart-type : 5,5 cm) (Sempé, Pédron, & Roy-Pernot, 1979). La taille correspondant à $z = -1,65$ est 154 cm. Considère-t-on que les femmes mesurant moins de 154 cm ont une taille pathologiquement petite ? On voit que le caractère pathologique ou non d'un comportement dépend largement de la nature du comportement mesuré lui-même.

Si 5% constitue encore le seuil de la rareté le plus consensuel chez la plupart des praticiens, un débat porte sur la manière de répartir ces 5% de performances les plus rares. En utilisant le seuil $z < -1,65$, le praticien considère uniquement les 5% des performances *basses* les plus rares. Mais de l'autre côté de la distribution normale se trouve des performances tout aussi rares, bien que *hautes* (cf. figure 4). Si le praticien considère aussi ces performances hautes

[8] Ce seuil étant déjà largement plus bas en sciences physiques.

comme pathologiques, le voilà qui classe 10% de la population comme ayant des performances pathologiques au lieu des 5% communément admis. Pour rester dans le périmètre consensuel des 5%, il doit abaisser son score seuil de manière à avoir 2,5% de performances rares-et-basses à gauche de la distribution ($z < -1,96$, cf. figure 4 et tableau 1) et 2,5% de performances rares-et-hautes à droite de la distribution ($z > 1,96$). Dans ce cas de figure, un patient ayant un score *z* compris entre -1,96 et -1,65 ne devrait pas être jugé pathologique. La décision de faire une lecture unilatérale de la distribution normale (considérer uniquement les performances basses, $z < -1,65$) ou une lecture bilatérale (considérer les performances basses, $z < -1,96$, *et* les performances hautes, $z > 1,96$) dépend de la dimension évaluée et des habitudes prises dans la communauté. Par exemple, la mesure de l'intelligence avec les échelles de QI est basée sur une lecture bilatérale de la distribution : un enfant a un retard mental quand il a un QI inférieur à 70[9] ($z < -2$) et une précocité quand il a un QI supérieur à 130 ($z > 2$).

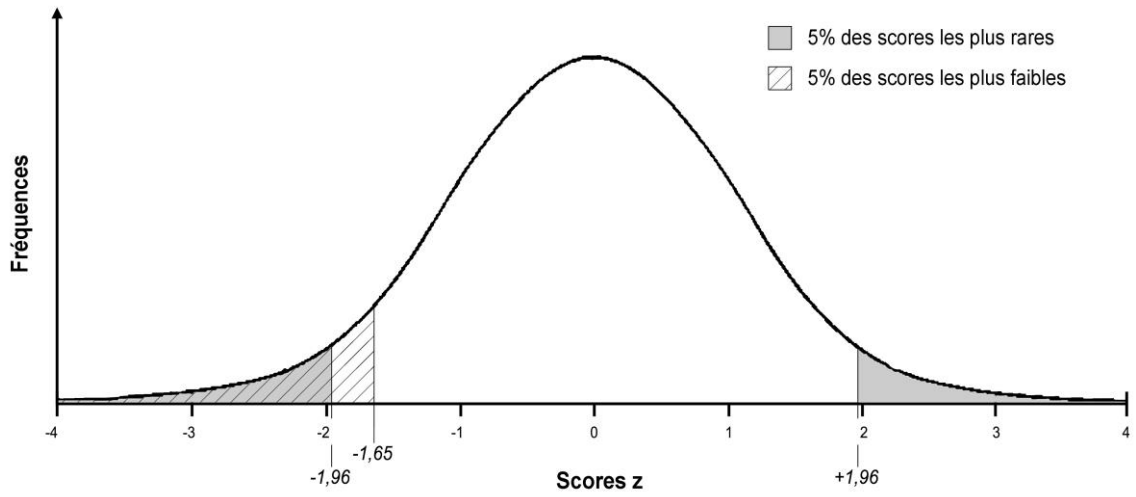

Figure 4. Scores rares et scores faibles sur une distribution normale théorique.

Un autre facteur influençant le choix d'un score seuil, outre le caractère mesuré, est l'objectif de la mesure. Il arrive parfois que les tests ne soient pas utilisés « pour eux-mêmes » mais pour dépister des maladies. Par exemple, les neuropsychologues utilisent souvent le RL/RI-16, un test initialement conçu pour évaluer la mémoire épisodique, pour dépister la maladie d'Alzheimer (Amieva, 2016). Si l'objectif premier n'est plus de situer la performance du patient par rapport à sa population de référence quant à la mémoire épisodique, mais de poser un diagnostic, ou a minima des hypothèses, quant à la présence de la maladie d'Alzheimer, il est clair que le score seuil n'implique plus uniquement des considérations sur la rareté des performances. Il faut procéder à des études poussées pour décider au cas par cas comment un test donné peut être prédicteur d'une maladie donnée. Il revient aux concepteurs du test de préciser les usages pouvant être faits du test, comparatif ou diagnostic, les scores seuils à utiliser

[9] Plus précisément, le diagnostic de retard mental se fait quand l'enfant cumule un QI < 70 et des difficultés objectivées d'adaptation sociale.

selon ces usages et éventuellement les normes à utiliser selon ces usages (cf. plus loin la section sur les normes).

