De l'usage des tests : Aspects métrologiques, statistiques et interprétatifs

Marc Aguert

PRE PRINT

Aguert, M. (2022). De l'usage des tests : Aspects métrologiques, statistiques et interprétatifs. In C. Sainson, C. Bolloré, & J. Trauchessec (Éds.), *Neurologie et orthophonie* (Vol 1. pp. 492-502). De Boeck.

De l'usage des tests : aspects métrologiques, statistiques et interprétatifs

Marc Aguert

1. Introduction à l'utilisation des tests

1.1. Pourquoi utiliser des tests ?

Une personne entre dans votre bureau avec une histoire, une plainte, éventuellement un diagnostic. Pour intervenir et aider cette personne, il vous faut recueillir rapidement une information fiable, aussi objective que possible, sur le fonctionnement de cette personne, à la fois les aspects quantitatifs (le « rendement ») et les aspects qualitatifs (les stratégies). C'est à ce moment et dans cet objectif que les praticiens, orthophonistes ou psychologues, ont fréquemment recours à des tests. Les tâches impliquées par le test, à défaut d'être toujours très écologiques, obligent le patient à mobiliser des compétences, des habilités auxquelles vous n'auriez pas eu accès autrement. Au fond, recourir à un test psychométrique ou orthophonique revient à échantillonner des comportements de la personne.

Les comportements ainsi échantillonnés ont deux caractéristiques importantes qui justifient l'utilisation du test. D'une part, ils sont des comportements révélateurs d'aspects-clés du fonctionnement du sujet, des comportements qui permettent d'identifier des dysfonctionnements, produire des hypothèses sur les processus, tirer des conclusions et établir des pistes de travail. Le fait que les tâches impliquées par le test mobilisent bien les processus cognitifs d'intérêt est en principe soutenu par des arguments théoriques et empiriques ; cela renvoie à la question de la validité du test sur laquelle nous reviendrons en détail dans la section suivante. D'autre part, ils sont des comportements pour lesquels le praticien a une idée précise de ce qui est attendu pour les personnes dont le fonctionnement est « normal », c'est-à-dire ici le fonctionnement qui n'appelle pas une prise en charge ou un accompagnement. Cette information, chiffrée, sur les performances attendues de la part des personnes « normales » est collectée et mise à disposition par les concepteurs du test. On dit alors du test qu'il est normé (ou étalonné) car on peut se référer aux normes, i.e. aux performances observées chez les

personnes « normales », pour situer la performance de la personne que l'on reçoit et tenter d'objectiver la plainte émise. Le test permet donc non seulement de recueillir la performance du patient pour des comportements clés (par ex. : inhiber une information, lire des pseudo-mots, abstraire une règle, etc.), mais également de comparer cette performance à la performance « normale », celle qui précisément ne nécessite pas de prise en soin. En ce sens, la plupart des tests psychométriques peuvent être qualifiés de « comparatifs ».

Ce recours à un échantillon-étalon est essentiel et distingue les tests psychométriques des tests scolaires ou des tests que l'on trouve dans les magazines. La performance de la personne testée n'est pas jugée à l'aune de standards absolus ou personnels (par ex. : les attendus du correcteur) mais en fonction de la manière dont la personne se situe par rapport à sa population de référence. Avoir 6/20 à un test n'est pas une mauvaise note en soi, en particulier si la grande majorité des personnes de notre âge et de notre niveau socio-culturel a une note encore plus basse ! Ainsi, l'idée de la comparaison de la performance du patient à la population de référence qui a servi à établir les normes est au cœur de l'utilisation des tests. A ce titre, deux remarques peuvent être formulées.

D'abord, un test sans norme ou dont les normes ne sont pas adaptées à la personne testée perd en grande partie sa légitimité de test. Si la personne testée est d'un âge non couvert par les normes, si celles-ci sont trop datées, si elles ont été conçues auprès de personnes d'une autre culture que le patient, si elles ont été élaborées sur la base d'un échantillon trop petit ou peu représentatif, etc., la comparaison du patient à sa population de référence, et à sa suite, la conclusion du praticien, sera hasardeuse voire trompeuse. Bien sûr, il est toujours possible d'utiliser un test dans le seul but de comparer le patient à lui-même au fil du temps et ainsi objectiver, par exemple, des progrès. Mais cet objectif peut en réalité être atteint avec n'importe quelle tâche. Ensuite, même avec un test correctement normé, la comparaison pourrait s'avérer faillible si le praticien n'est pas attentif à la question de la standardisation. En effet, la performance du patient n'est comparable à celle de sa population de référence que si la tâche a été réalisée rigoureusement dans les mêmes conditions. Si le patient n'a pas bénéficié des mêmes consignes, du même étayage, d'un environnement de travail équivalent, la comparaison pourrait, là encore, être trompeuse. Le manque de standardisation va produire de l'erreur de mesure et réduire la fidélité du test, notion que nous développons maintenant.

1.2. Les trois qualités d'un bon test

Un test, comme le qualificatif « psychométrique » l'indique, est un outil de mesure et doit avoir à ce titre des qualités métrologiques (i.e. être un bon outil de mesure). Trois qualités sont très classiquement requises : la sensibilité, la fidélité et la validité.

La sensibilité. Imaginez que vous vous pesiez et que votre balance ne vous renvoie qu'une information binaire sur votre masse du type « gros vs. maigre ». Vous pourriez considérer que votre balance n'est pas très sensible ! De même, une balance électronique qui vous indique votre masse au gramme près est plus sensible qu'une balance à aiguille. De la même manière, un test est sensible s'il est capable de différencier des sujets dont les performances sont proches. Pour cela, il faut que le nombre de valeurs que peut prendre le résultat du test soit suffisamment important (un test classant les enfants en deux catégories « stupides » vs. « intelligents » serait aussi peu sensible que votre balance binaire) et que la difficulté de la tâche croisse très progressivement. Si tous les sujets testés, quelle que soit leur compétence propre, réussissent les 5 premiers items et échouent aux 5 derniers, le test n'est pas sensible. Il n'est pas discriminant et échoue à atteindre l'objectif recherché : situer la personne par rapport à sa

population de référence. Le choix des items constituant un test est un exercice notoirement délicat à cause de l'exigence de sensibilité.

La fidélité. Imaginez que vous vous pesiez le lundi matin, que votre balance vous indique 55 kg et que le mercredi matin de la même semaine, votre balance vous indique 82 Kg. Dans cette situation, vous auriez deux possibilités : soit acter que vous avez pris 27 Kg en 2 jours ; soit remettre en doute la fiabilité de votre outil de mesure, c'est-à-dire considérer que la mesure qu'il produit n'est pas suffisamment fidèle. Si tester une personne, c'est échantillonner ses comportements, il est important que tous les échantillons réalisés donnent la même image de la personne, que ce soit le lundi matin ou le mercredi matin ! Malheureusement, de nombreux facteurs, liés à la personne testée (stress, fatigue, disposition émotionnelle, etc.), à la situation de test (attitude du testeur, présence d'une tierce personne, bruit, température, etc.) ou à l'instrument de mesure lui-même (ambiguïté dans la formulation des consignes, des questions, dans la manière d'évaluer les réponses, etc.) vont générer de l'erreur de mesure et donc réduire la fidélité du test. En somme, la performance observée avec le test n'est que la réalisation circonstanciée de la performance « vraie », i.e. de la compétence du sujet. La performance observée est la performance vraie brouillée, biaisée par de l'erreur de mesure. Là encore, on comprend qu'un test trop peu fidèle nous empêche d'atteindre l'objectif fixé, à savoir situer la performance (vraie) du sujet par rapport à sa population de référence.

