La rigueur clinique face aux limites psychométriques des tests diagnostiques en lecture

par
Julie Cattini,
Logopède, Luxembourg-ville, Luxembourg
Collaboratrice, RUCHE, Université de Liège, Liège, Belgique

Marion Balla,
Doctorante, Physiology of Cognition Lab, GIGA Institute, Université de Liège, Liège, Belgique

&
Guillaume Duboisdindien,
Maître de conférences, INSERM UMR 1322 LINC, Université Marie et Louis Pasteur, Besançon, France
Collaborateur scientifique, Département de Psychologie, Neuropsychologie de l'adulte, Université de Liège, Liège, Belgique

Dans le numéro 4 de l'UPLF-Info de 2025, un article présentait le consensus international sur la dyslexie (Carroll et al., 2025 ; Holden et al., 2025). Ce travail de concertation proposait plusieurs recommandations relatives à l'évaluation diagnostique, en insistant sur la nécessité de croiser différentes sources d'information (APA, 2023) :

- une synthèse clinique des antécédents médicaux, familiaux, développementaux et pédagogiques ;
- une analyse des relevés de notes et des rapports scolaires ;
- une évaluation normée ou critériée, logopédique et/ou psychoéducative.

Dans le diagnostic de la dyslexie, l'évaluation logopédique vise à objectiver des performances faibles en littératie, en décalage avec l'âge, le niveau scolaire attendu et/ou les autres compétences cognitives. La démarche diagnostique ne nécessite pas d'en rechercher la cause, mais bien de comprendre la persistance et l'impact fonctionnel des déficits objectivés sur le quotidien de la personne.

Si les tests normés offrent un cadre d'évaluation rigoureux et objectif, c'est uniquement à condition qu'ils soient sélectionnés, administrés et interprétés de manière appropriée (Holden et al., 2025). Dans la pratique clinique, cela implique de rester attentif à la qualité psychométrique des outils disponibles et à leur adéquation avec le profil du patient. L'évaluation fondée sur les preuves (*Evidence-Based Assessment*) met l'accent sur l'utilisation des données actuelles de la recherche pour guider le choix des cibles, des méthodes et des outils d'évaluation, ainsi que le processus d'interprétation. Dans cet article, nous proposons quelques pistes de réflexion et des précautions à adopter afin de tirer le meilleur parti des tests disponibles, même lorsque ceux-ci présentent des limites.

Caractéristiques psychométriques essentielles et importantes pour les outils diagnostiques

Lors de la sélection d'un outil d'évaluation, on ne doit pas s'appuyer uniquement sur son apparence, sa réputation ou sa date de publication. Ce qui fait la qualité d'un outil, ce sont ses caractéristiques psychométriques : autrement dit, tout ce qui permet de dire qu'un test mesure bien ce qu'il est censé mesurer et qu'il le fait de manière fiable.

Deux grandes questions guident l'analyse psychométrique :

1. le test mesure-t-il réellement ce qu'il prétend mesurer ? Il s'agit des preuves de validité ;
2. donne-t-il des résultats stables et cohérents ? Il s'agit des preuves de fidélité.

Pour qu'un test soit utile dans une démarche diagnostique, certaines caractéristiques sont essentielles (Burnay et al., 2024). Elles peuvent sembler nombreuses, mais chacune d'elles répond à une question très concrète de notre pratique.

- Validité théorique et de contenu
  ▶ *Les items couvrent-ils vraiment le domaine que je veux évaluer ?*
- Validité convergente, divergente et concourante
  ▶ *Ce test se comporte-t-il comme on s'y attend lorsqu'on le compare à d'autres outils ?*
- Pouvoir discriminant (sensibilité et spécificité)
  ▶ *Le test repère-t-il correctement les enfants dyslexiques ? Et évite-t-il d'en identifier à tort ?*
- Fidélité test-retest et fidélité interjuge
  ▶ *Obtiendrait-on le même résultat si l'enfant repassait l'épreuve ? Et si c'était un autre clinicien qui la proposait ?*
- Consistance interne
  ▶ *Les items vont-ils dans le même sens ? Mesurent-ils bien la même compétence ?*
- Distribution des scores
  ▶ *Les étalonnages sont-ils interprétables ? Les auteurs fournissent-ils des percentiles ou des normes utiles ?*
- Erreur standard de mesure
  ▶ *Quelle est la marge d'incertitude autour du score observé ?*