Gare aux faux positifs

Une fois établi le score seuil séparant les performances normales des performances pathologiques, il est en apparence simple de conclure quant au caractère normal ou pathologique de la performance d'un patient. Cependant, avant de conclure, le praticien doit se montrer attentif à ne pas produire des conclusions erronées. Trois causes possibles pouvant conduire à produire des faux-positifs ou des faux-négatifs sont discutées dans cette section : quand la distribution des scores ne suit pas la loi normale, quand on ne dispose pas de normes adaptées et enfin quand on néglige l'erreur de mesure.

Quand la distribution des scores ne suit pas la loi normale

Une idée répandue parmi les praticiens est qu'il n'est pas licite de calculer des scores *z* lorsque la distribution des scores bruts ne suit pas une loi normale mais qu'heureusement, la plupart des variables psychologiques, à l'instar de la grande majorité des caractères biologiques, se distribue selon une loi normale, à la manière des QI. Cette idée est doublement trompeuse. Premièrement, il est toujours possible de transformer des scores bruts en scores *z*, quelle que soit la forme de la distribution comme nous l'avons vu plus haut. Mais cette opération ne change en aucun cas la forme de la distribution. Deuxièmement, la distribution normale n'est que le modèle *postulé* de la distribution naturelle des caractères biologiques et psychologiques. L'observation plus détaillée de données réelles montre que les distributions véritablement normales sont plutôt rares (Gould, 1997). La réalité est pleine de « barrières » biologiques, physiques, sociales qui empêchent les scores de se distribuer à l'infini des deux côtés de la moyenne et en conséquence, les distributions asymétriques sont nombreuses. Un cas d'école est la distribution des salaires en France : cette distribution est asymétrique, bloquée sur la gauche par l'obligation qu'ont les entreprises de s'acquitter d'un salaire minimum (le SMIC) et tirée sur la droite par une petite minorité de très haut revenus. Cette asymétrie vers la droite est attestée par le décalage entre la valeur du salaire mensuel net médian[10] qui s'élevait à 1675€ en 2010 en euros constant et la valeur du salaire net mensuel moyen qui s'élevait elle à 2082€.

Dans le cadre plus restreint de la neuropsychologie, les cas de distributions asymétriques sont communs parce que les tests ont généralement pour objectif de mettre en évidence des performances déficitaires, i.e. inférieures à la moyenne. Le test de Stroop (Stroop, 1935), le Trail Making Test (Reitan, 1958), le Wisconsin Card Sorting Test (Anderson, Damasio, Jones, & Tranel, 1991), le test de Brixton ou encore le Hayling test (Burgess & Shallice, 1997) sont des exemples de tests pour lesquels certains indicateurs de performance, calculés en nombre d'erreurs, sont proches de 0 pour les sujets sains et augmentent en fonction du déficit des patients. Il y a ainsi un important « effet plancher » qui a pour conséquence de rendre la

[10] Le salaire tel que 50% de la population touche moins et 50% de la population touche plus.

distribution des scores fortement asymétrique vers la droite. C'est également le cas de la plupart des tests mesurant des temps de réponse ou de réaction : il y a toujours un temps incompressible pour effectuer une tâche qui constitue une barrière physique sur laquelle viennent buter les performances des sujets les plus rapides. Les choses se compliquent d'autant plus que les temps de réponses peuvent se distribuer d'une manière approchant la normalité statistique pour les sujets jeunes et devenir largement non normaux chez les sujets âgés (e.g., les distributions des temps de réalisation au TMT B selon l'âge, figure 5). A l'inverse des outils de diagnostic rapide de l'efficience globale comme le MMSE (Folstein, Folstein, & McHugh, 1975) ou la MoCA (Nasreddine et al., 2005) sont asymétriques vers la gauche, les sujets sains « plafonnant » près du score maximal.