La fidélité d'un test se quantifie assez simplement en calculant la corrélation (la plus ou moins grande concordance) entre deux passations du test par la même personne. Il peut s'agir tantôt de la corrélation entre deux versions dites « parallèles » du test (deux versions en principe interchangeables) réalisées au même moment par la personne et alors on teste plutôt l'impact de l'erreur de mesure induite par les items sur la fidélité. Ou il peut s'agir tantôt de la corrélation entre deux passations du même test par la même personne à quelques semaines ou mois d'intervalle et on teste alors plutôt l'impact de l'erreur de mesure induite par les conditions de passation sur la fidélité. Dans les deux cas, on obtient un coefficient de fidélité noté r et qui s'interprète comme un coefficient de corrélation : plus le coefficient est proche de 1, plus la fidélité est bonne, plus le coefficient est proche de 0, moins la fidélité est bonne. Un coefficient de fidélité inférieur à 0,80 doit conduire le praticien à interpréter avec prudence les mesures effectuées avec le test.

La validité. Cette fois, imaginez simplement que vous vous pesiez avec tous vos vêtements trempés. Vous pensiez peser votre masse corporelle et en réalité, vous pesez votre masse corporelle… plus vos vêtements mouillés ! Le résultat est donc trompeur, la mesure n'est pas valide. Un test est valide s'il mesure bien le construit théorique qu'il prétend mesurer et non autre chose (ou ce qu'il prétend mesurer + autre chose). C'est parfois facile à établir (un test qui prétend mesurer la capacité à réaliser des additions et qui implique de réaliser des additions est certainement très valide) et parfois bien plus difficile (quand votre test prétend mesurer l'intelligence par exemple). Quand la validité dite « de contenu » ne saute pas aux yeux (comme dans le cas des additions, où le contenu du test correspond clairement avec le construit théorique que l'on souhaite mesurer), il est nécessaire d'établir des preuves de la validité du test. Deux chemins, non mutuellement exclusifs, sont le plus souvent empruntés par les concepteurs de tests : la validité de construction et la validité critériée.

La validité de construction vise, dans une démarche « bottom-up », à qualifier ce que mesure le test. On part de ce qui est mesuré et on cherche à le définir plutôt que l'inverse. Pour cela, on utilise des techniques statistiques comme l'analyse factorielle qui permettent de mettre en évidence la ou les dimensions latentes mesurées par le test. C'est ainsi qu'on a établi que la WISC-V, dont les tâches sont pourtant quasi-inchangées depuis des décennies, mesure

aujourd'hui 5 construits distincts[1]. Une autre technique, plus accessible, consiste à établir des corrélations avec d'autres tests… dont la validité est bien établie ! La WISC-V est très bien corrélée avec la WISC-IV qui, tout le monde s'accorde là-dessus, mesure l'intelligence. C'est donc la preuve que la WISC-V mesure aussi l'intelligence. Les coefficients de corrélation obtenus font alors office de coefficient de validité.

La validité critériée est basée sur un raisonnement plus pragmatique. Plutôt qu'établir des preuves que le contenu du test (les items, les tâches) mobilise bien un construit théorique un peu abstrait, on établit des preuves que le contenu du test est bien associé avec un critère externe objectif facilement mesurable. Par exemple, si le test vise à sélectionner les meilleurs candidats à l'entrée d'une école, on garantit la validité du test en croisant ses résultats avec les résultats des étudiants à leur sortie de l'école ! La corrélation entre les résultats au test et ceux à la sortie de l'école fait office de coefficient de validité. Autre exemple : on tient généralement pour acquis qu'il faut être intelligent pour réussir à l'école. La réussite scolaire est donc souvent retenue comme critère pour garantir la validité de tests d'intelligence : si l'enfant réussit le test et que par ailleurs, il réussit à l'école, c'est que l'école et le test mobilisent la même ressource, à savoir l'intelligence.

La validité d'un test, on le voit, peut être difficile à établir tant certains construits théoriques sont difficiles à définir et mouvants avec les années. Quand on demandait à Alfred Binet, le fondateur de la psychométrie moderne, ce qu'était l'intelligence, il répondait de manière facétieuse : « L'intelligence ? Mais c'est ce que mesure mon test ! » La boutade en dit long sur le problème que pose la validité. Assurer qu'un test mesure bien ce qu'il prétend mesurer est pourtant une étape essentielle, dite étape de validation. En effet si un enfant échoue à un test d'intelligence juste parce qu'il ne maîtrise pas bien la langue dans laquelle sont présentées les consignes, il serait tout à fait problématique de conclure qu'il présente une déficience intellectuelle. On notera au passage combien la validité est liée à la fidélité puisque par définition, un test peu fidèle qui génère beaucoup d'erreur de mesure ne mesure pas ce qu'il prétend mesurer et perd en validité.

En conclusion de cette section, l'intérêt des tests réside dans leur capacité à fournir rapidement au praticien une information objective sur la performance du patient en la situant par rapport à ce qui est attendu « normalement ». On retrouve cette idée clé dans la vieille mais toujours pertinente définition que René Zazzo (1969) donnait des tests psychométriques : « Un test correspond à une épreuve strictement définie dans ses conditions et dans son mode de notation qui permet de situer le sujet par rapport à une population elle-même bien définie biologiquement et socialement. » Attention cependant, cette visée comparative n'est pas l'objectif de tous les tests psychométriques. La troisième et dernière section de ce chapitre permettra d'aborder une famille de tests avec un objectif différent : les tests diagnostiques. Les concepteurs de tests doivent apporter des preuves que leur test est sensible, fidèle et valide, preuves auxquelles les utilisateurs des tests doivent être attentifs. Sans ces garanties, les conclusions tirées par le praticien seront probablement fallacieuses. Au-delà de la métrologie, les concepteurs de tests doivent également garantir, on l'a vu plus haut, que leur test est correctement normalisé (étape de normalisation). En effet, sans norme bien faite (actualisée et représentative d'une population de référence bien identifiée), même une mesure sensible-fidèle-valide pourra conduire à une conclusion erronée au moment de la comparer avec la population de référence. L'utilisation de tests pour recueillir une information rapide et objective sur les comportements du patient est

[1] Indice visuospatial, indice de raisonnement fluide, indice de vitesse de traitement, indice de mémoire de travail, indice de compréhension verbale

généralement utile et fructueuse… à condition de bien s'assurer au préalable que les conditions (sensibilité, fidélité, validité et qualité des normes) sont réunies, au risque que l'opération se révèle contre-productive !

2. Situer statistiquement la personne par rapport à sa population de référence

On l'a vu, le cœur de la démarche psychométrique est de positionner la performance d'une personne par rapport à sa population de référence. D'un point de vue statistique, chiffré, cela peut se faire selon deux méthodes distinctes qui ont chacune des avantages et des inconvénients propres que nous verrons successivement : la méthode des centiles[2] et la méthode des scores standards.

