



UNIVERSITÉ DE NANTES

Unité de Formation et de Recherche de Médecine et des Techniques Médicales

Année Universitaire 2017/2018

Mémoire

pour l'obtention du

Certificat de Capacité en Orthophonie

**Analyse psychométrique d'outils utilisés dans
l'évaluation orthophonique du langage écrit**

présenté par *Alexandra DEMOULIN*

Née le 23/01/1994

Présidente du jury : Madame Martinage Valérie – Orthophoniste, chargée de cours,
directrice pédagogique

Directrice du mémoire : Madame Calvarin Suzanne – Orthophoniste, chargée de cours

Membre du jury : Madame Desvé Mélinda, Orthophoniste, chargée de cours

Remerciements

A Mme Suzanne Calvarin qui a accepté d'encadrer ce projet et y a apporté un éclairage précieux.

Au CFUO de Nantes pour les outils disponibles.

Aux orthophonistes qui m'ont accueillie en stage pour tous ces échanges.

A ma famille et mes amis pour leur soutien toutes ces années, à Robin et à Christopher en particulier.

Merci.



UNIVERSITÉ DE NANTES
FACULTÉ DE MÉDECINE
ET DES TECHNIQUES MÉDICALES

Centre de Formation Universitaire en Orthophonie
Directeur : Dr Florent ESPITALIER
Directrice Pédagogique : Mme Valérie MARTINAGE
Directrice des Stages : Mme Christine NUEZ

ANNEXE 9 ENGAGEMENT DE NON-PLAGIAT

« Par délibération du Conseil en date du 7 Mars 1962, la Faculté a arrêté que les opinions émises dans les dissertations qui lui seront présentées doivent être considérées comme propres à leurs auteurs et qu'elle n'entend leur donner aucune approbation ni improbation ».

Engagement de non-plagiat

Je, soussigné(e)
déclare être pleinement conscient(e) que le plagiat de documents ou d'une partie d'un document publiés sur toutes ses formes de support, y compris l'Internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée. En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour écrire ce mémoire.

Fait à : Le

Signature :

Sommaire

I- Introduction	1
1. <u>Evaluation et pratique professionnelle</u>	2
1.1 Questionnements professionnels autour de l'évaluation chiffrée : intérêts et limites..	2
1.2 Evaluation et responsabilité	4
1.3 Choisir les outils d'évaluation.....	5
1.4 Connaître les outils d'évaluation.....	7
2. <u>Psychométrie et perspectives en orthophonie</u>	8
2.1 Définitions.....	8
2.2 Echantillon de normalisation	11
2.3 Etudes menées en orthophonie et constats	12
II- Méthode	15
1. <u>Recensement des outils d'évaluation</u>	15
2. <u>Matériel</u>	16
3. <u>Choix et définitions des critères psychométriques retenus pour l'analyse</u>	16
4. <u>Traitement des données</u>	22
4.1. Participants au traitement des données	22
4.2. Procédure de traitement.....	22
4.3. Mode de présentation des résultats	23
5. <u>Diffusion des résultats</u>	25
III- Résultats	27
1. <u>Résultats globaux</u>	27
2. <u>Résultats par critère</u>	28
2.1 Critères généraux	29
2.2 Critères relatifs à l'échantillon	30
2.3 Validité.....	31
2.4 Fidélité	32
2.5 Sensibilité et spécificité.....	33
3. <u>Synthèse des résultats</u>	33
4. <u>Données complémentaires</u>	34
4.1 Outils et date	34
4.2 Population visée	34
4.3 Disponibilité des outils.....	35
IV- Discussion	37
1. <u>Constats</u>	37
2. <u>Aux orthophonistes</u>	39
2.1 L'importance du manuel d'utilisation.....	39
2.2 Des revendications de la profession ?	41
3. <u>Aux concepteurs et éditeurs d'outils d'évaluation</u>	42
4. <u>Limites du travail effectué</u>	43
5. <u>Perspectives</u>	43
6. <u>Conclusion</u>	45

