**Réflexions sur le choix de tests standardisés lors du diagnostic de dysphasie**

A.-L. Leclercq1 & E. Veys²

1Docteur en Logopédie, Unité de Logopédie Clinique, Université de Liège, Belgique

²Docteur en Logopédie, Haute Ecole de la Ville de Liège, Belgique

*Correspondance à adresser à :* Anne-Lise Leclercq, Unité de Logopédie Clinique, 30 rue de l’Aunaie, B.38, 4000 Liège, al.leclercq@ulg.ac.be

*Résumé* : Cet article s’intéresse aux caractéristiques psychométriques des outils d’évaluation utilisés en clinique francophone lors du diagnostic des troubles du langage oral chez l’enfant d’âge scolaire. L’objectif est d’aider les cliniciens à adopter une pratique réflexive lors du choix des tests diagnostiques dans une démarche d’Evidence-Based Practice, ainsi que de fournir des recommandations aux concepteurs de tests afin de continuer à développer les standards de qualité des outils normés.

*Abstract*: This paper addresses the psychometric properties of the French standardized tools used in the diagnosis of language impairment in school-age children. The aim is to assist the speech-language pathologists in the selection of assessment tools following an Evidence-Based Practice approach, and to give standardized tests developers psychometric gold standards for future tools.

*Mots-clés :* psychométrie, test, diagnostic, sensibilité, spécificité

*Key words*: psychometric, assessment tool, diagnosis, sensitivity, specificity

Diagnostiquer la dysphasie est une tâche complexe nécessitant de recourir à la fois à une analyse qualitative clinique des comportements et à l'utilisation de tests standardisés. L'utilisation de tests standardisés est en effet recommandée lors du diagnostic de troubles du langage : elle est évoquée dans le guide de bonnes pratiques de l'American Speech Hearing Association, le DSM-5 évoque la nécessité de mettre en évidence des habiletés langagières quantitativement inférieures à ce qui est attendu étant donné l'âge de l'enfant et l'ICD-10 précise que "*les habiletés langagières, telles qu'évaluées par des tests standardisés, doivent être inférieures à deux déviations standards de l'âge de l'enfant*". En Belgique, l'accès au remboursement de la prise en charge en nomenclature est également conditionné par les résultats obtenus aux tests standardisés : "*Ce trouble doit être démontré par des tests normés, effectués individuellement dont le score doit être inférieur ou égal à deux écarts-types en dessous de la moyenne dans au moins un versant (expressif ou réceptif) pour trois domaines minimum (phonologie (y compris la métaphonologie), lexique/sémantique, morphologie, syntaxe).* (A.R. 19.2.2008)". Dans une démarche d'Evidence-Based Practice (EBP), les cliniciens sont encouragés à baser leur pratique sur des données empiriques émanant de la littérature scientifique pour prendre leurs décisions cliniques et les justifier (Dollaghan, 2007; Gillam & Gillam, 2006). Dans le cadre du choix des tests employés lors du diagnostic de dysphasie, il s'agit essentiellement des propriétés psychométriques des tests. Celles-ci sont consultables dans leur manuel.

**Caractéristiques psychométriques des outils diagnostiques**

McCauley & Swisher (1984) ont défini des critères de qualité (repris plus récemment par Friberg, 2010) à prendre en compte dans l'évaluation des tests standardisés. Tout d’abord, ils insistent sur l’importance d’une explicitation claire des *qualifications requises* pour l’évaluateur. Le manuel du test doit donc spécifier les connaissances ou l’entraînement requis pour l’administration du test et le recueil de données afin que celles-ci puissent être considérées comme valides. Ensuite, les auteurs évoquent la présence de *consignes d’administration et de cotation*suffisamment détaillées afin qu’elles soient réplicables, similaires aux conditions d’étalonnage. Concernant *l’échantillon d’étalonnage*, les auteurs insistent sur sa *taille* et une description claire de ses caractéristiques. D’une part, les auteurs évoquent fréquemment le nombre de 100 participants par sous-groupe d’échantillon afin de garantir une certaine stabilité et une représentativité suffisante des normes obtenues. D’autre part, les *caractéristiques* socioéconomiques, ethniques et géographiques de l’échantillon d’étalonnage doivent être clairement explicitées, en plus de l’âge, du sexe et de son caractère monolingue ou multilingue afin d’estimer si les normes sont représentatives des patients dont on souhaite comparer les scores. La pertinence d’inclure à ces normes des enfants présentant un trouble langagier dépendrait de l’objectif du test : l’inclusion de tels enfants a tendance à réduire l’efficacité diagnostique des tests, mais serait intéressante si leur objectif est de déterminer un niveau de sévérité aux troubles déjà diagnostiqués (Pena, Spaulding, & Plante, 2006).

Les outils doivent également démontrer une *analyse* approfondie *des items* qu’ils contiennent afin de s’assurer que tous les items utilisés mesurent bien ce qu’ils sont censés mesurer. Il s’agit donc de s’assurer de la fidélité interne du test, c’est-à-dire du contrôle via des analyses statistiques, des items qui y ont été inclus, au niveau de leur cohérence et de leur difficulté. Certains auteurs proposent un Alpha de Cronbach de .70 pour qu’il soit considéré comme acceptable (George & Mallery, 2003).