2.1. La méthode des centiles

Imaginons un test de lecture où le concepteur du test a mesuré la vitesse de lecture d'un texte chez 100 enfants de 7 ans représentatifs de la population des enfants de 7 ans[3]. L'enfant le plus lent lit le texte en 241 secondes tandis que l'enfant le plus rapide lit le même texte en 69 secondes (cf. figure 1). Imaginons que 160 secondes soit le temps de lecture tel que 50 enfants sur 100 lisent le texte plus lentement et 50 enfants sur 100 lisent le texte plus rapidement. Cent soixante secondes serait alors le temps de lecture correspondant au 50ème centile, c'est-à-dire le temps de lecture tel que 50 % des personnes de l'échantillon-étalon sont plus lents et 50 % sont plus rapides (cf. figure 1a). Avoir un temps de lecture proche du 50ème centile signifie qu'on lit à une vitesse « moyenne », au sens où il y a autant de gens plus rapides que nous que de gens plus lents. On appelle aussi ce 50ème centile la médiane de la distribution. Imaginons maintenant que 223 secondes soit le temps de lecture correspondant au 95ème centile. Cela signifie que 95 % des enfants de l'échantillon-étalon seraient plus rapides que 223 secondes pour lire le texte et seulement 5 % seraient plus lents (cf. figure 1b). Ainsi, avoir un temps de lecture proche du 95ème centile signifie qu'on lit très lentement. On ne lit pas très lentement parce que 223 secondes, c'est beaucoup de temps dans l'absolu. On lit très lentement parce que sur un texte identique, 95 % des personnes qui nous ressemblent (des enfants de 7 ans) font mieux. Exprimer la performance d'un patient, non pas avec un score brut souvent abstrait, mais par son rang sur l'échelle des centiles permet de le situer immédiatement par rapport à sa population de référence : l'objectif du test est atteint.

[2] L'anglicisme « percentile » est souvent utilisé en lieu et place du mot français « centile »

[3] Nous avons pris l'effectif N = 100 par commodité du raisonnement mais pour constituer de bonnes normes, davantage de participants serait souhaitable.

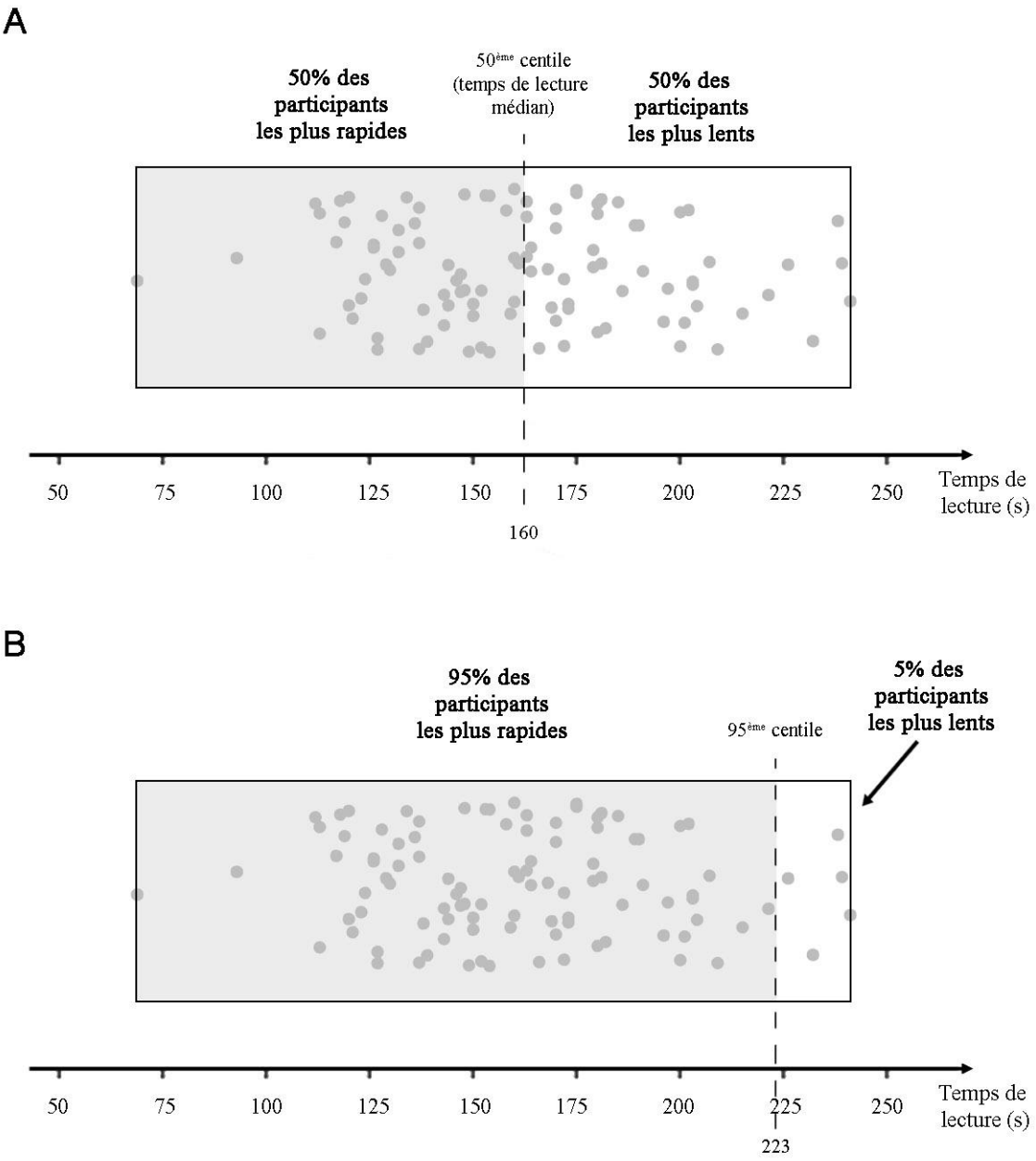


Figure 1 : Scores de 100 enfants de 7 ans à un test de vitesse de lecture. La « boîte » qui représente 100 % des scores est tantôt divisée en 2 paquets de 50 % des effectifs grâce au 50ème centile (panel A, en haut), tantôt en 2 paquets de 95 % et 5 % des effectifs respectivement, grâce au 95ème centile (panel B, en bas).

Imaginons un test de rétention de mots où la variable mesurée est le nombre de mots rappelés parmi une liste de 50 mots. Qu'importe que le 90ème centile corresponde à 23 mots où à 42 mots : avoir un score correspondant au 90ème centile exprime dans tous les cas que l'on a une excellente mémoire puisque seulement 10 % des gens rappellent plus de mots que nous quand 90 % en rappellent moins ! Dans un test, les scores bruts sont souvent peu informatifs ; ce qui importe, c'est le positionnement de la personne testée par rapport à sa population de référence, son rang, soit précisément ce qu'expriment les centiles.

L'utilisation des centiles présente un avantage majeur et trois inconvénients. Le premier inconvénient est que l'information sur les centiles n'est pas toujours mise à disposition par les concepteurs du test ! Il arrive que l'utilisateur du test ignore à quel centile correspondent tels ou tels scores bruts. Parfois les concepteurs du test ne communiquent que quelques centiles plus « significatifs » et utiles à l'interprétation : le 5ème centile, le 10ème, le 50ème, etc. Si les concepteurs du test faillissent à fournir cette information, il est bien sûr impossible de raisonner sur la performance du patient en termes de centiles. Le deuxième inconvénient est que les centiles ne sont plus des chiffres à proprement parler. En transformant le score brut d'un patient en son rang dans la distribution des scores dans l'ensemble de l'échantillon-étalon, on passe d'une échelle numérique à une échelle ordinale. La conséquence directe de cette transformation est l'impossibilité d'effectuer des opérations arithmétiques sur les centiles. Par exemple, si on dispose de trois scores exprimés en centiles pour un patient, il est impossible d'en faire la moyenne. Enfin, le troisième inconvénient est que les centiles sont une manière d'exprimer les performances des patients peu sensible aux extrêmes de la distribution (voir aussi la figure 2 pour une illustration). Au centre de la distribution, une seconde de temps de lecture supplémentaire suffit à passer du 50ème au 51ème centile ; à l'extrême droite de la distribution, il faut trois secondes pour passer du 95ème au 96ème centile. Ainsi aux extrêmes, des enfants avec des temps de lecture différents peuvent se voir attribuer le même rang. Finalement, l'avantage majeur des centiles et qu'ils peuvent s'utiliser, dès lors qu'on dispose de l'information, quelle que soit la forme de la distribution des scores bruts dans l'échantillon-étalon. On verra dans la section suivante que ce n'est pas le cas des scores standards. En cela, les centiles sont souvent une solution de repli pour les utilisateurs de scores standards dès lors que la distribution des scores dans l'échantillon-étalon ne suit pas une certaine distribution : la distribution normale.