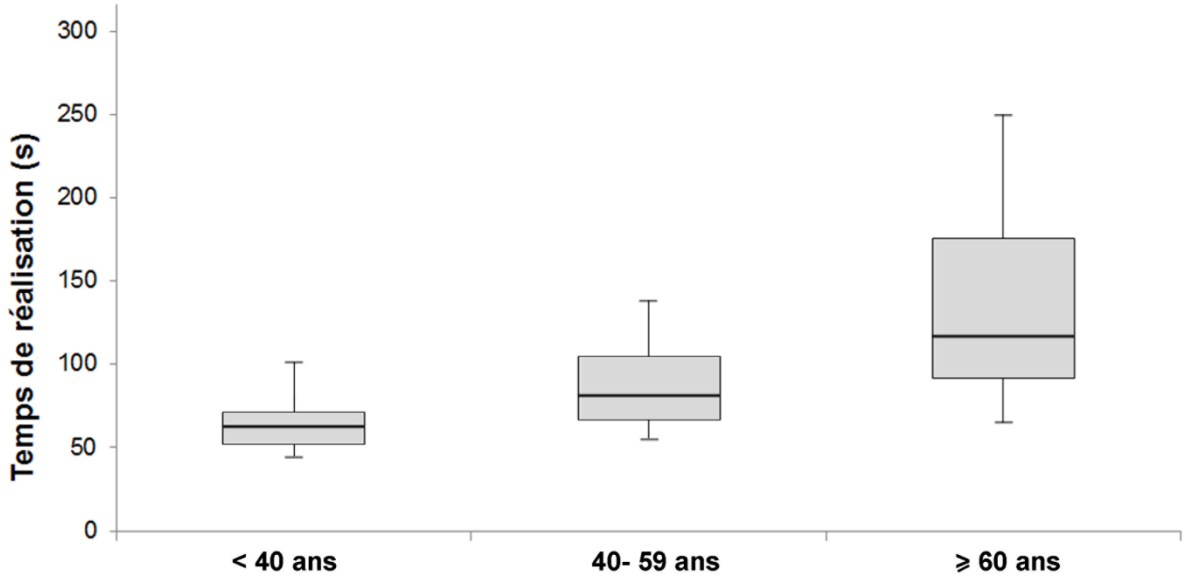

Figure 5. Boîtes à moustaches (déciles / quartiles / médiane) représentant la distribution des temps de réalisation (en secondes) du parcours au TMT B pour 3 classes d'âge, niveau d'éducation 2. Normes du GREFEX.

S'il est toujours possible de transformer ces scores bruts en scores *z* alors que la distribution des scores bruts n'est pas normale, il est parfaitement illicite d'utiliser la table de probabilité de la loi normale. Lorsque la distribution des valeurs est asymétrique, les correspondances rapportées dans le tableau 1 cessent d'être valides. Il faut alors utiliser les **quantiles**. Un quantile est un score tel qu'on a une proportion déterminée de la population (*p*) qui précède ce score et une proportion déterminée (1-*p*) qui suit cette valeur. Le quantile le plus connu est la médiane qui coupe la distribution de fréquence en deux moitiés égales : si le score médian à un test est 12 alors cela signifie que 50% des sujets ont un score inférieur ou égal à 12 et 50% des sujets ont un score supérieur ou égal à 12. Un découpage en deux moitiés ne permet pas de situer finement une personne donnée, c'est pourquoi les concepteurs des tests utilisent plus souvent les déciles (9 déciles qui découpent la distribution en 10 dixièmes de taille égale) ou les centiles (99 centiles qui découpent la distribution en 100 centièmes de taille

égale). Le premier décile (noté D1) est le score tel que 10% des sujets ont un score inférieur ou égal et 90% des sujets ont un score supérieur ou égal. Le 95ème centile (noté C95) est le score tel que 95% des sujets ont un score inférieur ou égal et 5% des sujets ont un score supérieur ou égal. On voit ainsi que la médiane est à la fois le 5ème décile (D5) et le 50ème centile (C50). Les quantiles ont l'avantage de pouvoir être utilisés quelle que soit la forme de la distribution et ils expriment directement la rareté d'un score. Cependant en transformant le score brut par une information sur son rang dans la distribution, on perd la possibilité de faire des calculs arithmétiques, par exemple de moyenner les différents scores d'un patient sur plusieurs tests. En effet, les quantiles sont une échelle ordinale et non une échelle numérique : l'écart de score bruts entre le 5ème et le 10ème centile n'est pas nécessairement le même que l'écart entre le 45ème et le 50ème centile.

Prenons un exemple pour illustrer le fait que lorsque la distribution de fréquence n'est pas normale, les scores *z* n'expriment plus les proportions indiquées dans la table de probabilités de la loi normale, ce qui peut augmenter considérablement le risque de faire des faux-positifs, c'est-à-dire de conclure que la performance du patient est pathologique alors que ce n'est pas le cas. Imaginons qu'un(e) orthophoniste utilise un test impliquant de mesurer le nombre d'erreurs que fait le patient. Le test est assez facile et on s'attend à trouver très peu d'erreurs chez les patients « normaux ». Cinq cents participants ont été testés, la distribution des résultats est représentée sur la figure 6. Cette distribution est caractérisée par une large asymétrie vers la droite. Cette asymétrie se retrouve dans le décalage des 3 indicateurs de tendance centrale vers la droite : le mode vaut 0, la médiane vaut 2 et la moyenne est égale à 3 (s = 3,15). Classiquement, si un patient est dans les 5% des patients faisant le plus d'erreurs alors sa performance est jugée pathologique. Une analyse de cette distribution montre que le 95ème centile, au-delà duquel se trouvent les 5% des scores les plus rares, correspond à une performance de 10 erreurs. Seuls 5% de notre échantillon font plus de 10 erreurs au test. Le score *z* correspondant à ce score brut est $\frac{10 - 3}{3,15} = 2,22$. Ainsi, le score *z* au-delà duquel se trouve 5% de la population est 2,22 et non pas 1,65 comme le prévoit la table (cf. figure 6). Cette différence est due à la non normalité de la distribution. Si un(e) orthophoniste utilisait le score *z* de 1,65 pour identifier les performances normales et les performances pathologiques, il ou elle classerait environ 9% des participants dans la pathologie au lieu de 5% !