Bien entendu, il est important que les tests fassent état des *mesures de tendance centrale*, c’est-à-dire la moyenne et l’écart-type (ou les percentiles de performances) de l’échantillon d’étalonnage afin d’avoir un repère quantitatif clair auquel comparer les performances de l’enfant et de pouvoir les situer dans la moyenne, ou de mettre en évidence une faiblesse ou un déficit. Selon les auteurs, les tests devraient également faire état de leur *validité concourante*, c’est-à-dire démontrer des corrélations significatives entre les scores obtenus à ce test et d’autres outils censés mesurer le même concept. La *validité prédictive* doit également être prise en compte : il faut pouvoir montrer que le résultat à un test dont l’objectif est d’évaluer une certaine compétence est significativement corrélé avec l’application fonctionnelle de cette compétence au quotidien (par exemple, pour pouvoir se fier à la réussite à un test de lecture, cette réussite doit aussi refléter les performances en lecture lorsqu’elles sont évaluées en classe). Dans le cas contraire, on peut s’interroger sur la pertinence d’utiliser un test qui ne refléterait pas le fonctionnement au quotidien.

La fidélité fait référence à la stabilité du test quand la procédure de testing est répétée. Dans le cas de la *fidélité test-retest*, elle est mesurée par un coefficient de fidélité qui indique le degré de similarité entre deux sets de scores à un test administré à deux moments séparés, un coefficient de .90 étant recommandé (Salvia, Ysseldyke, & Bolt, 2010). Elle permet donc de s’assurer que plusieurs mesures successives conduiront, chez un même individu, au même résultat si cet individu n’est pas soumis à un programme d’apprentissage. Comme cette fidélité n’est malheureusement jamais de 100% et que tout test est entaché d’erreurs de mesure, il est possible de déterminer un intervalle de confiance autour de la note observée grâce à l’erreur standard de mesure. Cela sera particulièrement intéressant dans le cas de la pose d’un diagnostic : j’ai observé un score déficitaire, mais quelle était la probabilité d’observer un score dans la norme ? De même, il sera intéressant de prendre en compte cet intervalle de confiance quand il s’agira d’évaluer les progrès après un an de prise en charge. L’évolution observée est-elle statistiquement significative ? Si le score observé est en-dehors de l’intervalle de confiance initial, il y a en effet de grandes chances pour que l’évolution observée soit bien réelle et pas exclusivement due à la variabilité attendue dans les mesures. Cette erreur de mesure se calcule facilement en divisant l'écart-type par la racine carrée de la taille de l'échantillon.

La *fidélité inter-juge* est également une caractéristique psychométrique fréquemment rapportée. Elle permet de s’assurer que deux personnes qui se sont familiarisées avec les procédures de cotation d’un test produisent des cotations similaires, réplicables et fiables. McCauley et Swisher proposent un coefficient de .90, mais d’autres auteurs proposent pour leur part un Kappa de Cohen de .60 (Fleiss, 1981). Il sera particulièrement important de s’en soucier dans le cas de jugements de réponses orales. Par contre, cette information n’aura que peu d’intérêt dans le cas de réponses données par désignation. En effet, si le test consiste pour l’enfant à pointer sa réponse parmi quatre images, il est difficile d’imaginer que deux évaluateurs puissent coter différemment la désignation des mêmes images. On voit donc qu’un score élevé en fidélité inter-juge peut être crucial et révéler une grande fiabilité au test, mais peut aussi simplement découler de la procédure d’administration et ne donner que peu d’information complémentaire sur la qualité d’un test. La fidélité doit donc être vue comme un prérequis nécessaire, permettant de tirer des conclusions sur les performances observées. Elle n’est néanmoins pas suffisante pour affirmer le pouvoir classificatoire du test.

Enfin, Friberg a ajouté aux critères définis par McCauley et Swisher un dernier critère qui est l’identification de *l’objectif du test*. En effet, l’utilisation de tests peut remplir différents objectifs. Un test de première ligne aura pour principal objectif de diagnostiquer la présence ou l’absence d’un trouble, mais il sera nécessaire de mener des évaluations différentes, plus approfondies, pour mettre en évidence le degré de sévérité d’un trouble déjà avéré ou planifier des objectifs clairs de prise en charge. Il faut donc être attentif à l’objectif dont se revendiquent les outils diagnostiques utilisés et être conscient qu’un même outil « *n’a vraisemblablement pas le pouvoir de remplir ces différents objectifs* » (Pena et al., 2006).

**Réflexions sur les caractéristiques psychométriques des outils diagnostiques**

La présence de bonnes capacités psychométriques ne suffit cependant pas à garantir l’efficacité diagnostique des tests que nous utilisons. Plante et Vance ont ainsi montré, il y a déjà 20 ans, qu’une bonne taille d’échantillon, un bon niveau de fidélité test-retest, de fidélité inter-juge et de validité concourante ne garantissent pas l’identification des troubles développementaux du langage (Plante & Vance, 1994). Ces auteurs ont en effet analysé le pouvoir discriminant des outils présentant les meilleures caractéristiques psychométriques et ont observé qu’un seul d’entre eux présentait un pouvoir diagnostique satisfaisant, concluant que la présence de bonnes capacités psychométriques n’est que « *de peu d’utilité* » dans le diagnostic des troubles développementaux du langage.