Bibliographie

Annexes

Introduction

En France, les actes correspondant à l'intitulé « Rééducation des pathologies du langage écrit : lecture et/ou orthographe » représentaient en 2010 27,6% du montant de l'ensemble des soins délivrés en orthophonie. (Observatoire national de la démographie des professions de santé, 2012, p.95). Il s'agit donc d'un domaine largement présent dans la pratique des orthophonistes. Le bilan constitue une tâche incontournable du travail du praticien. Le professionnel ne peut conclure à la nécessité ou non d'une prise en charge qu'après avoir rencontré le patient et procédé à une évaluation. L'orthophoniste va devoir, à partir des données recueillies durant l'anamnèse, orienter la suite de son bilan et investiguer les domaines pour lesquels il juge cela nécessaire. La structure rédactionnelle du bilan s'inscrit dans un cadre légal : l'orthophoniste doit suivre le plan « Objet du bilan - Anamnèse - Bilan - Diagnostic orthophonique - Projet thérapeutique ». Cependant si un cadre est donné l'orthophoniste reste libre quant au choix des techniques et outils employés lors du bilan. Ainsi afin de répondre à l'objectif de diagnostic la question du choix des outils peut se poser. Le référentiel d'activités de l'orthophoniste (Ministère de l'enseignement supérieur et de la recherche, 2013) indique que l'orthophoniste peut utiliser des épreuves standardisées. Bien que l'utilisation de ces épreuves n'apparaisse en rien comme une obligation légale pour le professionnel, leur emploi peut s'avérer pertinent voire nécessaire. Elles contribueront au diagnostic. D'après un rapport de l'INSERM, reprenant les critères du DSM-V :

tout diagnostic d'un trouble des apprentissages doit se baser non seulement sur l'entretien et l'observation informelle pratiquée par le clinicien, mais également sur des mesures plus objectives telles que celles obtenues grâce à des batteries de tests dûment validées et étalonnées (INSERM, 2007).

Que dire des outils d'évaluation utilisés en France dans le champ du langage écrit ? Sont-ils « dûment » validés et étalonnés ? Dans quelle mesure se fier aux outils et aux résultats qu'ils apportent ?

Le présent travail a pour objectif de proposer une analyse de tests et batteries utilisés dans le domaine du langage écrit en France. Les outils d'évaluation seront évalués à partir de critères psychométriques.

Evaluation et pratique professionnelle

En France, la pratique du bilan orthophonique s'inscrit dans un cadre légal attestant la nécessité de celui-ci avant toute prise en charge :

Le bilan est l'outil indispensable à la pose du diagnostic orthophonique, à la décision thérapeutique et à la conduite du traitement; il en est le fondement. (Ministère de l'enseignement supérieur et de la recherche, 2013).

Si l'utilisation d'outils standardisés n'est pas une obligation, il apparaît cependant qu'ils sont largement employés par les orthophonistes. Lors d'une expérimentation sur l'Evaluation des Pratiques Professionnelles, 91% des orthophonistes participant à l'étude ont répondu utiliser des outils comportant une cotation, un étalonnage ou une grille d'analyse clinique (Bétrancourt et Lederlé, 2017). Dans le cadre plus précis de l'évaluation du langage écrit, le même constat est fait : les orthophonistes utilisent principalement des outils normés (Brin-Henry, 2015).

Questionnements professionnels autour de l'évaluation chiffrée : intérêt et limites

Une épreuve étalonnée est définie par le fait qu'elle a été administrée à un groupe de référence afin d'obtenir des données chiffrées. Ce genre d'outil se révèle utile à l'orthophoniste. En effet les tests et batteries étalonnés sont essentiels car ils apportent au professionnel des mesures objectives et lui permettent notamment de situer le patient par rapport à ses pairs (INSERM, 2007 ; Casalis, Leloup et Bois Parriaud, 2013). En objectivant un score déficitaire à une épreuve le test atteste la plainte du patient et offre des arguments précieux au diagnostic (Launay, 2015 ; Perdrix, 2015). Les données objectives apportent en conséquent du crédit à l'évaluation qui dès lors ne se limite plus à l'observation subjective du professionnel.

Les outils d'évaluation normés permettraient également, selon des orthophonistes, de légitimer la profession en apportant une certaine rigueur scientifique. Le test viendrait alors tracer la frontière entre « pédagogie » et « pathologie » et délimiter le champ d'intervention de l'orthophoniste dont les soins sont pris en charge par la sécurité sociale (Auzias et Le Garsmeur, 2012).