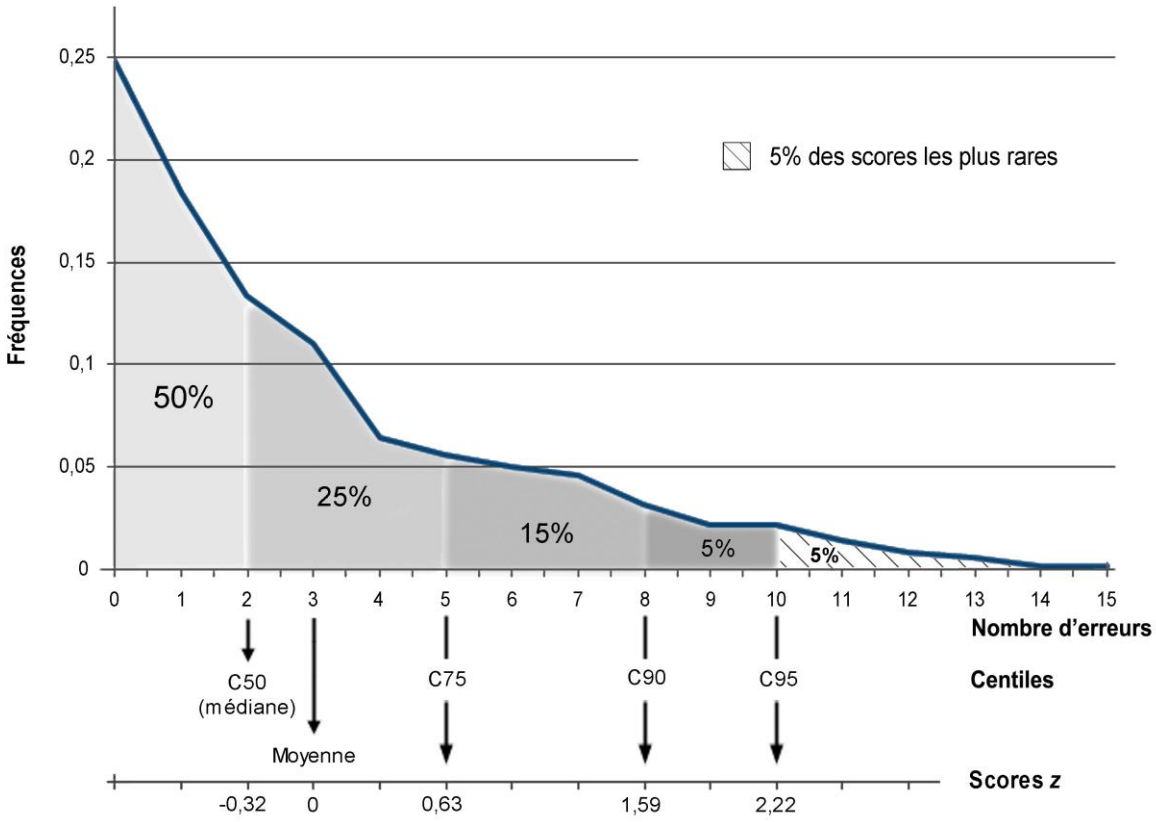

Figure 6. Distribution de fréquence pour le nombre d'erreurs à un test fictif (*m* = 3 ; *s* = 3,15 ; N = 500). Le nombre d'erreurs est également exprimé en centiles et en scores *z*

Cet exemple démontre le risque important de faire des faux-positifs qu'il y a à utiliser des scores *z* sans s'être auparavant assuré que la distribution des scores était normale[11]. L'absence de normalité de la distribution peut conduire le praticien à élargir substantiellement, sans qu'il en ait même conscience, l'étendue classiquement acceptée de l'anormalité des performances psychologiques. Pourtant, le présupposé selon lequel la distribution des scores suit une loi normale est tellement prégnant que de nombreux professionnels de la santé n'expriment plus ce seuil entre normalité et pathologie en centiles (il faut être dans les 5% des scores les plus extrêmes) mais directement en nombre d'écart-types (il faut être au-delà de 2 écart-types de la moyenne). Ce raccourci pose évidemment problème dès que les distributions sont non-normales. Dans ces cas, l'utilisation des scores *z* doit être résolument évitée.