En effet, le pouvoir diagnostique des outils n’est garanti que par de hauts niveaux de sensibilité et de spécificité. La *sensibilité* fait référence à la capacité d’un outil à mettre en évidence la présence d’un trouble chez les enfants qui en sont atteints. La *spécificité* fait quant à elle référence à la capacité d’un outil à rejeter la présence d’un trouble chez les enfants qui n’en sont pas atteints. Pour pouvoir se fier aux résultats d’un test, il faut regarder ces deux indices de manière conjointe. Quelles conclusions pourrions-nous tirer des résultats d’un test qui présente un haut niveau de sensibilité (90%) mais un faible niveau de spécificité (25%) ? De telles caractéristiques garantiraient en effet la détection de difficultés chez 90% des enfants avec trouble mais conduiraient également à une mise en évidence (erronée) de difficultés chez 75% des enfants sans trouble ! De même, quel crédit pouvons-nous accorder à un outil présentant un haut niveau de spécificité (95%) mais un faible niveau de sensibilité (20%) ? Cet outil rejetterait la présence de difficultés chez 95% des enfants sans trouble, mais également (à tort) chez 80% des enfants avec trouble. Il est donc important de pouvoir garantir de bons niveaux conjoints de spécificité et de sensibilité. L’importance de ces critères est directement dépendante de l’impact social associé aux erreurs d’identification. En effet, si notre objectif premier en tant qu’orthophoniste-logopède est de pouvoir identifier les troubles langagiers chez les enfants qui en sont atteints, il y a également des conséquences indéniables à sur-identifier des troubles chez des enfants qui n’en présentent pas.

Concernant la sensibilité et la spécificité, il n’y a pas de guideline clairement établie pour délimiter les seuils d’acceptabilité comme c’est le cas pour la significativité statistique, par exemple. Nous suggérons donc d’adopter ceux qui ont été proposés par Plante et Vance (1994) : 80% seraient acceptables et 90% seraient de bons niveaux. Néanmoins, ces caractéristiques sont les grandes oubliées des concepteurs de tests qui ne les rapportent en général pas. Or, comme l’ont montré Plante et Vance, il est à craindre que l’étude des qualités psychométriques des tests, sans tenir compte de leurs sensibilité et spécificité, se révèle être « *une perte de temps* » (Spaulding, Plante, & Farinella, 2006) dans l’étude de leur pouvoir diagnostique. Dès lors, comment expliquer l’absence de ces informations cruciales dans les manuels des outils diagnostiques ? Selon Betz et ses collaborateurs (Betz, Eickhoff, & Sullivan, 2013), ce serait lié à un manque d'information des utilisateurs. Selon ces auteurs, sensibilité et spécificité sont des caractéristiques moins souvent évoquées que la fidélité et la validité dans le champ des troubles de la communication chez l'enfant. Par conséquent, les cliniciens pourraient ne pas savoir comment les interpréter et les développeurs de tests ne seraient pas conscients de l'importance de ces informations. Dans leur étude, moins de 40% des tests anglophones analysés présentaient des mesures de sensibilité et de spécificité.

Notre profession évolue vers la mise en place de standards de qualité visant à garantir une pratique basée sur les meilleures preuves issues de la clinique et de la recherche, tant pour le choix des prises en charge que pour celui des outils d’évaluation : c’est l’Evidence-Based Practice (Dollaghan, 2004, 2007; Gillam & Gillam, 2006). Cette démarche est similaire à celle adoptée par nos collègues médecins. En effet, confirions-nous un diagnostic médical important à des outils qui n’ont pas prouvé leur pouvoir discriminant mais qui sont les seuls dont notre médecin dispose ? Nous pensons qu'il est important de prendre nos responsabilités et de revendiquer des tests psychométriques de qualité. En effet, peu de pression existe actuellement de la part des utilisateurs de tests quant à la définition et au respect de standards de qualité pour les outils qu’ils utilisent.

**Etat des lieux des outils francophones**

Nous nous sommes intéressées à la présence ou à l’absence des caractéristiques psychométriques proposées par Friberg (2010) dans cinq batteries d’évaluation couramment utilisées dans le diagnostic langagier des enfants francophones d’âge scolaire : l’ELO (Evaluation du langage oral, Khomsi, 2001), la NEEL (Nouvelles épreuves pour l’évaluation du langage, Chevrie-Muller & Plaza, 2001), l’Exalang 5-8 (Batterie informatisée pour l’examen du langage oral et écrit chez l’enfant de 5 à 8 ans, Thibault, Helloin, & Croteau, 2010), l’Exalang 8-11 **(Bilan informatisé pour l’examen du langage et des compétences transversales chez l’enfant de 8 à 11 ans, Thibault, Lenfant, & Helloin, 2012)** et la L2MA2[[1]](#footnote-1) (Batterie Langage oral, langage écrit, mémoire et attention – 2e édition, Chevrie-Muller, Maillart, Simon, & Fournier, 2010). En plus de ces caractéristiques, nous avons évalué la présence ou l’absence de données relatives au pouvoir discriminant de ces batteries, c’est-à-dire leur sensibilité et leur spécificité. Enfin, nous avons ajouté le critère de validité théorique, c’est-à-dire *« l’évaluation du sens à attribuer aux scores à l’échelle sur la base d’un modèle théorique* » (Laveault & Grégoire, 2002, p.165). Il nous paraît en effet important que les épreuves sur lesquelles nous basons notre diagnostic, et les items qui les composent, aient été construits à partir de modèles théoriques du fonctionnement cognitif de l’habileté que nous souhaitons évaluer. Cela doit avant tout permettre un diagnostic éclairé quant à l’origine du déficit, mais également par la suite guider l’orientation de la prise en charge.

Le tableau 1 présente une synthèse des caractéristiques psychométriques pour chaque batterie : le signe « + » indique que le critère est rempli, le signe « - » signale que le critère n’est pas rempli, et le signe « +/- » a été attribué lorsque le critère a été mentionné mais qu’il n’est pas suffisamment détaillé (les statistiques et leurs résultats chiffrés ne sont par exemple pas disponibles).