2.2. La méthode des scores standards

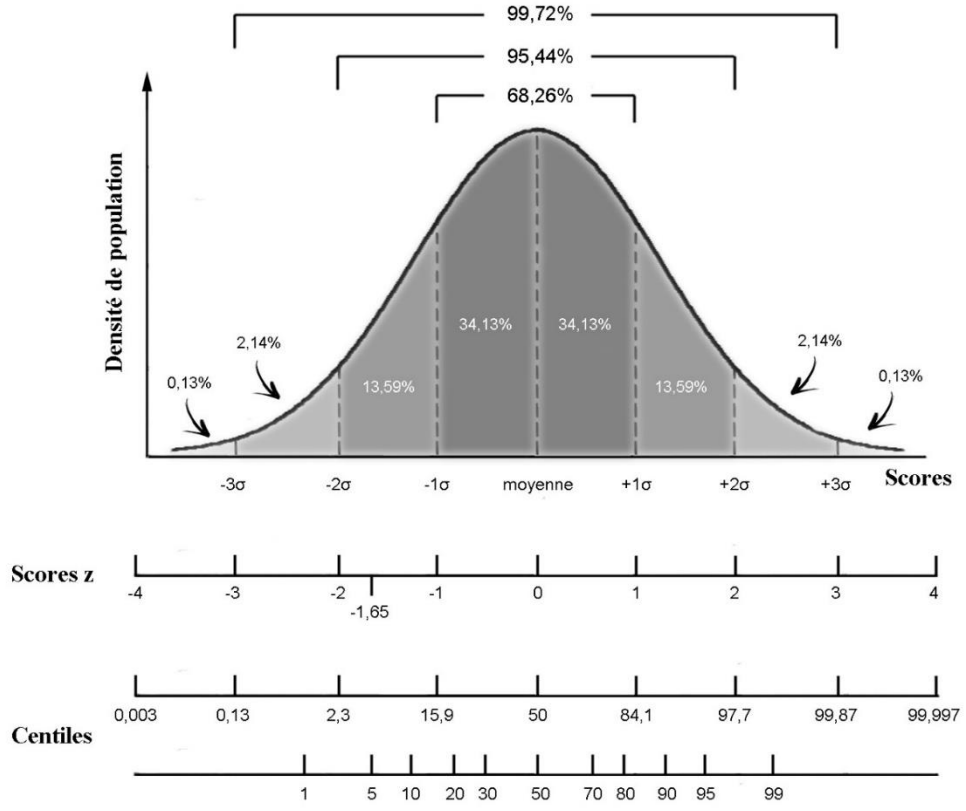
Comme la méthode des centiles, la méthode des scores standards suppose de transformer le score brut de la personne testée car ce dernier ne permet pas facilement de situer la personne par rapport à sa population de référence. Mais plutôt que de remplacer le score brut par son rang dans la distribution, le praticien va transformer le score brut en un score dit « standard » avec une formule mathématique (cf. infra). Les scores standards sont des scores dont la distribution a toujours la même moyenne et toujours le même écart-type[4]. 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 ($\mu = 0$; $\sigma = 1$)[5]. La connaissance qu'a le praticien de la moyenne et de l'écart-type de la distribution des scores standards lui facilite grandement l'interprétation de la performance du patient. Imaginons un enfant qui a 85 de QI. Connaissant la moyenne et l'écart-type de cette distribution de scores standards (même si le score brut vient initialement d'un test qui lui est tout à fait inconnu), le praticien peut immédiatement situer l'enfant par rapport à sa population de référence : la performance est inférieure à la moyenne et, plus exactement, se situe à un écart-type en dessous de la moyenne. Tous les scores standards sont équivalents et peuvent être convertis : 85 de QI correspond à un score T = 40 et à un score z = -1. Dans la suite de ce chapitre, par souci de simplicité, nous nous focaliserons sur les scores z dans la mesure où 1

[4] L'écart-type est un indicateur de dispersion des scores qui exprime l'écart moyen entre chaque score individuel et la moyenne de tous les scores.

[5] 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* ».

point sur l'échelle des scores z correspond précisément à 1 écart-type (1 σ) mais la logique est identique pour les autres scores standards.

Le lecteur peu habitué à l'utilisation des scores standards pourrait légitimement continuer à se poser la question : en quoi savoir que l'enfant de l'exemple ci-dessus a un score z = -1, situé donc à un écart-type sous la moyenne, nous informe-t-il précisément sur la manière de situer cet enfant par rapport à sa population de référence ? Les praticiens plus avertis savent que se situer à 1 écart-type de la moyenne, dans un sens ou dans l'autre, c'est être moyennement éloigné de la moyenne ; être à 2 écart-types de la moyenne c'est être éloigné de la moyenne ; enfin être à 3 écart-types de la moyenne ou plus, c'est être très éloigné de la moyenne. Mais d'où viennent ces ordres de grandeur ? Si l'on est très éloigné de la moyenne quand on se situe à 3 écart-types, quid quand on se situe à 7 ou 15 écart-types ? La réponse nous vient… des centiles ! En effet, quand la distribution des scores z suit une distribution normale, c'est-à-dire suit la loi de probabilité de Laplace-Gauss, il existe une correspondance entre ces scores z et les centiles. Ces correspondances sont rapportées dans la fameuse « table de probabilité de la loi normale » que l'on trouve à la fin de tous les bons manuels d'introduction aux statistiques. Cette table spécifie par exemple[6] qu'un score z = 1 correspond au 84ème centile. Cela signifie que 16 % seulement des personnes de la population de référence ont un score z supérieur à 1. Autre exemple : seulement 5 % des personnes de la population de référence ont un score z inférieur à -1,65. La figure 2 ci-dessous présente visuellement ces correspondances entre scores z et centiles.



[6] Dans sa version unilatérale, la plus commune

Figure 2 : Correspondances entre proportions de la population (représentées sous la courbe normale et sur les 2 axes de centiles) et distance à la moyenne en écarts-types (représentée sur l'axe des scores z).

Du fait des caractéristiques de la distribution normale la quasi-totalité de la population de référence (99,94 %) a un score compris entre z = -4 et z = 4. Donc plus on s'éloigne de la moyenne (z = 0), dans un sens ou l'autre, plus on a un score rare, atypique par rapport à la population de référence, dans les limites indiquées ci-dessus, z = 4 et z = -4 fonctionnant respectivement comme un plafond et un plancher qu'il est très improbable de dépasser. Dans ce cadre, mesurer chez un sujet un score z = -6 par exemple devra forcément éveiller les soupçons du praticien. En effet, ce score est si improbable que soit le sujet mesuré est effectivement un spécimen très très rare[7], soit c'est qu'il faut 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 (par ex. : une performance d'enfant comparée à des performances d'adultes).

L'utilisation des scores z présente plusieurs avantages et un inconvénient majeur. Parmi les avantages, ils sont simples à calculer dès lors que l'on dispose de la moyenne et de l'écart-type de la distribution des scores bruts. Il suffit d'appliquer la formule (1) ci-dessous où x est le score brut du patient, m_x la moyenne des scores bruts dans l'échantillon-étalon et s_x, l'écart-type des scores dans l'échantillon-étalon.