Quand on ne dispose pas de normes adaptées

Lorsque l'on calcule un score *z*, on confronte en principe le score du sujet aux scores obtenus dans sa population de référence. Pour cela, il faut disposer de la moyenne et de l'écart-type des scores de la population à laquelle appartient le patient. En pratique, nous ne

[11] L'information sur la normalité de la distribution des scores devrait être indiquée clairement avec le manuel ou les normes du test.

connaissons que très rarement la moyenne et l'écart-type d'une population. A la place, les concepteurs de tests constituent des normes sur la base d'un échantillon représentatif de la population d'intérêt. C'est l'étape d'étalonnage du test. Le praticien doit être très vigilant à la manière dont cet échantillon-étalon a été constitué et toujours se poser la question de si la comparaison entre cet échantillon et le patient qu'il a en face de lui est pertinente. La règle d'or est que votre patient aurait pu potentiellement faire partie de l'échantillon-étalon. Plusieurs paramètres doivent être considérés :

- La taille de l'échantillon-étalon est-elle suffisamment importante ? Ci-dessus (et dans la note de bas de page N°3), nous avons indiqué qu'à défaut de connaître les paramètres de notre population d'intérêt (moyenne μ et écart-type σ), nous utilisions dans la formule des scores *z* les statistiques de l'échantillon-étalon (moyenne *m* et écart-type *s*). Mais si cet échantillon-étalon est de trop petite taille, la moyenne (*m*) et l'écart-type (*s*) utilisés dans la formule seront une mauvaise estimation des paramètres réels dans la population (μ et σ) et le calcul du score *z* conduira à une évaluation biaisée. Cela peut être particulièrement le cas quand la population totale est divisée en sous-groupes sur la base de variables démographiques. Dans ces cas, des solutions statistiques accessibles existent consistant à non pas baser le calcul sur l'assomption d'une distribution des scores suivant la loi normale mais suivant la loi de Student (voir Atzeni, 2009 ; Crawford & Howell, 1998).
- A quelle date et dans quel pays ont été constituées les normes ? Il n'est pas rare de voir des praticiens utiliser des tests traduits de manière artisanale pour ensuite comparer la performance d'un français en 2018 à celles d'américains en 1972. Calculer le score *z* d'un individu sur base d'une norme issue d'un autre pays, d'une autre culture, risque de conduire là encore à une conclusion erronée (Amieva et al., 2011).
- Les normes constituées sont-elles vraiment représentatives de la population générale quant à l'âge des individus et leur niveau d'étude ? Constituer des normes se fait souvent sur la base du volontariat des sujets et ce mode de recrutement « sélectionne » un profil particulier, généralement plus favorisé sur le plan culturel et cognitif (Amieva et al., 2011). Or, on comprend bien qu'un élève « normal » sera jugé déficitaire s'il est comparé à un groupe de très bons élèves. Là encore, le praticien s'expose à faire des faux positifs, *i.e.* voir un déficit là où il n'a qu'un échantillon-étalon non représentatif. Il arrive parfois que l'on compare les performances du patient à celles d'un sous-échantillon qui lui est apparié sur la base de caractéristiques démographiques : âge, sexe, niveau d'étude. Etre comparée à des femmes âgées et diplômées quand on est soi-même une femme âgée et diplômée plutôt qu'à la population tout venante augmente la précision de la comparaison en contrôlant la variance indésirable liée au sexe, à l'âge ou au niveau d'étude. Attention cependant, cette technique a pour corollaire de réduire encore la taille des échantillons. De plus, dans un objectif diagnostic et non comparatif, la variance liée à ces variables démographiques n'est pas toujours indésirable, l'âge et le niveau d'étude étant par exemple prédicteurs de la maladie d'Alzheimer (Amieva, 2016).

- Les personnes incluses dans l'échantillon-étalon représentent-elles la population « saine » ou la population tout-venante ? Il est fréquent que pour mesurer de manière « pure » une dimension cognitive (par exemple les fonctions exécutives), les concepteurs du test écartent de l'échantillon-étalon toutes les personnes à risque de perturbation. C'est le cas par exemple pour la batterie GREFEX, utilisée en neuropsychologie (Roussel & Godefroy, 2008) : la liste des critères d'exclusion de l'échantillon-étalon est très longue et comprend par exemple « éthylisme actuel ou antécédent de sevrage ». Ces exclusions sont nécessaires pour établir les performances « normales » mais est-il pertinent alors de comparer la performance de votre patient, « éthyliste » à ses heures, à cet échantillon-étalon qui n'est délibérément pas représentatif de personnes comme lui ?

En bref, dès que le praticien ignore comment ont été constituées les normes qu'il utilise ou qu'il doute de leur pertinence au regard du profil particulier de son patient, il s'expose à produire des faux-positifs ou des faux-négatifs en utilisant ces normes.