Enfin, il est toujours appréciable qu'un outil diagnostique documente aussi sa validité prédictive (est-ce que les scores annoncent bien des difficultés futures ?) et la structure interne (les items se regroupent-ils en dimensions cohérentes ?). Apporter des preuves de validité et de fidélité est primordial pour s'assurer que les données collectées avec l'outil sont fiables. L'ensemble de ces éléments constitue un socle indispensable pour garantir une interprétation rigoureuse des performances observées.

Les logopèdes sont familiers des outils normés, qui permettent de situer les performances de leurs patients par rapport à un échantillon représentatif. Toutefois, les études montrent que cette pratique s'accompagne souvent d'une expertise limitée dans le choix des outils les plus pertinents et, plus largement, dans la compréhension des principes psychométriques (Betz et al., 2013 ; Cattini & Willems, 2024 ; Ogiela & Montzka, 2021).

Pour aider à clarifier ces enjeux, la section suivante dresse un panorama des preuves psychométriques disponibles concernant les outils d'évaluation de la lecture.

Caractéristiques psychométriques des outils évaluant les compétences en lecture dans une visée diagnostique

Dans la mesure du possible, il est recommandé de s'appuyer sur des outils dont les données de fidélité et de validité sont à la fois disponibles et robustes. Cependant, dans la pratique, le logopède n'a souvent pas le choix : il doit composer avec des tests dont la qualité psychométrique est partiellement, voire pas du tout documentée. L'étude de Cattini et al. (soumis), présentée lors du Congrès BeLogo 2024, a montré que le niveau de preuve concernant la qualité psychométrique des instruments proposés comme outils diagnostiques de la dyslexie reste globalement faible (moyenne de 4.4 (± 1.73) pour une note totale sur 10). Sur les 20 outils répertoriés, la totalité fournissait des preuves de validité théorique mais les autres preuves de validité, de fidélité et de mesures de tendance centrale étaient limitées, voire absentes[1]. Cette situation rend l'interprétation des scores délicate, puisqu'une partie de leur signification repose sur des hypothèses non vérifiées.

Même si certains outils, comme l'Alouette-R (Lefavrais, 2005), disposent de données convaincantes (note de 7 sur 10 dans l'étude de Cattini et al. (soumis)), aucune épreuve ne peut à elle seule suffire à établir un diagnostic de dyslexie. C'est la convergence de plusieurs résultats, issus d'épreuves complémentaires, qui renforce la fiabilité du jugement clinique. À cela s'ajoutent parfois des contraintes environnementales (budgétaires, institutionnelles ou logistiques) qui rendent difficile, voire impossible, l'accès à certains outils.

Ni les patients ni les logopèdes ne peuvent se permettre d'attendre la publication de nouveaux tests répondant pleinement aux standards psychométriques

1 Pour une présentation plus détaillée des résultats, nous invitons le lecteur à consulter l'article de Cattini et al. (soumis), le PowerPoint de la présentation en libre accès et également, la plateforme de Tool2Care.

actuels, ou encore des études complémentaires évaluant les caractéristiques psychométriques des outils disponibles. Dès lors, comment faire ?

Dans la section suivante, nous développons quelques principes de précaution à prendre en fonction du manque de preuves disponibles concernant l'outil choisi, ou encore de l'inadéquation des outils existants au regard du profil du patient (par exemple, chez les patients présentant un trouble du développement intellectuel). L'enjeu n'est donc pas seulement de choisir un test, mais de savoir comment en interpréter les résultats avec nuance et discernement.