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

Il faut noter que cette opération arithmétique ne change pas la nature de l'échelle qui reste numérique. Cela revient, par analogie, à transformer des degrés Celsius en degrés Fahrenheit : la température reste inchangée mais elle est exprimée sur une échelle différente. L'échelle restant numérique, il reste possible de faire des opérations arithmétiques et on pourra par exemple moyenner plusieurs scores z. De manière encore plus intéressante, la nature numérique des scores z (et des scores standards en général) va rendre possible le calcul d'un intervalle de confiance pour estimer la performance « vraie » du patient sur la base de sa performance observée, qui tienne compte de l'erreur de mesure. En d'autres termes, il n'est par définition pas possible de connaître la performance « vraie » du patient mais on peut l'approcher en produisant, avec la technique des intervalles de confiance, un intervalle dans lequel elle sera par exemple dans 95 % des cas. Pour calculer l'intervalle de confiance (IC) à 95 % du score z d'un patient à un test, on utilise la formule (2) ci-dessous où r est le coefficient de fidélité du test utilisé[8].

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

En plus de l'information selon laquelle le score observé du patient est z = -1,56, le praticien aura l'information selon laquelle, avec 95 % de confiance par exemple, son score vrai est situé dans l'intervalle [-1,83 ; -1,08]. Il est souvent fortement préconisé d'utiliser ces intervalles de confiance car ils rappellent que la performance observée n'est pas la performance vraie et évitent que les personnes testées n'attachent trop d'importance au résultat chiffré du test. On remarquera que logiquement, plus le coefficient de fidélité r du test est petit, plus l'intervalle

[7] En réalité, la probabilité de tomber sur un individu avec un tel score est infime : approximativement une chance sur un milliard ! Donc il existe sur Terre environ 7 personnes comme celle-ci et vous seriez tombé dessus…

[8] 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. Le chiffre « 1,96 » garantit un IC à 95% de confiance mais un autre niveau de confiance est possible en modifiant ce chiffre. La valeur du coefficient de fidélité r est en principe indiquée dans le manuel du test.

de confiance à 95 % sera grand : plus il y a de l'erreur de mesure, plus l'estimation du score vrai du patient est imprécise.

Un autre avantage des scores z est qu'ils sont plus sensibles aux extrêmes de la distribution qu'en son centre (et accessoirement, plus sensibles aux extrêmes que les centiles, cf. figure 2), ce qui est généralement un intérêt pour le praticien, plus intéressé par les scores atypiques que les scores typiques. Finalement, soulignons aussi vigoureusement que possible l'inconvénient majeur des scores z : leur interprétation telle que décrite jusqu'à maintenant dépend d'une condition déjà mentionnée : que les scores dans la population de référence se distribuent en suivant une loi normale (i.e. une loi de Laplace-Gauss). Si ce n'est pas le cas, les correspondances entre scores z et centiles rapportées dans la table de probabilité de la loi normale ne sont plus valables et les conclusions du praticien seraient erronées. En somme, l'utilisation des scores z nécessite de connaître, non pas deux caractéristiques statistiques des scores bruts dans la population de référence, mais trois : leur moyenne, leur écart-type et également la forme de la distribution. Cette information sur la forme de la distribution, normale ou pas, n'apparaît malheureusement que très rarement dans les manuels des tests. Elle est pourtant une condition sine qua non et est loin d'être évidente. Il est notamment assez improbable que des variables comme des temps de réaction ou des nombres d'erreurs, variables assez communément mesurées dans les tests, se distribuent en suivant une loi normale (voir Aguert & Capel, 2018, pour une illustration).

2.3. De l'utilisation d'un score seuil

Pour des raisons historiques et pratiques, la psychométrie mesure souvent les construits qu'elle approche de manière dimensionnelle, avec des échelles. Ces échelles permettent de positionner le sujet testé à un endroit donné de la distribution des scores de l'échantillon-étalon, par la méthode des centiles ou par le calcul d'un score standard, dans une visée comparative.

La littérature contemporaine montre que cette approche dimensionnelle présente de bons résultats du point de vue clinique, pour formuler des hypothèses de remédiation par exemple ou pour communiquer avec les patients (Bornstein & Natoli, 2019). Et pourtant, à la fin du processus, les praticiens sont souvent tentés de rompre avec cette conception dimensionnelle du construit mesuré, riche en information, pour basculer sur une conception catégorielle consistant à distinguer de manière binaire les personnes qui seraient « normales » à l'aune du construit mesuré et les personnes qui seraient « atypiques » ou « déficitaires ». Par analogie, cela revient à mesurer précisément la masse d'un échantillon de personnes avec un outil sensible, fidèle, valide pour finalement classer un peu grossièrement ces personnes en deux catégories : les obèses (mettons ceux dont la masse dépasse le seuil de 90 Kg) et les « poids normaux » (ceux dont la masse reste inférieure au seuil de 90 Kg). Même si cette opération constitue un appauvrissement de l'information quant à la performance de la personne, elle peut s'avérer utile dans des contextes de prise de décision pour donner suite à la relation de soin. La difficulté de l'opération consiste à savoir quel score seuil (« cut-off score » en anglais) utiliser pour dichotomiser la performance entre les personnes « normales » et les personnes « déficitaires » et quel sens accorder à ce score seuil.

Sur la question du choix du seuil, le consensus actuel est de considérer qu'un score est atypique quand il se situe parmi les 5 % les plus rares de la distribution statistique[9]. Il faut insister sur le

[9] Ce qui correspond, pour une distribution de scores normale (gaussienne) et dans une perspective bilatérale, au fait de se situer à plus de 2 écart-types de la moyenne ($z = \pm 2$), cf. figure 2. La question, importante, du choix

caractère arbitraire de ce consensus autour de la proportion 5 % qui n'a comme légitimité que le crédit que lui donnent les professionnels du domaine. Il est par ailleurs susceptible de varier, d'une époque à l'autre et d'un construit à l'autre. Ainsi, on trouve fréquemment dans les pays anglophones des tests qui utilisent comme score seuil le 10ème, voire le 16ème centile (le fait de se trouver à un écart-type de la moyenne de la population générale). En dernière analyse, même si des préconisations existent (par ex. : Colombo et al., 2016) le choix du score seuil relève de la responsabilité du praticien qui doit pouvoir défendre ses choix d'outils et ses interprétations. Sur la question du sens à accorder à ce seuil, la réserve est là aussi de mise. D'abord, soulignons que tout ce qui est rare n'est pas forcément préoccupant pour le praticien. Doit-on s'inquiéter du fait qu'un enfant dispose d'un stock de vocabulaire à +2 écart types de la moyenne des enfants de son âge ? Ensuite, comme indiqué plus haut, la psychométrie mesure généralement les construits de manière dimensionnelle et il convient de ne pas réifier des catégories de personnes juste sur la base d'une distribution de scores dichotomisée. Comme le soulignent Guilmette et al. (2020), un score inférieur à 2 écart-types de la moyenne du groupe de référence doit être qualifié de score « exceptionnellement bas » mais pas de score « déficitaire ». Les auteurs indiquent « The labels do not convey impairment or other evaluative judgments ; scores in isolation cannot be impaired or deficient » (p. 445). Il existe une exception à cette règle, c'est le cas des tests diagnostiques que nous abordons dans la section suivante.