Quand on néglige l'erreur de mesure

Un test ne mesure jamais la compétence d'un sujet mais sa performance à un moment t et dans un contexte donné. Ainsi 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. La sensibilité des tests à l'erreur de mesure est mesurée avec le coefficient de fidélité *r*, coefficient qui varie de 0 à 1 et qui n'est ni plus ni moins qu'un coefficient de corrélation entre deux mesures par le 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. En psychologie, on considère généralement que la fidélité d'un test commence à être acceptable à partir de *r* = .70 (Vautier & Gaudron, 2002). La fidélité des tests est un indicateur important qui devrait toujours être mentionnée dans les manuels accompagnant les tests. Parce qu'un test ne mesure jamais la performance « vraie » d'un patient mais cette performance « vraie » mélangée à de l'erreur de mesure, il est déconseillé de communiquer au patient un chiffre isolé qui sera souvent interpréter comme le reflet fiable et direct de sa performance. A la place, mieux vaut calculer un intervalle de confiance autour de la performance observée, basé sur l'écart-type des scores au test et le coefficient de fidélité[12]. Pour calculer un intervalle de confiance à 95% de score *z* (un intervalle dans lequel devrait se trouver la performance « vraie » dans 95 cas sur 100), on utilise la formule (2) ci-dessous[13].

[12] Certains auteurs préconisent de plutôt faire l'intervalle de confiance autour de la présumée performance vraie, voir note de bas de page suivante pour plus de détails.

[13] Pour calculer un intervalle de confiance (IC) sur la base de scores bruts (et non des scores *z*), il faut utiliser la formule suivante : $IC_{95} = x \pm 1{,}96 \times s \times \sqrt{1-r}$ où *s* est l'écart-type de la distribution des scores (omis de la

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

Illustrons ce calcul avec un exemple. Marina est une petite fille de 10 ans qui passe le subtest « Information » de la WISC-III. Les scores aux subtests de la WISC ont une moyenne m = 10 (s = 3). Le manuel d'interprétation de la WISC-III précise qu'à 10 ans, le coefficient de fidélité de ce subtest est r = 0,79 (soit une fidélité plutôt bonne). Marina a un score brut de 5. Converti en scores *z*, Marina a un score z = -1,67. Elle se situe donc au-delà du seuil de la rareté de -1,65 (5%) et le praticien serait tenté de conclure à une performance pathologique. Si l'on calcule l'intervalle de confiance à 95% dans lequel se situe la performance « vraie » de Marina (l'intervalle dans lequel tombera 95 fois Marina si on lui faisait passer le test 100 fois), on obtient le résultat suivant : z = -2,56 pour la borne inférieure et z = -0,77 pour la borne supérieure[14]. Il est donc tout à fait probable que Marina ait un score supérieur au seuil z = -1,65 ce qui conduirait le praticien à une conclusion totalement différente. Pour pouvoir conclure de manière certaine quant au caractère normal ou pathologique d'une performance, le praticien devrait s'assurer, respectivement, que la borne inférieure de l'intervalle de confiance est supérieure à z = -1,65 ou que la borne supérieure de l'intervalle de confiance est inférieure à z = -1,65. Si l'intervalle de confiance est à cheval sur le score seuil choisi pour distinguer le normal du pathologique, alors le praticien prend le risque de faire un faux positif ou un faux négatif en classant la performance du patient dans l'une de ces deux catégories.

Conclusion

L'utilisation des scores *z*, couplée avec celle de scores-seuils, permet d'aboutir assez facilement à la conclusion qu'une performance est normale ou pathologique. Cependant, outre les considérations cliniques qui doivent permettre de nuancer ce type de conclusion binaire, des limites inhérentes à l'utilisation même des scores *z* doivent rester sous le radar du praticien.

- La première de ces limites concerne la distribution des scores dans l'échantillon-étalon : cette distribution doit suivre, au moins approximativement, la loi normale. Malheureusement, l'information sur la forme de la distribution est rarement mentionnée dans les manuels des tests, quand ces manuels existent. Heureusement, il n'est pas très difficile pour un praticien expérimenté qui connaît son outil

formule (2) car dans le cas des scores *z*, s = 1). 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. La partie « $s \times \sqrt{1-r}$ » de la formule désigne la SE_M, l'erreur-type de mesure qui est parfois également indiquée dans les manuels des tests. Pour calculer l'IC autour de la note vraie estimée plutôt qu'autour de la note observée, on utilise l'erreur-type d'estimation (SE_E) plutôt que la SEM dans la formule, celle-ci valant : « $s \times \sqrt{1-r} \times r$ ».