Évaluer de manière rigoureuse malgré les limites des outils d'évaluation

« Je n'ai pas pu respecter la standardisation de l'outil lors de la passation. »

Lorsque nous utilisons un test normé, nous sommes censés le faire dans les mêmes conditions que celles décrites dans le manuel : matériel, consignes, temps, ordre des items, modalités de réponse… C'est ce qu'on appelle la standardisation.

La standardisation correspond à l'ensemble des règles qui garantissent que chaque personne est évaluée exactement de la même manière, afin que ses scores soient comparables aux normes établies.

Si la standardisation n'est pas respectée, la comparabilité aux normes est compromise. Dans ce cadre, on ne peut plus interpréter le score de manière normative, c'est-à-dire en termes de percentiles, de moyenne ou d'écart-type. Il devient alors nécessaire de décrire qualitativement la performance (par exemple, le type d'erreurs observées, les stratégies utilisées, le soutien nécessaire ou les comportements pertinents).

En résumé, un test administré hors standardisation devient un outil d'observation et non un outil normatif. Il conserve une valeur clinique, mais ne peut plus être utilisé comme preuve diagnostique basée sur les normes.

« Mon patient ne correspond pas à l'échantillon normatif ? »

Lorsqu'on utilise un test normé, les scores sont interprétés en les comparant à un échantillon normatif, c'est-à-dire un groupe de personnes qui ont servi de référence pour construire les normes. Cet échantillon normatif doit être représentatif de la population pour laquelle le test est conçu. Les scores des patients ne prennent sens que s'ils sont comparés à un groupe qui leur ressemble réellement. Lorsque le profil d'un patient s'éloigne de celui de l'échantillon normatif, par exemple en cas de bilinguisme, la validité de la comparaison aux normes diminue fortement. Un score faible peut alors refléter un décalage entre le profil du patient et celui du groupe de référence, plutôt qu'un véritable trouble.

Dans ce cas, les recommandations sont similaires à celles formulées lorsqu'on ne parvient pas à respecter la standardisation : les résultats ne peuvent pas être interprétés de manière normative. Il est alors nécessaire d'adopter une démarche descriptive détaillée, centrée sur les types d'erreurs, les stratégies employées, les comportements observés et la convergence des informations issues d'autres sources.

« L'outil utilisé n'explicite pas les modèles cognitifs sur lesquels il repose ou ceux-ci semblent obsolètes. »

Un test diagnostique doit reposer sur des modèles cognitifs explicites et actuels, car cela conditionne sa validité théorique : c'est ce qui permet de savoir si l'épreuve mesure réellement le construit qu'elle prétend évaluer. Cependant, la validité d'un test ne dépend pas uniquement de ses concepteurs : le raisonnement clinique joue lui aussi un rôle central pour garantir une interprétation adéquate des résultats.

Par exemple, des théories récentes comme la *phonological decoding self-teaching* (PDST) décrivent l'apprentissage de la lecture comme un processus fondé sur le décodage grapho-phonologique, qui permet ensuite l'autoapprentissage d'un lexique orthographique (pour un article de vulgarisation, voir : https://cuitdanslebec.wordpress.com/2021/10/18/comment-les-enfants-apprennent-a-lire-des-ordinateurs-testent-lhypothese-du-decodage-phonologique-et-du-mecanisme-dapprentissage-de-la-lecture/).

Lorsque l'outil repose sur un modèle daté ou inadapté au développement (par exemple, le modèle à double voie issu de l'aphasiologie adulte), l'interprétation risque de s'appuyer sur des hypothèses théoriques

non pertinentes pour la population concernée. Cela peut conduire à des choix interventionnels mal ajustés, comme considérer deux voies de lecture indépendantes et cibler prioritairement un entraînement de la « voie d'adressage ».

Dans ce genre de situation, il revient au logopède d'interpréter les résultats à la lumière de modèles cognitifs contemporains, tels que la PDST (Ziegler et al., 2013), afin de garantir une analyse cohérente avec les connaissances actuelles sur l'apprentissage de la lecture.