En conclusion de cette section, le praticien recourt à des tests pour situer, pour une habileté donnée, la performance de son patient par rapport à sa population de référence. Cela se réalise statistiquement par deux méthodes distinctes (quoique fondées toutes les deux sur le même principe) : la méthode des centiles et la méthode des scores standards (par ex. : les scores z). Cette dernière méthode est plus populaire car elle laisse la possibilité d'effectuer des calculs arithmétiques sur les scores et notamment l'estimation du score vrai par intervalle de confiance. Mais elle souffre d'une limite trop souvent négligée : les scores dans l'échantillon-étalon doivent se distribuer, au moins approximativement, selon une loi normale. La méthode des centiles est une alternative parfaitement satisfaisante lorsque cette condition n'est pas remplie. Quelle que soit la méthode utilisée et dès lors que le test a les qualités métrologiques requises, le praticien dispose maintenant d'une information objective sur la performance du patient pour une habileté donnée : dans la norme, (très) supérieure à la norme ou (très) inférieure à la norme. Cette information chiffrée n'a pas forcément beaucoup de sens « en soi » et il convient de l'interpréter dans le contexte de la passation, de la recouper avec des informations d'autre nature (observation, entretien, etc.). Une manière prisée des praticiens de donner du sens à la performance du patient est de la confronter à un seuil censé distinguer la performance « normale » de la performance « déficitaire ». Cette pratique est critiquée car elle conduit à envisager la performance de manière binaire, jusqu'à parfois créer des « pathologies » ex nihilo.

3. Le cas particulier des tests diagnostiques

Les tests diagnostiques forment une famille de tests avec un objectif spécifique : détecter une pathologie chez la personne testée, pathologie psychologique, psychiatrique ou neurologique, définie indépendamment du test. Il peut s'agir de la dyslexie, l'autisme, la dépression, etc. Ces tests ont un fonctionnement particulier que nous décrivons dans cette section.

---

d'une perspective unilatérale ou bilatérale est délibérément esquivée dans ce chapitre. Faute de place et par souci de lisibilité, nous nous sommes restreints dans le raisonnement et les exemples, à une perspective unilatérale. Cela implique d'exclure a priori, sur la base d'hypothèses théoriques ou cliniques, que le sujet testé puisse être positionné sur l'un des deux versants de la distribution des scores pour se focaliser sur le versant pertinent.

3.1. Un raisonnement par test d'hypothèses

Les tests diagnostiques nous viennent du modèle médical de la maladie dans lequel il y a fondamentalement deux catégories de personnes : les personnes en bonne santé, « saines », et les personnes malades, « pathologiques ». Tout comme un test pour la Covid-19 par exemple, le test diagnostique doit permettre de dire si la personne est malade (positive à la pathologie) ou pas (négative à la pathologie). Dans le domaine qui nous occupe, comme il n'existe pas de marqueurs biologiques fiables de la dyslexie, l'autisme ou la dépression, le test peut produire des erreurs de catégorisations. La première erreur consiste à conclure à la suite du test que la personne est positive à la pathologie alors que ce n'est pas le cas, il s'agit donc d'un faux-positif. Dans le domaine de la santé, un test qui produit peu de faux-positifs est dit « spécifique » : il identifie bien les personnes positives à la pathologie en évitant les personnes négatives. La deuxième erreur possible consiste à conclure à la suite du test que la personne est négative à la pathologie alors que ce n'est pas le cas, il s'agit donc d'un faux-négatif. Un test qui produit peu de faux-négatifs est dit « sensible » : il identifie bien les personnes positives dès lors qu'elles sont effectivement positives. Il est important de noter que la sensibilité d'un test diagnostique n'est pas la même que celle d'un test psychométrique classique (cf. section 1.2.) ! La sensibilité diagnostique (ou statistique) n'a de sens que dans une perspective catégorielle (puisqu'il s'agit de la capacité du test à limiter une erreur de catégorisation) alors que la sensibilité psychométrique (définie comme la capacité du test à discriminer des sujets ayant des performances proches) n'a de sens que dans une perspective dimensionnelle. La figure 3 schématise la différence entre une perspective psychométrique classique, dimensionnelle (panel A), où l'ensemble des personnes de la population se distribue sur un continuum de performances allant de performances atypiquement faibles à des performances typiques, et une perspective diagnostique, catégorielle (panel B), où la population totale est divisée entre les personnes saines et les personnes pathologiques.

Pour atteindre l'objectif de classer correctement la personne testée dans l'une des deux catégories (saine vs. pathologique) en limitant les erreurs de catégorisation, les tests diagnostiques utilisent le même raisonnement que les tests d'hypothèses en statistiques. Il s'agit d'abord de poser une hypothèse dite « nulle » selon laquelle la personne testée est saine puis dans un deuxième temps, de rejeter cette hypothèse si la performance de la personne testée est improbable au regard de cette hypothèse nulle. On accepte alors l'hypothèse alternative : la personne n'étant pas saine, elle est pathologique, c'est-à-dire porteuse d'une maladie, un trouble, une lésion cérébrale, etc. Concrètement, cela implique de construire un échantillon-étalon particulier en ce qu'il est uniquement composé de sujets sains, et non de sujets représentant l'ensemble de la population de référence des sujets testés. Le recrutement de l'échantillon-étalon implique donc tout un éventail de critères d'exclusion de manière à garantir que les normes sont représentatives de la catégorie « personnes saines », évitant les personnes dont les performances pourraient être dégradées 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 étalon.

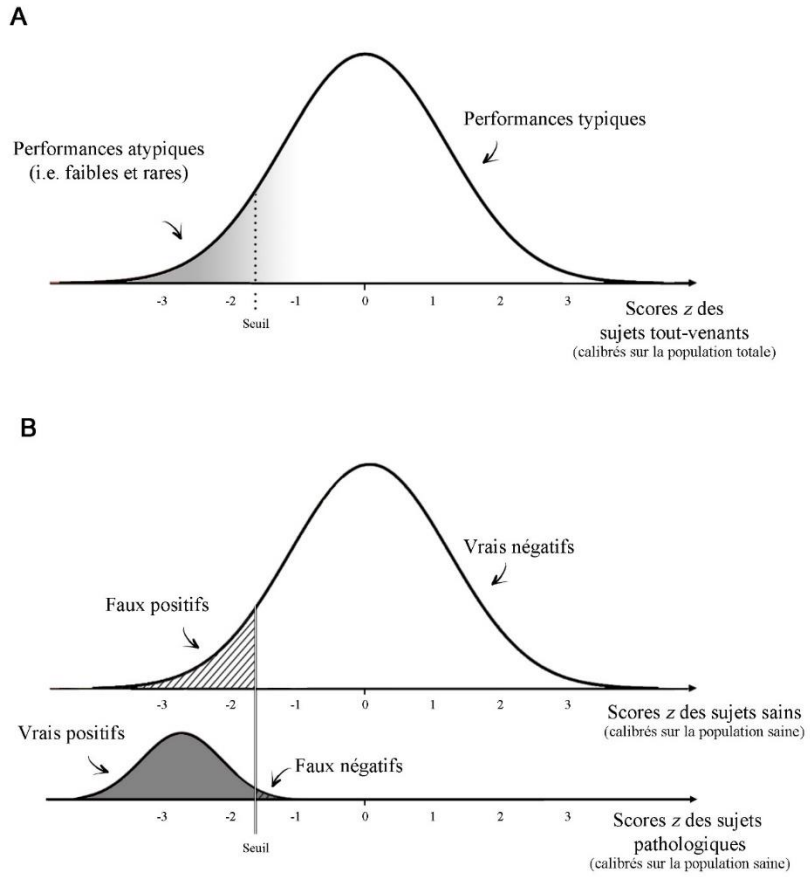


Figure 3 : Scores z de la population totale conceptualisés respectivement pour un test comparatif (panel A) et pour un test diagnostique (panel B). Panel A : l'ensemble des personnes de la population se distribue sur un continuum de performances allant de performances atypiques (en gris, performances à la fois rares et faibles) à des performances typiques (en blanc). Le dégradé de couleur indique que le passage des performances typiques aux performances atypiques est progressif. Panel B : la population totale est divisée entre les personnes saines (en blanc, distribution du haut) et les personnes pathologiques (en gris, distribution du bas). 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 : Dans le panel B, les performances des sujets pathologiques ont été représentées suivant une distribution normale (de moyenne $\mu \approx -2,8$) pour des raisons pratiques mais d'autres formes et d'autres positions de la distribution sont parfaitement possibles.