[14] Notons que cet intervalle, déjà très grand, est calculé sur la base du coefficient de fidelité fournit par les concepteurs du test. Ce coefficient est calculé sur la base de conditions de passation optimales, dans le but précis de réduire l'erreur de mesure et d'avoir le meilleur coefficient de fidélité possible. Dans sa pratique quotidienne, le praticien offrira souvent des conditions moins favorables à son patient, générant davantage d'erreur de mesure ce qui nous laisse penser qu'en pratique, l'intervalle de confiance dans lequel se trouve la « performance vraie » du patient pourrait être encore plus grand.

d'imaginer quelle pourrait être la distribution théorique des scores dans la population. Avec un peu d'habitude, on comprend pourquoi il est probable que des tailles ou des QI peuvent se distribuer de manière normale mais pas des temps de réponse ou un nombre d'erreurs. En cas de doute, le praticien doit absolument se rabattre sur l'utilisation des centiles dont l'utilisation n'est pas dépendante du type de distribution.

- La deuxième limite a trait à la taille et la constitution de l'échantillon-étalon. Même si la distribution théorique des scores dans la population de référence suit une loi normale, la distribution des scores dans l'échantillon-étalon n'a que peu de chance d'avoir une allure « normale » si cet échantillon compte moins d'une centaine d'individus ce qui nous ramène à la première limite. Finalement, même si l'échantillon-étalon est d'une taille correcte, la conversion de la performance d'un patient en scores *z* n'aura de sens que si le patient est représentatif des personnes constituant l'échantillon-étalon. Dans le cas contraire, la conversion en score *z* reviendra à comparer des pommes avec des oranges et probablement à faire des faux négatifs ou plus grave, des faux positifs.
- La troisième limite touche au caractère relatif du score individuel mesuré chez le sujet et au caractère arbitraire du score-seuil auquel le premier est comparé. Le score $z < -1{,}65$ jouit d'une telle notoriété qu'il est parfois directement synonyme de pathologie dans la tête des cliniciens. Il est utile de se rappeler que ce seuil est davantage le résultat d'une histoire et d'habitudes que le fruit d'une réflexion concertée sur ce que doit être la normalité et la pathologie dans le domaine de la psychologie ou de l'orthophonie. Nous avons par ailleurs vu que si le praticien prend en compte l'erreur de mesure dans son évaluation de la performance du patient, ce qu'il devrait bien sûr faire autant que possible, il va fréquemment se trouver avec une performance « vraie » à cheval sur la zone « pathologique » et sur le zone « normale » de la distribution des scores. Une situation délicate à interpréter mais qui reflète mieux la réalité et la complexité du monde qu'un découpage binaire entre le normal et le pathologique.
- La quatrième limite concerne l'échelle des scores *z*. Avec une moyenne de 0 et un écart-type de 1, la conversion de scores bruts en scores *z*, en particulier la conversion des scores inférieurs à la moyenne, aboutit toujours à « un chiffre » et rien ne vient alerter le praticien en cas de bizarrerie. Le praticien qui utilise les scores *z* devrait être attentif au fait que la distribution de ces scores a un plancher ($z \approx -4$) et un plafond ($z \approx +4$) qui, s'ils sont dépassés, doivent agir comme un signal sur la probable invalidité de la mesure. En plus de n'avoir ni décimale, ni signe négatif, ce qui les rendent plus facilement communicables aux patients, les scores T évitent l'écueil susmentionné. En effet, avec une moyenne de 50 et un écart-type de 10, la conversion d'un score brut se trouvant au-delà de 5 écart-types de la moyenne aboutit à… un chiffre négatif. Ce qui précisément n'est pas censé se produire.

Ces limites, par définition, restreignent le champ d'utilisation des scores *z* mais cette restriction est positive si elle améliore la pratique *in fine*. La mesure quantitative des

performances à l'aide d'un test est souvent utile mais la marque d'un bon praticien est certainement de savoir y renoncer quand les conditions ne sont pas réunies. Même quand les conditions sont réunies, la mesure ne saurait être interprétée de manière décontextualisée. Dans cet article, nous avons délibérément focalisé notre propos sur des aspects psychométriques et statistiques. Pour autant, il nous paraît essentiel d'insister ici en conclusion sur le fait qu'un simple score *z* inférieur à un seuil quel qu'il soit ne doit jamais permettre d'affirmer que le patient est pathologique ou déficitaire. Le clinicien réunit des informations par différents moyens : l'observation, l'entretien, les tests en font partie. Le diagnostic est le fruit de l'intégration et de l'interprétation de toutes ces informations. Une littérature importante, que nous rejoignons, a mis l'accent sur la nécessité absolue de ne pas réduire le patient à une performance chiffrée. L'évaluation, en orthophonie ou en psychologie est avant tout une relation de soin et elle suppose d'abord de reconnaître l'autre comme une personne à part entière, avec son identité, son histoire et ses particularités. Concluons avec un patient qui viendrait voir le professionnel avec une plainte de perte d'efficience suite à un accident ou un traitement. Une évaluation standardisée et un calcul de scores *z* place le sujet au-dessus du score seuil de la pathologie que vous avez défini (par exemple, $z < -1{,}65$). Que faire de la plainte du patient si elle n'est pas objectivée par le test ? Il faut se souvenir que toute la mécanique des scores *z* repose sur une vision statistique de la normalité qui n'est pas la seule possible. Vue par le prisme de la normalité fonctionnelle que nous avons évoquée au début de cet article, la plainte du patient se comprend plus facilement : celui-ci avait une excellente efficience cognitive ou langagière avant son traitement ou son accident et était au-dessus de la moyenne de son groupe de référence. Suite au traitement ou à l'accident, la perte d'efficience est bien réelle, d'où la plainte, mais « invisible » pour le praticien qui s'en tiendrait simplement à ses scores *z* et à la normalité statistique.