<Placer le Tableau 1 ici>

Un premier constat est qu’aucune de ces grandes batteries ne remplit l’ensemble des critères recommandés. Seuls trois critères sont systématiquement validés : il s’agit des *consignes d’administration et de cotation*, des *mesures de tendance centrale* et de la *spécification de l’objectif du test*. Concernant le premier point, il faut toutefois noter qu’il existe de grandes disparités dans la précision avec laquelle ces consignes sont détaillées, la NEEL et la L2MA2 fournissant les consignes d’administration et de cotation les plus détaillées. Or, la précision des consignes d’administration et de cotation est inévitablement liée à la fidélité inter-juge. Concernant le second point, des moyennes et écart-types sont fournis pour chaque épreuve dans chaque tranche d’âge. Néanmoins, la pertinence de l’utilisation des moyennes et écart-types est directement dépendante de la normalité de la distribution : définir un score-seuil, par exemple, à -2 écart-types n’est pertinent que si on a la garantie que seuls 2,5 % de l’échantillon se trouvent en effet sous ce seuil. Par conséquent, il faudrait d’abord s’assurer que la distribution des scores de l’échantillon d’étalonnage est normale afin d’utiliser à bon escient les moyennes et écart-types. Dans certains cas, l’utilisation des percentiles pourrait en effet s’avérer plus adaptée aux normes d’étalonnage. Enfin, concernant la *spécification de l’objectif du test*, nous avons validé le critère lorsque le manuel évoquait cet objectif. Il est toutefois étonnant de relever des objectifs très différents d’un manuel à l’autre. Le manuel de l’ELO précise que « *cette batterie ELO vise donc à décrire divers aspects du fonctionnement langagier oral des enfants […] on notera que son but ne peut être, en aucun cas, la description des différences individuelles avec une finalité de classement*. » et celui de la NEEL évoque « *l’exploration clinique des différentes composantes du langage » ;*  les auteurs ne parlent donc pas directement d’un objectif diagnostique. L’Exalang 5-8 parle clairement d’un objectif de diagnostic d’un trouble spécifique du langage oral et/ou écrit et l’Exalang 8-11 parle à la fois d’un objectif diagnostique et d’établissement d’un profil thérapeutique pour les épreuves proposées. Enfin, la L2MA2 a construit son évaluation en deux niveaux : un bilan diagnostique rapide, puis un approfondissement de l’évaluation pour orienter la prise en charge. Ce choix de deux niveaux distincts semble le plus raisonnable si l’on veut remplir ce double objectif puisqu’une même épreuve n’a vraisemblablement pas la capacité de remplir les deux (Pena et al., 2006).

Deux autres critères remplis par 4 des 5 batteries sont la description de la *population d’échantillon*etla *validité théorique*. Pour remplir le premier critère, les tests devaient préciser les caractéristiques socioéconomiques, géographiques, linguistiques, d’âge et de genre de l’échantillon d’étalonnage (nous n’avons par contre pas pris en compte le caractère ethnique généralement mentionné aux Etats-Unis). La plupart des concepteurs de tests a donc veillé à fournir à l’utilisateur les moyens de vérifier la représentativité des normes de comparaison en regard du profil de ses patients. Concernant la *validité théorique*, nous avons validé le critère si les manuels des tests en faisaient mention. Il s’agit en général d’éléments de théorie liés aux fonctions cognitives ciblées par les tâches, mais pas d’une réelle justification des items à proprement parler.

Vient ensuite la *taille de l’échantillon*. Nous avons ici appliqué strictement le critère de Friberg en ne le validant que si l’échantillon de *chaque tranche d’âge* comporte un minimum de 100 participants. Seules deux des cinq batteries remplissent entièrement ce critère. Nous souhaitons néanmoins nuancer ces résultats en précisant que l’ELO le remplit pour 5 des 8 tranches d’âge proposées, alors que l’Exalang 5-8 et la L2MA2 présentent les échantillons variant de 80 à 95 par tranche d’âge, ce qui est proche du critère.

Bien que ce critère ne pose généralement pas de problème, il est étonnant de constater qu’aucun outil ne stipule clairement les *qualifications requises* pour la personne qui peut administrer la batterie, et que seules les Exalang 5-8 et 8-11 précisent la formation des personnes qui les ont administrées lors de l’étalonnage. Concernant la *validité concourante*, bien que l’Exalang 8-11 relate que « *la corrélation est bonne entre les résultats d’Exalang 8-11 et ceux obtenus à divers autres tests évaluant des fonctions comparables* […] », il est regrettable que ces corrélations et les tests de comparaison ne soient pas précisés. De même, l’Exalang 5-8 évoque « *des résultats proches ou semblables aux résultats obtenus pour les items comparables d’autres batteries* ». En l’absence de données chiffrées précises, on ne peut valider la présence de ce critère. Par ailleurs, ces informations ne sont mentionnées que pour une petite partie des épreuves. Il faut toutefois relativiser la pertinence de ce critère : seule une corrélation importante avec un outil qui a démontré son efficacité diagnostique apporterait du crédit à l’outil évalué. Or, cette pertinence diagnostique doit encore être démontrée pour la plupart des outils francophones existants.