« L'outil que j'utilise ne fournit pas de données sur son pouvoir discriminant. »

Le pouvoir discriminant d'un test correspond à sa capacité à distinguer correctement les personnes qui présentent un trouble de celles qui n'en présentent pas. Il repose sur deux indices : la sensibilité et la spécificité. Ces données sont essentielles dans un contexte diagnostique, car elles indiquent si l'outil permet réellement de repérer un trouble de manière fiable.

Or, dans notre étude, 60 % des outils évaluant la lecture ne documentaient ni leur sensibilité ni leur spécificité, ce qui est représentatif d'une situation fréquente chez les outils francophones (Cattini et al., soumis). Sans ces informations, il devient difficile d'estimer la capacité réelle de l'épreuve à contribuer au diagnostic.

Dans ce cas, il faut se montrer prudent dans l'interprétation des scores, en particulier lorsqu'ils sont comparés aux normes. L'épreuve peut être utilisée, mais uniquement comme indice parmi d'autres. Il est alors essentiel de :

- croiser les résultats avec plusieurs outils, idéalement plus robustes ;
- analyser finement le profil d'erreurs et les stratégies ;
- recueillir un maximum d'informations cliniques : historique développemental, évolution des apprentissages, retours de l'école, observations en séance ;
- mettre en cohérence l'ensemble des données plutôt que de s'appuyer sur une performance isolée.

En résumé, lorsque les outils disponibles n'offrent aucune donnée de pouvoir discriminant, le diagnostic repose avant tout sur la cohérence du tableau clinique global, plutôt que sur la performance à une épreuve isolée. S'il est certain qu'un diagnostic ne devrait jamais reposer uniquement sur des scores normés, la prudence doit être encore renforcée en cas d'absence d'informations sur la précision diagnostique de l'outil.

« Je veux interpréter les scores de mon patient mais je ne sais pas si les données normatives suivent une distribution normale. »

Pour interpréter correctement un score normé, il est important de connaître la forme de la distribution des données normatives. De nombreux tests supposent une distribution normale : une répartition en « cloche » (ou courbe de Gauss), dans laquelle la moyenne et l'écart-type sont des indicateurs fiables du niveau relatif d'un patient.

Cependant, lorsque la distribution n'est pas normale (ce qui est souvent le cas), ces indicateurs deviennent moins pertinents, car la moyenne ou l'écart-type peut être influencé(e) par des valeurs extrêmes, une difficulté trop élevée ou trop faible de l'épreuve, ou des effets plafond/plancher.

Dans ce cas, il est préférable d'utiliser les percentiles, car ils décrivent directement la position du patient, sans faire d'hypothèse sur la distribution des données. En cas d'absence d'informations, il est également recommandé d'interpréter les scores sur base des percentiles.

« L'outil que j'utilise ne fournit aucune preuve de fidélité. »

La fidélité correspond au degré de stabilité, de cohérence et de précision d'un test. Elle inclut plusieurs aspects :

- la fidélité test–retest (le score reste stable si on repasse le test dans les mêmes conditions) ;
- la fidélité inter-juges (deux évaluateurs obtiennent le même score) ;
- la consistance interne (les items mesurent bien la même compétence).

Lorsque la fidélité n'est pas documentée, il devient impossible de savoir si le score reflète réellement les compétences du patient… ou s'il résulte du hasard, de la fatigue, ou de variations dans la passation. Autrement dit : un score faible ou élevé n'est pas forcément fiable (ou autrement dit « fidèle »).

Dans ce cas, plusieurs précautions s'imposent :

- ne pas oublier que le score obtenu est uniquement une estimation, et non une mesure précise du « score vrai » du patient ;
- ne pas conclure à des progressions ou régressions entre deux passations car les variations peuvent relever de l'erreur standard de mesure.

Cet aspect est particulièrement important dans le diagnostic de la dyslexie, où la "réponse à l'intervention" et la persistance des difficultés constituent des critères centraux (Holden et al., 2025). Sans données de fidélité, il est impossible de statuer de manière fiable sur l'évolution, ce qui limite fortement l'interprétation du changement à partir de scores normés.