Bibliographie

AMIEVA, Hélène. Normal, pas normal ? In : BELIN, Catherine, MAILLET, Didier, AMIEVA, Hélène (éds.). *L'évaluation neuropsychologique : De la norme à l'exception*. Louvain la Neuve : De Boeck Université, 2016, p. 1-10.

AMIEVA, Hélène., MICHAEL, George A., ALLAIN, Philippe. Les normes et leur utilisation. In THOMAS-ANTÉRION, Catherine, BARBEAU, Emmanuel (éds.). *Neuropsychologie en pratique(s)*. Marseille : Solal, 2011, p. 75-85.

ANDERSON, Steven W., DAMASIO, Hanna, JONES, R. Dallas, *et al.* Wisconsin Card Sorting Test performance as a measure of frontal lobe damage. *Journal of Clinical and Experimental Neuropsychology*, 1991, vol. 13, no 6, p. 909-922.

ATZENI, Thierry. Statistiques appliquées aux études de cas unique : méthodes usuelles et alternatives. *Revue de neuropsychologie*, 2009, vol. 1, no 4, p. 343-351.

BENJAMIN, Daniel J., BERGER, James O., JOHANNESSON, Magnus, *et al.* Redefine statistical significance. *Nature Human Behaviour*, 2018, vol. 2, no 1, p. 6.

BURGESS, Paul W. et SHALLICE, Tim. The hayling and brixton tests. 1997.

BUTTON, Katherine S., IOANNIDIS, John PA, MOKRYSZ, Claire, *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 2013, vol. 14, no 5, p. 365.

CANGUILHEM, Georges. *Le normal et le pathologique*. Paris : Presses Universitaires de France – PUF, 2013, 12ème édition (1ère édition1966).

CRAWFORD, John R. Psychometric foundations of neuropsychological assessment. In : GOLDSTEIN, Laura H., McNEIL, Jane, E. (éds). *Clinical neuropsychology: A practical guide to assessment and management for clinicians*. Chichester : Wiley, 2004, p. 121-140.

CRAWFORD, John R. et HOWELL, David C. Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist*, 1998, vol. 12, no 4, p. 482-486.

FIELD, Andy. *Discovering Statistics Using SPSS*. London : Sage publication, 2009, 3ème édition.

FOLSTEIN, Marshal F., FOLSTEIN, Susan E., et MCHUGH, Paul R. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 1975, vol. 12, no 3, p. 189-198.

FRANCES, Allen. Saving normal: An insider's revolt against out-of-control psychiatric diagnosis, DSM-5, big pharma and the medicalization of ordinary life. *Psychotherapy in Australia*, 2013, vol. 19, no 3, p. 14.

GODEFROY, Olivier, DIOUF, Momar, BIGAND, Charlotte, *et al.* Troubles neurocognitifs d'intensité légère ou performances normales basses?. *Revue de neuropsychologie*, 2014, vol. 6, no 3, p. 159-162.

GOULD, Stephen. J. *L'éventail du vivant : le mythe du progrès*. Paris : Editions du Seuil, 1997.

NASREDDINE, Ziad S., PHILLIPS, Natalie A., BÉDIRIAN, Valérie, *et al.* The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 2005, vol. 53, no 4, p. 695-699.

REITAN, Ralph M. Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and motor skills*, 1958, vol. 8, no 3, p. 271-276.

ROUSSEL, Martine et GODEFROY, Olivier. La batterie GREFEX: données normatives. In GODEFROY, Olivier & GREFEX (éd.), *Fonctions exécutives et pathologies neurologiques et psychiatriques*. Marseille : Solal, 2008, p. 231–52.

SCHWARTZ, Claudine. La preuve par les chiffres (evidence based): de quoi s' agit-il?. *Statistique et enseignement*, 2012, vol. 3, no 2, p. 3-21.

SEMPÉ, Michel, PÉDRON, Guy, et ROY-PERNOT, Marie-Paule. *Auxologie: méthode et séquences*. Théraplix, 1979.

STROOP, J. Ridley. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 1935, vol. 18, no 6, p. 643.

VAUTIER, Stéphane et GAUDRON, Jean-Philippe. Intégrer l'erreur de mesure dans l'interprétation quantitative des scores individuels. *Pratiques psychologiques*, 2002, no 2, p. 97-108.