Concernant la *validité prédictive*, les Exalang 5-8 et 8-11 précisent que « *les résultats scolaires […] reflètent les compétences mises en lumière par la batterie* ». Néanmoins, il serait intéressant de disposer des mesures détaillées des « résultats scolaires » ainsi que de leurs corrélations avec les résultats aux épreuves de la batterie. A nouveau, en l’absence de données chiffrées précises, on ne peut valider la présence de ce critère. Cette caractéristique nous paraît cruciale car il s’agit de la finalité de l’orthophonie/logopédie : l’efficacité fonctionnelle au quotidien. Cependant, il nous faut admettre qu’il peut être difficile d’obtenir des données objectives de fonctionnement dans la vie de tous les jours.

Seules les batteries Exalang 5-8 et 8-11 font état des *fidélités test-retest et inter-juge*. Néanmoins, les corrélations n’atteignent pas le seuil de r = .90 (recommandé par Friberg) pour toutes les épreuves de la batterie. En ce qui concerne *l’analyse des items*, aucune batterie n’évoque un contrôle de la cohérence des items inclus dans chaque sous-test afin d’en garantir l’homogénéité. Néanmoins, la L2MA2 fait état d’une pré-expérimentation ayant fait l’objet d’analyses statistiques approfondies afin de choisir les items pour la version définitive. Nous n’avons toutefois pas validé la présence de ce critère en l’absence de données chiffrées précises. Par ailleurs, l’ELO détaille, pour toutes les épreuves, le niveau de réussite de chaque item pour les différentes tranches d’âge. Cela permet une première estimation de la cohérence interne des items, sans toutefois avoir exclu les items dont le taux de réussite se démarque fort des autres.

Enfin, seuls les manuels des Exalang 5-8 et 8-11 mentionnent le pouvoir discriminant de ces outils, à savoir leur *sensibilité et leur spécificité*. Cependant, cette mention reste partielle car les auteurs évoquent « *les enfants diagnostiqués comme présentant une pathologie ou déjà suivis en rééducation orthophonique et répertoriés comme présentant un trouble du langage oral ou d’entrée dans l’écrit apparaissent également comme pathologiques ou en dessous des normes d’Exalang 5-8* » (Exalang 5-8, p.16) ; « *de bonnes correspondances entre le diagnostic, les observations cliniques des orthophonistes et les résultats à la batterie* » (Exalang 8-11, p.16). Il est néanmoins dommage que les auteurs ne mentionnent pas les statistiques généralement utilisées : analyse discriminante, likelihood ratio, pourcentage de sensibilité et de spécificité.

Notre analyse de cinq grandes batteries d’examen du langage chez les enfants francophones d’âge scolaire reflète les constats précédemment réalisés pour les outils anglophones. D’une part, les outils diagnostiques présents sur le marché ne rencontrent pas l’ensemble des critères psychométriques recommandés pour une pratique de qualité (Friberg, 2010 ; Plante & Vance, 1994). D’autre part, un critère diagnostique crucial, le pouvoir discriminant des outils, reste le grand oublié des concepteurs de tests, peut-être par manque d’information (Betz et al., 2013 ; Spaulding et al., 2006).

**Comment appliquer ces concepts à notre pratique clinique ?**

Selon une démarche d’Evidence-Based Practice, ce sont les tests présentant de bonnes valeurs de fidélité, de validité et de pouvoir discriminant qui devraient être choisis dans la pratique clinique. Si les recommandations vont clairement dans le sens d'une prise en compte de ces caractéristiques psychométriques lors du choix des outils diagnostiques, la pratique clinique s'en éloigne. Betz, Eickhoff et Sullivan (2013) ont évalué la corrélation entre les qualités psychométriques des tests standardisés et leur fréquence d'utilisation auprès de 364 cliniciens à travers les Etats-Unis. Leurs conclusions sont édifiantes : ni la fidélité, ni la validité, ni le pouvoir discriminant des tests ne semble influencer leur fréquence d'utilisation par les cliniciens. Seule l'année de publication est statistiquement corrélée avec l’utilisation. Les auteurs en concluent que la section technique des manuels est trop souvent oubliée par les cliniciens lors du choix des tests et par conséquent, que l'importance de certaines caractéristiques est parfois minimisée par les développeurs de tests. Ces considérations sont vraisemblablement transposables aux pratiques cliniques francophones. On peut même supposer que les cliniciens ne cherchent pas ces indices de qualité diagnostique dans les manuels de tests.

Comme nous l’avons développé, seul un bon pouvoir discriminant permet de garantir la véracité du diagnostic posé. On ne pourra donc pas faire l’économie d’études approfondies sur la spécificité et la sensibilité des outils d’évaluation. En attendant, pour une bonne utilisation des outils dont nous disposons, nous pouvons veiller à prendre en compte certains critères qui risquent d’influencer fortement la capacité diagnostique des outils.

Une première manière d’approcher ces informations serait d’identifier ce qui, dans les normes, nous donne une indication du niveau possible de sensibilité et de spécificité. Rappelons que la sensibilité fait référence à la capacité d’un test à détecter la présence de difficultés chez les enfants qui en sont atteints. En termes de critères normatifs, c’est lié à la capacité d’observer un score significativement inférieur à la moyenne chez les enfants dysphasiques. Par conséquent, un test présentant une moyenne trop basse ou un écart-type trop grand ne permettra pas facilement de détecter une faiblesse ou un déficit chez ces enfants. Par exemple, dans un test de 15 items dont la moyenne est de 6,15 avec un écart-type de 4,45, la seule possibilité d’observer une faiblesse chez un enfant est que celui-ci ne donne qu’une seule réponse juste. Le déficit ne sera quant à lui pas observable puisque même un score nul ne sera pas déficitaire.