Des critiques sont cependant manifestées à l'égard des outils étalonnés. Ces critiques sont principalement de deux sortes :

- 1) Relatives à l'utilisation des outils d'évaluation normés en elle-même
- 2) Relatives aux résultats obtenus par ces outils

Concernant le premier point, certains professionnels craignent que l'utilisation de données chiffrées se fasse au détriment de l'évaluation qualitative et de la démarche clinique ; voire que le quantitatif se substitue au qualitatif (Ferrand, 2007 ; Auzias et Le Garsmeur, 2012). A cela il est possible de répondre que le test n'est ni plus ni moins qu'un outil à disposition de l'orthophoniste (Lederlé et Maeder, 2016). Le regard clinique du praticien pourra porter sur les résultats fournis par le test. Ces résultats ne feront sens que parce qu'ils auront été interprétés par le professionnel et remis dans le contexte particulier de l'évaluation d'un patient précis. Il faut donc « faire parler les tests » (Perdrix, 2016). Les démarches qualitative et quantitative en orthophonie sont complémentaires et vont s'alimenter l'une l'autre (Ferrand, 2007 ; Hernandez et Ribaga, 2010 ; Witko, 2013 ; Leclercq et Veys, 2014). Ces outils d'évaluation ne constituent donc pas en eux-mêmes une menace pour l'analyse clinique puisque cette compétence clinique de l'orthophoniste portera également sur les résultats chiffrés obtenus. L'utilisation des tests et batteries doit se faire de manière sensée, elle s'inscrit dans une démarche clinique plus générale (Perdrix, 2016) où la réflexion du praticien doit primer. L'évaluation est une partie seulement de la démarche diagnostique qui doit être hypothético-déductive (Launay, 2015). La corroboration des éléments qualitatifs et quantitatifs va asseoir la crédibilité du diagnostic posé.

L'utilisation d'outils normés doit aussi dépendre de l'objectif du professionnel : l'évaluation normée n'est pas la plus adaptée à chaque situation. Si elle est intéressante pour attester de la présence d'un trouble, elle ne doit pas être utilisée pour évaluer l'efficacité de la thérapie ou définir les objectifs thérapeutiques (Cattini et Clair-Bonaimé, 2017).

Au sujet du second point de nombreux auteurs/orthophonistes, avertissent les professionnels quant au poids à accorder à l'évaluation chiffrée :

D'une part, les résultats obtenus doivent être analysés en regard du contexte clinique du patient. Une autre cause qu'une compétence faible dans la mesure visée peut expliquer un

score chuté à l'épreuve : le résultat ne rendra pas compte des capacités réelles du patient (McCauley et Swisher, 1984a ; Marin-Curtoud, Rousseau et Gatignol, 2010). Tenir compte du caractère artificiel de la situation d'évaluation et de la réalité clinique du patient est indispensable : il peut refuser l'épreuve, être fatigué, avoir mal compris la consigne (Lerderlé et Maeder, 2016). De même, afin de juger de la sévérité d'un trouble par exemple il apparaît préférable de se fier aux répercussions de ce trouble dans le quotidien du patient plutôt qu'aux résultats fournis par un test (Spaulding, 2011 ; Launay, 2015). Le patient et sa plainte doivent primer sur les chiffres obtenus. Les outils d'évaluation, bien qu'utiles, restent liés et soumis à des situations cliniques singulières.

D'autre part, les propriétés et la construction de certains outils sont sujettes à des interrogations de la part des professionnels. Les résultats d'un test sont donc à prendre avec un certain recul. Il arrive en effet qu'un outil d'évaluation annonce un objectif, comme déterminer un degré de sévérité ou poser un diagnostic, et ne soit en réalité pas apte à le remplir. Si la construction d'un outil est discutable alors les résultats obtenus par cet outil le seront aussi.

Afin de répondre aux limites de l'évaluation normalisée ou de compléter celle-ci, d'autres méthodes d'évaluation sont applicables en orthophonie : l'évaluation critériée et l'évaluation dynamique (McCauley et Swisher, 1984a ; Cattini et Clair-Bonaimé, 2017).