Une fois obtenue une estimation fiable de ce que sont les performances de la catégorie des personnes saines, le praticien sera en mesure de déterminer si la performance du sujet testé est très improbable, sous l'hypothèse qu'il est sain. En particulier, si sa performance est très faible par rapport à la moyenne des sujets sains constituant l'échantillon-étalon, il est tentant de conclure qu'il est pathologique. Mais certains sujets sains peuvent aussi avoir des performances très faibles au test. Comment contrôler qu'on ne produit pas un faux-positif en concluant que le sujet testé est pathologique (alors qu'il est sain) ? C'est là qu'intervient le score seuil qui va avoir un rôle radicalement différent d'avec les tests comparatifs décrits plus haut. Dans les tests diagnostiques, le score seuil sert à contrôler le risque de faire des erreurs de catégorisation, en particulier des faux-positifs. Le praticien va d'abord poser la probabilité maximale de commettre un faux-positif qu'il est prêt à assumer. En statistiques, cette probabilité est appelée

« seuil α » et est généralement fixée à 5 %[10] : sur 100 personnes saines, en moyenne 5 seront identifiées à tort comme pathologiques. Si l'on regarde la distribution des scores des personnes saines (cf. figure 3b), on voit que le score z permettant de délimiter ces 5 % de sujets sains dont les scores sont les plus faibles et qui sont donc susceptibles d'être considérés comme des sujets pathologiques (conduisant ainsi à un faux-positif) est z = -1,65 (cf. table de probabilités de la loi normale et figure 2). Ce sera donc le score seuil retenu par le praticien. En effet, il garantit que jamais plus de 5 % des sujets sains seront catégorisés comme pathologiques puisque pour toutes les personnes dont le score est supérieur à ce seuil, le praticien se gardera de conclure que la personne est pathologique. En fait, un score supérieur à ce seuil de -1,65 pourra aussi bien être obtenu par des sujets pathologiques ou des sujets sains (cf. figure 3b). De la même manière, si un sujet obtient un score observé inférieur à ce seuil de -1,65, il pourra aussi bien être pathologique ou sain. Mais dans ce dernier cas, le praticien sait que la probabilité que ce sujet est sain est de 5 chances sur 100 au maximum. Cette probabilité étant faible, il va assumer de rejeter l'hypothèse nulle que la personne testée est saine et conclure qu'elle appartient à la catégorie des personnes « pathologiques » (sans perdre de vue qu'il commettra un faux-positif pour 5 personnes saines sur 100). Le lecteur qui souhaite reprendre ce raisonnement de manière plus détaillée peut se référer à Aguert et al. (en révision).

3.2. Interpréter un test diagnostique

Trop souvent les praticiens ignorent la différence entre tests comparatifs et tests diagnostiques et les utilisent de manière identique. Après tout, « dans les deux cas, on teste le sujet et si celui-ci a un score z inférieur à z = -1,65, on le prend en charge car il est pathologique ou déficitaire ». Pourtant, ces deux familles de tests s'interprètent de manière très différente. Les tests diagnostiques ne visent pas à positionner la performance du sujet par rapport à sa population de référence. Au contraire, il s'agit de comparer le sujet à une population (les personnes saines) pour finalement conclure qu'il n'appartient pas à cette population ! Ceci acté, il n'est plus question d'utiliser le score z du sujet pour qualifier sa performance puisque celui-ci a été calculé sur la base de normes qui ne représentent pas la population de référence du sujet ! Ce score z n'aura servi qu'au test d'hypothèses et il est erroné de l'interpréter dans une perspective psychométrique. Avec un test diagnostique, que le sujet testé ait z = -6 ou z = -2, la conclusion du test est identique et doit se limiter à : rejet de l'hypothèse nulle que le sujet est sain, acceptation de l'hypothèse alternative que le sujet est pathologique. Puisque l'échantillon-étalon d'un test diagnostique est constitué de sujets sains, il est inadéquat pour décrire la performance d'un sujet pathologique. Le contenu du test lui-même, la ou les tâches demandées, sont peut-être adaptées pour les personnes saines mais pas forcément pour les personnes pathologiques (manque de sensibilité, au sens psychométrique du terme). Pour positionner la performance du sujet avec pathologie sur une dimension donnée, il faut la comparer à un échantillon-étalon pertinent, soit représentatif de la population générale (i.e., incluant des personnes avec la pathologie, à la hauteur de la prévalence de la pathologie en question), soit uniquement constitué de personnes avec la pathologie[11].

Il faut souligner que les tests diagnostiques ne sont pertinents que lorsque la pathologie qu'ils sont censés détecter est clairement définie indépendamment du test lui-même. Si la pathologie à détecter est définie de manière ad hoc par le fait d'avoir un score z < -1,65 au test, ce dernier

[10] Il faut souligner que cette probabilité est tout aussi arbitraire que celle déterminant la rareté dans l'approche psychométrique classique (cf. section 2.3). Elle est susceptible de varier d'une époque à l'autre et d'un domaine à l'autre. En physique des particules par exemple, le seuil α communément utilisé est non pas 5% mais 0,00003% (soit z = ±5) !

[11] Ces deux comparaisons sont possibles et valides mais répondent évidemment à des questions différentes.

n'est plus « diagnostique » au sens strict de la reconnaissance d'une maladie ou d'une condition d'après ses manifestations observables. La définition et la détection de la pathologie se faisant dans un même mouvement, c'est la démarche même de diagnostic qui est rendue caduque. Bien connaître la population pathologique permet de constituer des groupes de personnes pathologiques qui pourront non seulement servir à étalonner des tests à destination de ces personnes mais également à établir la sensibilité et la spécificité du test diagnostique. L'utilisation du score seuil z = -1,65 assure seulement à l'utilisateur du test qu'il ne fera pas plus de 5 % de faux-positifs (spécificité) et n'apporte aucune garantie sur le risque de faire des faux-négatifs (sensibilité diagnostique) ! Bien connaître la sensibilité et la spécificité du test permet d'établir le score seuil qui donne le meilleur rapport sensibilité / spécificité, en fonction des objectifs des concepteurs du test. Ce n'est pas forcément z = -1,65 (voir Aguert et al., en révision, pour une discussion plus détaillée de cet aspect). Outre l'interprétation du score z du sujet testé, on voit que le rôle du score seuil n'est pas du tout le même pour les tests comparatifs et les tests diagnostiques. Pour les premiers, le seuil est un simple repère pour souligner la rareté du score du sujet, son caractère atypique, et prendre des décisions. C'est un repère facultatif et souvent critiqué pour son caractère binaire. Pour les seconds, il s'agit d'un ingrédient tout à fait indispensable du raisonnement : pour prendre une décision sur la catégorie d'appartenance du sujet dans un contexte d'incertitude, il faut contrôler le risque de faire des faux-positifs et le score seuil établit ce niveau de risque. A ce titre, il est utile de rappeler que remonter le seuil d'un test diagnostique, par exemple à z = -1 ne va pas permettre d'identifier des personnes « moins pathologique », « borderline » où à « trouble léger ». Le test se borne à classifier des personnes dans deux catégories qui existent indépendamment de lui et remonter le seuil va juste augmenter le risque de faire des faux-positifs. De la même manière, un sujet avec un score z juste au-dessus du seuil n'est pas « presque pathologique » tout simplement parce que le score seuil n'indique pas que l'on est au seuil de la pathologie : il quantifie le risque de qualifier une personne saine de « pathologique ». Derrière la distinction entre tests comparatifs et tests diagnostiques se niche un débat qui dépasse largement le cadre du présent chapitre : la manière dont le concepteur du test ou son utilisateur conceptualise le passage entre le normal et le pathologique. Deux grandes positions structurent ce débat : la position dimensionnelle selon laquelle normalité et pathologie sont des différences de degrés sur une dimension donnée (cf. figure 3, panel A) et la position catégorielle selon laquelle normalité et pathologie caractérisent des fonctionnements structurellement différents (cf. figure 3, panel B). La logique des tests comparatifs est davantage compatible avec la position dimensionnelle tandis que la logique des tests diagnostiques est clairement plus compatible avec la position catégorielle (Aguert et al., en révision).