De même, la spécificité fait référence à la capacité d’un test à rejeter la présence de difficultés chez les enfants qui ne les présentent pas. En critères normatifs, cela implique de pouvoir objectiver un score dans la norme ou supérieur à celle-ci pour les enfants sans troubles. Par conséquent, si les scores de l’échantillon normatif plafonnent et que la moyenne est trop élevée ou que l’écart-type est trop étroit, le score d’un enfant peut, par distraction ou par fatigue, rapidement se retrouver en-deçà des normes alors qu’il ne présente pas de trouble avéré. Par exemple, dans le cas d’un test sur 15 items dont la moyenne est de 14,65 avec un écart-type de 0,60, l’enfant qui fait une erreur tombera en-deçà des normes et celui qui fait deux erreurs aura un score considéré comme déficitaire. Etant donné les considérations précédemment explicitées sur les erreurs mesures possibles dont est entaché tout test, il est primordial de limiter cet effet au maximum. Il serait donc intéressant de regarder systématiquement l’écart entre le score maximum et la moyenne d’une part, l’écart entre le score maximum et le score seuil d’autre part. La psychométrie ne définit pas de critère strict sur l’écart minimal à respecter : il s’agit de considérer la significativité clinique d’un tel écart. En l’absence de consensus à ce niveau, nous proposons de considérer inacceptables des normes proposant moins de deux points d’écart entre le score seuil et le score maximal, de considérer limites des normes ne proposant que 2 ou 3 points d’écart et de considérer comme acceptables des normes proposant 4 points d’écart et au-delà. Par ailleurs, une évolution significative des scores en fonction de l’âge limite le risque d’être en présence d’une épreuve présentant des scores « plafond » (ou « plancher »). Il s’agirait donc d’un argument en faveur d’une certaine spécificité (ou sensibilité) de la tâche.

**Quels seuils de pathologie adopter ?**

Les psychométriciens définissent la significativité statistique à p = .05, c’est -à-dire qu’on peut considérer une performance comme étant significativement différente de celle attendue si elle avait moins de 5% de chance d’être à ce point différente de la moyenne. On peut donc considérer que cela correspond à -2 écart-types de la moyenne de l’échantillon d’étalonnage, si les 5% se répartissent de part et d’autre de la moyenne. Dans ce cas, l’enfant n’avait même que 2,5% de chance d’avoir un score si faible. Mais on peut également ne considérer que la partie faible de la courbe et prendre un seuil de significativité aux 5% les plus faibles, soit alors -1,65 écart-types en-dessous de la moyenne. Il n’existe en fait pas de consensus à ce sujet : si dans nos pays francophones la tendance est à considérer pathologique une performance langagière inférieure à -2 écart-types en-dessous de la moyenne (y compris au Québec, Elin Thordardottir et al., 2011), une récente étude en la matière (Spaulding et al., 2006) a montré qu’aux Etats-Unis les critères de diagnostic appliqués en clinique varient en fonction de l’Etat (de -1,5 à -2 écart-types). Il en va de même pour les critères internationaux de recherche qui varient en fonction de l’étude, du pays ou des chercheurs (de -1 à -1,5 écart-types). Cependant, appliquer un même score seuil à l’ensemble des outils d’évaluation pourrait s’avérer inapproprié. En analysant 43 outils normés utilisés dans le diagnostic des troubles du langage chez les enfants et adolescents anglophones, Spaulding et ses collègues (Spaulding et al., 2006) ont montré que l’utilisation d’un même score seuil pour l’ensemble des outils amenait un taux non négligeable d’erreurs diagnostiques, dans un sens comme dans l’autre.

En outre, ces scores seuils sont définis à partir du présupposé selon lequel la distribution des scores de l’échantillon d’étalonnage est suit une courbe normale, ce qui est en fait rarement le cas. Par conséquent, il est possible que l’on évolue vers des critères normatifs qui ne soient pas unifiés mais plutôt vers l’utilisation de scores seuils qui seraient fonction du test et directement dérivés à partir de l’échantillon d’étalonnage. De plus en plus d’auteurs, s’inscrivant dans le courant de l’EBP, recommandent en effet d’abandonner l’utilisation de scores seuils généralisés au profit de scores-seuils individualisés à chaque test, dérivés des données probantes issues d’analyses discriminantes (Plante & Vance, 1994; Spaulding et al., 2006; Spaulding, Szulga, & Figueroa, 2012). Seules ces analyses garantiraient le pouvoir discriminant de ces tests, soient de bons niveaux de sensibilité et de spécificité.

**Quels outils privilégier ?**

Une manière avisée de choisir les outils pertinents dans le diagnostic des enfants dysphasiques est de cibler certains outils qui évaluent spécifiquement les zones de faiblesse chez ces enfants. En effet, plusieurs auteurs ont évoqué le concept de « marqueurs diagnostiques » pour certaines épreuves systématiquement échouées par ces enfants: on relève notamment la répétition de pseudo-mots, la répétition de phrases et l’utilisation de la morphologie verbale (notamment, Conti-Ramsden, Botting, & Faragher, 2001; Stokes, Wong, Fletcher, & Leonard, 2006; Elin Thordardottir et al., 2011). Des auteurs ont en effet montré que les différences de moyennes groupes entre les enfants dysphasiques et leurs contrôles sont plus importantes pour ces épreuves ciblées que pour d’autres évaluant également le langage oral (comme le lexique par exemple, Spaulding et al., 2006). Néanmoins, choisir une épreuve ciblant une faiblesse connue des enfants dysphasiques ne garantit pas la qualité du diagnostic. En effet, tous les tests évaluant la morphosyntaxe ne montrent pas le même écart de performances entre les enfants contrôles et les enfants dysphasiques, notamment en fonction de la complexité de leurs items. Certains items restent plus discriminants que d’autres, y compris au sein d’un même domaine langagier (Spaulding et al., 2006). Une bonne connaissance de la pathologie mais également du développement langagier typique doit donc être une aide précieuse dans la sélection des outils d’évaluation et dans la prise de décision lors du diagnostic. Quelle que soit la norme de l’échantillon d’étalonnage, on ne considère par exemple pas problématique le fait d’avoir des difficultés à réaliser une épreuve de fluences phonétiques ou de jugement morphosyntaxique en maternelles.