Les tests et batteries présentent donc certaines limites. En être conscient permet d'en avoir une utilisation éclairée et de ne pas leur accorder plus de poids que nécessaire ; c'est aussi une question de responsabilité.

Evaluation et responsabilité

Laveault et Grégoire expriment la notion d'une double responsabilité : celle de l'auteur du test et celle du praticien utilisant le test. L'auteur doit s'assurer que le manuel fourni avec le test comporte suffisamment d'indications précises à sa bonne passation. Il a également la responsabilité de préciser les qualifications requises pour l'utilisation du test (Laveault et Grégoire, 2014). En effet, un professionnel qui n'aurait pas les connaissances nécessaires à la bonne interprétation des résultats obtenus par l'évaluation effectuerait de

mauvaises conclusions. La seconde responsabilité est celle de l'utilisateur du test : le professionnel doit avoir lu le manuel du test et s'assurer de la bonne application des consignes de passation. Si la procédure de passation n'est pas respectée alors les résultats obtenus sont faussés. Une mauvaise utilisation d'un outil d'évaluation aura pour conséquences une compréhension erronée du problème du patient ainsi que l'établissement d'une thérapie inappropriée (McCauley et Swisher, 1984a). Au-delà de cette responsabilité de l'auteur du test d'une part et de son utilisateur d'autre part, une troisième responsabilité peut être questionnée : celle de l'éditeur du test. En février 2018 environ 250 orthophonistes manifestent sur les réseaux sociaux leur étonnement et leur mécontentement suite à une découverte : du matériel spécialisé, dont des outils d'évaluation, est en vente et accessible à des non-orthophonistes. La Fédération Nationale des Orthophonistes et une maison d'édition concernée ont vite réagi en promettant de mettre en place des mesures de contrôle plus strictes (Annexe 1). Si dans le manuel l'auteur précise que seuls certains professionnels sont habilités à la bonne utilisation de l'outil, il apparaît alors important que l'éditeur s'assure que la vente soit exclusivement réservée à ces professionnels. La responsabilité d'une utilisation consciencieuse de l'outil d'évaluation est donc partagée par plusieurs acteurs, cela étant dans l'intérêt du patient.

L'orthophoniste a une responsabilité lors de l'utilisation de l'outil mais également lors du choix de cet outil : « afin de délivrer le meilleur soin possible, l'obligation de moyens des praticiens repose sur la vérification des instruments utilisés en situation clinique » (Witko, 2014, p.13).

Choisir les outils d'évaluation

L'orthophoniste est seul juge et responsable du choix de l'outil d'évaluation à utiliser. Dans cette décision l'anamnèse revêt une importance particulière : c'est à partir des éléments recueillis lors de celle-ci que l'orthophoniste pourra décider du test le plus adapté. La plainte, la demande, les mots du patient vont ainsi guider le professionnel vers l'outil le plus en adéquation avec les paroles formulées. Cependant, avant la rencontre avec le patient, il est un moment où le choix d'un outil peut être difficile : lors de son achat. Plusieurs critères vont en effet entrer en ligne de compte : le prix de l'outil, sa disponibilité, l'usage qu'en aura le

praticien (va-t-il investir dans un test onéreux si la majorité de sa patientèle n'entre pas dans la catégorie ciblée par le test ?), la possession ou non d'un outil similaire au préalable, etc. Pour Perdrix « au final, l'offre de tests à un âge donné est en soi un puissant facteur de tri » (Perdrix, 2016). L'orthophoniste choisit donc son outil en fonction de ce qui est disponible d'une part, et de critères divers d'autre part.