En conclusion de cette section, interpréter un test diagnostique est simple : si la personne testée a une performance sous le score seuil quantifiant la probabilité de faire un faux-positif (généralement 5 %), alors le praticien peut rejeter l'hypothèse que cette personne est saine et considérer qu'elle fait partie des personnes avec la pathologie. La difficulté réside plutôt dans le fait de résister à la tentation d'interpréter davantage cette performance ! Un test diagnostique n'est pas un test comparatif et ne permet pas de situer une personne par rapport à sa population de référence pour la simple et bonne raison que les personnes « saines » de l'échantillon-étalon ne sont pas la population de référence du patient. Il ne dit rien des caractéristiques des sujets pathologiques, ni au niveau individuel, ni au niveau de la population (quantité de personnes concernées, performances moyennes, etc.) Une autre difficulté est d'être capable de bien discerner les tests diagnostiques des tests comparatifs car un certain flou règne. En principe, un test diagnostique est construit spécialement en référence à une pathologie qu'il est censé détecter, et a un échantillon-étalon constitué de personnes dites « saines ».

Conclusion

Un test psychométrique ou orthophonique sert avant tout à comparer la performance de la personne testée à la performance type, attendue, dans la population de référence. Cette comparaison peut permettre d'objectiver, le cas échéant, que la personne nécessite un accompagnement car sa performance, pour une habileté critique dans son fonctionnement, une habileté bien définie, diffère d'une performance typique, normale (au sens de la norme statistique). Cette comparaison peut aussi servir à identifier des points forts du fonctionnement de la personne, qui serviront de points d'appui dans la prise en soin. Une exception à cet usage dominant des tests est la réalisation de diagnostics. Les tests diagnostiques reposent sur une logique interprétative différente : il n'est pas question de situer la personne par rapport à sa population de référence mais de détecter, grâce à la performance du sujet, la présence éventuelle d'une pathologie, définie indépendamment du test lui-même.

Que le test soit comparatif ou diagnostique, une utilisation adéquate et optimale suppose d'être vigilant à de multiples endroits. Concluons sur les principaux points de vigilance :

Le praticien doit avoir une idée précise de ce qu'il mesure avec le test et pourquoi il le mesure : la mesure doit être sensible, valide et fidèle. Il doit être conscient que la mesure entraine inévitablement avec elle de l'erreur de mesure qui l'éloigne de la performance « vraie » du sujet. Il peut être attentif à réduire cette erreur de mesure en standardisant au maximum la passation et utiliser des intervalles de confiance qui relativisent l'exactitude de la performance observée.

De manière cruciale, la démarche psychométrique implique une comparaison. Cela suppose une « démarche qualité » non seulement du côté de la mesure de la performance du sujet mais également du côté de la constitution des normes du test ! La démarche psychométrique n'est valide que si l'on s'assure de ne pas comparer des choux et des carottes. Ainsi, le praticien doit disposer d'informations sûres et complètes sur l'échantillon de normalisation du test : taille de l'échantillon, date de constitution, statistiques sur les données recueillies (forme de la distribution, valeurs des centiles, etc.), critère d'inclusion et d'exclusion des participants afin de savoir si les normes sont constituées de personnes « saines » (tests diagnostiques) ou de personnes tout-venantes (tests comparatifs), etc. Certaines de ces informations font malheureusement souvent défaut.

Si le praticien appuie sa démarche sur des scores seuils, il ne doit pas oublier que leur interprétation est très différente selon qu'il recourt à un test comparatif ou à un test diagnostique. Pour ces derniers, le seuil est indispensable en ce qu'il permet de prendre une décision sur la catégorie à laquelle appartient le sujet en contrôlant le risque de faux-positifs. Pour des tests comparatifs en revanche, situer le sujet au-dessus ou au-dessous d'un seuil n'a rien d'obligatoire, cela constituerait même plutôt une traduction simpliste de l'information qualitative fournie par le test.

Le praticien doit toujours avoir en tête que les performances qu'il mesure ne sont pas le reflet objectif de la réalité. Le score brut est le produit d'une mesure par un outil de mesure qui peut ne pas avoir les qualités métrologiques requises. Quant au score transformé en centile ou en score standard, il peut être doublement compromis : parce que la mesure initiale est de mauvaise qualité et/ou parce que l'opération de transformation est erronée (parce que la distribution des scores bruts dans l'échantillon-étalon ne suit pas une loi normale, parce que les normes sont trop anciennes, etc.) ! Le praticien doit être absolument attentif et toute donnée aberrante ou

même suspecte (par ex. : un score z = -6, cf. supra) doit l'amener à remettre en question la validité de la mesure.

Finalement, le praticien doit avoir en tête que la démarche psychométrique est largement basée sur des construits. Construits théoriques comme les habiletés que l'on mesure, les dysfonctionnements que l'on traque ou plus largement la conception, plutôt dimensionnelle ou plutôt catégorielle, que l'on se fait de la distinction entre normalité et pathologie. Ces construits, qui sont par définition mouvants, pour partie arbitraire et généralement peu interrogés, impactent massivement nos pratiques et nos conclusions. Construits méthodologiques également comme les centiles, les scores standards, les intervalles de confiance, les scores seuils qui ont un impact tout aussi important. L'absence de justification autre que « c'est le consensus actuel » à l'utilisation répandue du score seuil z = -1,65 en est une illustration. Se poser la question du sens de notre démarche, du sens et de la validité de nos outils est un impératif : la qualité de nos prises en soin en dépend.

Déclaration de conflits d'intérêts : l'auteur déclare n'avoir aucun conflit d'intérêts avec le contenu de cet article.

Note de l'auteur : Ce chapitre contient peu de références bibliographiques car il est surtout le produit de mes 10 années à enseigner la psychométrie à l'Université de Caen Normandie. Les informations rapportées dans ce chapitre qui ne sont pas l'objet d'une référence bibliographique spécifique peuvent très probablement être retrouvées dans la plupart des bons ouvrages sur la psychométrie donc voici quelques exemples ci-dessous.

Bibliographie

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

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

Anastasi, A. (2005). *Introduction à la psychométrie*. Editeur Guérin.

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.

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

Guillevic, C. & Vautier, S. (1998). *Diagnostic et tests psychologiques*. Paris : Nathan.

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.

Laveault, D. & Grégoire, J. (2002). *Introduction aux théories des tests.* Bruxelles : De Boeck

Roussel, M., & Godefroy, O. (2008). La batterie GREFEX : Données normatives. In O. Godefroy & GREFEX (Éds.), *Fonctions exécutives et pathologies neurologiques et psychiatriques* (p. 231–52). Solal.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of Neuropsychological Tests : Administration, Norms, and Commentary* (3e édition). Oxford University Press.

Zazzo, R. (1969). *Manuel pour l'examen psychologique de l'enfant*. Delachaux et Niestlé.