En résumé, sans preuve de fidélité, un score ne doit jamais être utilisé pour appuyer une conclusion forte ni pour évaluer un changement. L'épreuve peut contribuer à l'observation clinique, mais elle ne permet pas d'inférer avec précision une évolution ou un niveau de performance, car les erreurs de mesure aléatoires sont méconnues.

En conclusion

Malgré des outils imparfaits et des données psychométriques parfois incomplètes, il est possible de maintenir une évaluation rigoureuse, à condition d'adopter une posture critique, de croiser les sources d'information et de privilégier la cohérence du tableau clinique global. Il ne s'agit donc pas d'attendre que les "bons tests" arrivent, mais de savoir travailler avec les outils disponibles, en ayant pleinement conscience de leurs limites.

Cette lucidité permet de mieux équilibrer le poids des différentes informations (observations cliniques, historique, résultats scolaires, résultats aux épreuves normées) et de formuler des conclusions prudentes, argumentées et transparentes. En ce sens, le logopède dispose de véritables garde-fous pour sécuriser son raisonnement clinique, même lorsque les preuves psychométriques sont partielles.

Enfin, rendre visibles ces limites, dans nos comptes rendus, nos échanges avec les collègues et nos instances professionnelles, contribue aussi à nourrir une demande collective pour des outils mieux étudiés et mieux adaptés à nos besoins cliniques. Nous ne sommes pas de simples utilisateurs passifs des tests : en questionnant, en documentant et en faisant remonter ces enjeux, nous contribuons activement à faire évoluer l'offre d'outils d'évaluation.

RÉFÉRENCES

American Psychiatric Association. (2023). *DSM-5-TR Manuel diagnostique et statistique des troubles mentaux, texte révisé*. Elsevier Masson. https://doi.org/10.1176/appi.books.9780890425787

Betz, S., Eickhoff, J., & Sullivan, S. (2013). Factors Influencing the Selection of Standardized Tests for the Diagnosis of Specific Language Impairment. *Language Speech and Hearing Services in Schools*, *44*(April), 133–147. https://doi.org/10.1044/0161-1461(2012/12-0093)

Burnay, J., Grégoire, J., Monseur, C., & Willems, S. (2024). Lutter contre les freins à l'Evidence-Based Assessment : Création d'une grille pour examiner la qualité des outils d'évaluation en psychologie, neuropsychologie et orthophonie. *ANAE : Approche Neuropsychologique Des Apprentissages Chez l'Enfant*, *190*.

Carroll, J. M., Holden, C., Kirby, P., Thompson, P. A., & Snowling, M. J. (2025). Toward a consensus on dyslexia: findings from a Delphi study. *Journal of Child Psychology and Psychiatry*, *66*(7), 1065–1076. https://doi.org/10.1111/jcpp.14123

Cattini, J., Balla, M., & Duboisdindien, G. (soumis). Caractéristiques psychométriques des outils d'évaluation des compétences en lecture dans une visée diagnostique : une revue.

Cattini, J., & Willems, S. (2024). Analyse des pratiques professionnelles et du sentiment de compétences des orthophonistes francophones pour la sélection des tests normés. *ANAE*, *190*, 237–246.

Holden, C., Kirby, P., Snowling, M. J., Thompson, P. A., & Carroll, J. M. (2025). Towards a consensus for dyslexia practice: Findings of a delphi study on assessment and identification. *Dyslexia*, *31*(1). https://doi.org/10.1002/dys.1800

Lefavrais, P. (2005). *Alouette-R*. Les éditions du centre de psychologie appliquée.

Ogiela, D. A., & Montzka, J. L. (2021). Norm-referenced language test selection practices for elementary school children with suspected developmental language disorder. *Language, Speech, and Hearing Services in Schools*, *52*(1), 288–303. https://doi.org/10.1044/2020_LSHSS-19-00067

Ziegler, J. C., Perry, C., & Zorzi, M. (2013). Modelling reading development through phonological decoding and self-teaching: implications for dyslexia. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120397–20120397. https://doi.org/10.1098/rstb.2012.0397