Par ailleurs, certaines de ces épreuves pourraient s’avérer particulièrement précieuses dans le cadre complexe du diagnostic dans un contexte multilingue. Nous disposons actuellement de peu de support et de recommandations pour faire face aux situations (de plus en plus fréquentes) d’évaluation d’enfants dont le français n’est pas la langue maternelle. En effet, il n’est pas toujours facile de distinguer, dans un score faible, ce qui peut être attribué à un manque d’exposition à la langue de ce qui peut être attribué à des problèmes de traitement ou de représentations langagières. Enfants dysphasiques et enfants bilingues peuvent présenter des performances comparables à différentes épreuves, pour des raisons différentes. Des auteurs ont commencé à s’intéresser à cette question au niveau international et, d’après les premiers résultats, certaines épreuves pourraient s’avérer plus appropriées que d’autres. Thordadottir et Brandeker (2013) ont ainsi montré que la répétition de non-mots était une épreuve peu influencée par l’exposition langagière (contrairement au vocabulaire, par exemple) et qu’elle permettait, comme la répétition de phrases, de distinguer adéquatement les enfants dysphasiques des enfants sans troubles, y compris chez les enfants bilingues anglais-français. De futures études dans ce sens doivent permettre de développer des recommandations précises sur l’évaluation langagière des enfants en contexte multilingue.

**Conclusions : Elaboration de standards de qualité pour les outils d’évaluation**

Nous souhaitons donc amener une réflexion en profondeur de la part des cliniciens et des concepteurs de tests afin de garantir la qualité des outils sur lesquels reposent (en partie) nos diagnostics. Le tableau 2 reprend une synthèse des critères de qualité sous forme de recommandations aux concepteurs d’outils d’évaluation. La prochaine étape sera de pouvoir définir des critères concernant les seuils minimalement acceptables pour affirmer qu’un test est approprié à la prise de décision diagnostique en clinique.

<Placer le Tableau 2 ici>

**Références**

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*, 133-146.

Chevrie-Muller, C., Maillart, C., Simon, A.-M., & Fournier, S. (2010). *Batterie Langage oral, langage écrit, mémoire, attention - 2ème édition*. Montreuil: ECPA.

Chevrie-Muller, C., & Plaza, M. (2001). *Nouvelles épreuves pour l'examen du langage*. Montreuil: ECPA.

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry, 42*(6), 741-748. doi: http://dx.doi.org/10.1111/1469-7610.00770

Dollaghan, C. A. (2004). Evidence-based practice in communication disorders: What do we know, and when do we know it? *Journal of Communication Disorders, 37*(5), 391-400. doi: http://dx.doi.org/10.1016/j.jcomdis.2004.04.002

Dollaghan, C. A. (2007). *The handbook for evidence-based practice in communication disorders*. Baltimore: Paul H. Brookes Publishers.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions. 2nd ed*. New York: John Wiley.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy, 26*(1), 77-92. doi: http://dx.doi.org/10.1177/0265659009349972

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.)* Boston: Allyn & Bacon.

Gillam, S. L., & Gillam, R. B. (2006). Making evidence-based decisions about child language intervention in schools. *Language, Speech, and Hearing Services in Schools, 37*(4), 304-315. doi: http://dx.doi.org/10.1044/0161-1461%282006/035%29

Khomsi, A. (2001). *Evaluation du langage oral*. Paris: ECPA.

Laveault, D., & Grégoire, J. (2002). *Introduction aux théories des tests en psychologie et sciences de l'éducation*. Bruxelles: De Boeck.

McCauley, R., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49*, 34-42.

Pena, E., Spaulding, T. J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech Language Pathology, 15*(3), 247-254.

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25*, 15-24.

Salvia, J., Ysseldyke, J., & Bolt, S. (2010). *Assessment: In special and inclusive education (11th ed.)*. Belmont, CA: Wadsworth.

Spaulding, T. J., Plante, E., & Farinella, K. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools, 37*, 61-72.

Spaulding, T. J., Szulga, M., & Figueroa, C. (2012). Using norm-referenced tests to determine severity of language impairment in children: Disconnect between U.S. policy makers and tests developers. *Language, Speech, and Hearing Services in Schools, 43*, 176-190.

Stokes, S. F., Wong, A. M., Fletcher, P., & Leonard, L. B. (2006). Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. *Journal of Speech, Language, and Hearing Research, 49*(2), 219-236.

Thibault, M.-P., Helloin, M.-C., & Croteau, B. (2010). *Batterie informatisée pour l'examen du langage oral et écrit chez l'enfant de 5 à 8 ans version 2*. Mont-Saint-aignan: Orthomotus.

Thibault, M.-P., Lenfant, M., & Helloin, M.-C. (2012). *Bilan informatisé pour l’examen du langage et des compétences transversales chez l’enfant de 8 à 11 ans*. Mont-Saint-Aignan: Orthomotus.

Thordardottir, E., & Brandeker, M. (2013). The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. *Journal of Communication Disorders, 46*, 1-16.

Thordardottir, E., Kehayia, E., Mazer, B., Lessard, N., Majnemer, A., Sutton, A., . . . Chilingaryan, G. (2011). Sensitivity and specificity of French language and processing measures for the identification of primary language impairment at age 5. *Journal of Speech, Language, and Hearing Research, 54*(2), 580-597.

**Tableau 1**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Critère** | **ELO** | **NEEL** | **Exalang 5-8** | **Exalang**  **8-11** | **L2MA2** | **Nombre de tests remplissant le critère** |
| 1. Qualifications de l’évaluateur | - | - | + | + | - | 2 |
| 1. Consignes d’administration et de cotation | + | + | + | + | + | 5 |
| 1. Taille de l’échantillon | - | + | - | + | - | 2 |
| 1. Description de l’échantillon | - | + | + | + | + | 4 |
| 1. Analyse des items | - | - | - | - | +/- | 0 (1) |
| 1. Mesures de tendance centrale | + | + | + | + | + | 5 |
| 1. Validité concourante | - | - | +/- | +/- | - | 0 (2) |
| 1. Validité prédictive | - | - | +/- | +/- | - | 0 (2) |
| 1. Fidélité test-retest | - | - | - | - | - | 0 |
| 1. Fidélité inter-juge | - | - | - | - | - | 0 |
| 1. Objectif du test précisé | + | + | + | + | + | 5 |
| 1. Sensibilité et spécificité | - | - | - | - | - | 0 |
| 1. Validité théorique | - | + | + | + | + | 4 |
| Nombre de critères remplis | 3/13 | 6/13 | 7/13 | 8/13 | 5,5/13 |  |

*Tableau 1. Relevé des caractéristiques psychométriques présentes dans les différentes batteries d’évaluation analysées.*

**Tableau 2**

|  |  |
| --- | --- |
| **Critère** | Explications |
| Qualifications de l’évaluateur | Les qualifications de la personne qui va administrer le test, le corriger et l’interpréter doivent être clairement explicitées afin de garantir la validité des résultats. |
| Consignes d’administration et de cotation | Les consignes d’administration et de cotation sont clairement spécifiées afin de minimiser la subjectivité lors de l’administration et de la cotation. |
| Taille de l’échantillon de l’étalonnage | La taille de l’échantillon d’étalonnage doit être d’au minimum 100 participants par tranche d’âge / sous-groupe. |
| Caractéristiques de l’échantillon d’étalonnage | Les caractéristiques géographiques, socioéconomiques, linguistiques, l’âge et le genre de la population de l’échantillon d’étalonnage doivent être clairement explicitées et représentatives de la population tout venant. |
| Analyse des items | Le manuel doit faire état d’une analyse statistique de la pertinence des items inclus dans les sous-épreuves, notamment en en démontrant la cohérence interne. |
| Mesures de tendance centrale | Les moyennes et écart-type de l’échantillon d’étalonnage doivent être mentionnées pour chaque tranche d’âge (et/ou les percentiles si la distribution des scores n’est pas normale). |
| Validité concourante | A l’avenir, chaque outil devra pouvoir montrer une bonne corrélation entre ses résultats et ceux d’autres épreuves mesurant les mêmes fonctions cognitives et ayant prouvé leur pertinence diagnostique. |
| Validité prédictive | La pertinence fonctionnelle de l’outil doit être attestée via une concordance entre les scores observés à l’outil et le fonctionnement dans les activités de vie quotidienne mettant en œuvre la fonction évaluée. |
| Validité théorique | La conception de l’outil et le choix des items qui le composent doivent reposer sur les modèles théoriques récents de la fonction cognitive évaluée. |
| Fidélité test-retest | L’outil doit faire preuve d’une fidélité test-retest suffisante afin de garantir la stabilité des résultats. Friberg (2010) recommande un coefficient de corrélation de .90. |
| Fidélité inter-juge | L’outil doit faire preuve d’une fidélité inter-juge suffisante afin de garantir des résultats les plus objectifs possibles, indépendamment de la personne qui l’a administré et corrigé. Friberg (2010) recommande un coefficient de corrélation de .90 ; un Kappa de Cohen de .60 étant le minimum souhaité (Fleiss, 1981). |
| Objectif du test précisé | Les concepteurs doivent poser un choix clair concernant l’objectif de leur outil (diagnostic, détermination d’un niveau de sévérité, orientation thérapeutique) et le préciser. |
| Sensibilité et spécificité | Le pouvoir discriminant de l’outil, c’est-à-dire sa sensibilité et sa spécificité, doit avoir fait l’objet d’analyses spécifiques incluant notamment une population pathologique, afin de garantir son pouvoir diagnostique. |

*Tableau 2. Synthèse des recommandations concernant les outils d’évaluation.*

1. Concernant cette dernière batterie, les auteurs précisent que la première version du manuel ne mentionne pas l’ensemble des caractéristiques psychométriques de l’outil et qu’une version plus complète devrait être mise en ligne avec les analyses psychométriques réalisées (notamment, analyses d’items, validité concourante, et intérêt dans l’évaluation du langage d’enfants dysphasiques). [↑](#footnote-ref-1)