C'est à cet instant que les inventaires d'outils révèlent tout leur intérêt : pour que le choix soit réel, il faut avoir connaissance de ce qui est à disposition. Plusieurs inventaires ont déjà été réalisés et permettent au professionnel qui les consulte d'avoir connaissance du matériel existant pour l'évaluation en général ou dans un domaine particulier. Dans la littérature disponible en orthophonie, deux types d'inventaires des instruments d'évaluation peuvent être distingués :

- Ceux qui recensent les outils et apportent des éléments d'information (domaine, population visée, objectif(s), taille de l'échantillon, année)
- Ceux qui en font une analyse critique

Concernant la première catégorie :

- Inventaire des outils pour évaluer le langage oral et le langage écrit, France (Belot et Tricot, 2001).
- Inventaire des outils pour évaluer les troubles d'origine neurologique de l'adulte, France. (Calvarin, 2013).
- Inventaire des outils pour évaluer le langage oral et la parole, Québec (Monetta, Desmarais, MacLeod, St-Pierre, Bourgeois-Marcotte et Perron, 2016 ; Monetta et tous les membres de l'axe 4 du REPAR, 2016).
- Liste d'outils utilisés par les orthophonistes français entre 2007 et 2011 (Brin-Henry, 2016). Ce travail-ci ne vise pas l'exhaustivité mais s'avère intéressant car il rend compte de ce qui est réellement utilisé sur le terrain par les professionnels.

Ces inventaires-ci ont pour objectif principal de fournir un catalogue aux orthophonistes et de répondre à cette demande de la profession (Belot et Tricot, 2001). Ils permettent également de faire une capture à un temps donné de ce qui existe et est utilisé réellement (Brin-Henry, 2016). Disposer d'inventaires à intervalles réguliers permet d'avoir connaissance de l'évolution des pratiques d'évaluation de la profession au fil du temps.

Concernant la seconde catégorie, à la différence des inventaires précédents cette fois les outils sont « décortiqués » et commentés. Il y a derrière ces études (commentées en p.12) l'idée de s'intéresser à la qualité des tests et batteries utilisés en orthophonie. Un constat commun à ces études est fait : celui du manque d'outils et d'un besoin de la profession d'avoir des outils répondant à ses attentes (McCauley et Swisher, 1984b ; Plante et Vance, 1994 ; Gaul Bouchard, Fitzpatrick et Olds, 2009 ; Friberg, 2010 ; Lafay, Saint-Pierre et Macoir, 2014 ; Leclercq et Veys, 2014).

S'intéresser aux tests et batteries disponibles présente les intérêts suivants :

- Mettre en évidence l'absence d'outils pour une tranche d'âge ou un domaine donné (Lafay *et al.*, 2014).
- En conséquence du point précédent, enclencher le développement de test et batteries pour répondre au besoin (Boudreault, Cabirol, Trudeau, Poulin-Dubois et Sutton, 2007).
- Argumenter auprès des concepteurs et éditeurs de tests la nécessité de faire figurer certaines informations dans les manuels des tests (McCauley et Swisher, 1984b). Rendre ces informations disponibles avant l'achat serait encore mieux (Cattini et Clair-Bonaimé, 2017).
- Les études réalisées aux Etats-Unis permettent d'argumenter en faveur de l'achat d'un test auprès des districts. Pour des raisons de budget il est courant que l'orthophoniste fasse avec ce qui lui est fourni. En France, il est envisageable que les études aient le même intérêt dans le cadre de l'exercice salarié : pouvoir démontrer qu'un outil est nécessaire constitue un argument en faveur de son financement.
- Informer les professionnels sur les outils qu'ils emploient.

Connaître les outils d'évaluation

Avoir une bonne connaissance des outils d'évaluation, de leurs limites et de la façon dont ils sont construits est nécessaire (Marin-Curtoud, Rousseau et Gatignol, 2010 ; Perdrix,

2016). C'est aussi un pas de plus vers la pratique fondée sur les preuves (Evidence Based Practice - EBP), de plus en plus présente dans le domaine de la santé et donc de l'orthophonie (Durieux, Pasleau, Vandenput et Maillart, 2013). L'EBP vise à aider le professionnel lors de la prise d'une décision par l'appui sur les meilleures preuves de la recherche scientifique. En ce sens, il est important pour les orthophonistes de connaître les outils d'évaluation et la façon dont ils sont construits (Skarakis-Doyle, Miller et Reichheld, 2000 ; Leclercq et Veys, 2014). Dans le cadre de l'évaluation, l'EBP réside aussi dans la considération du niveau de preuve afin d'évaluer la validité des résultats (Witko, 2013).

Ainsi, pour être dans l'EBP il est nécessaire de s'intéresser aux qualités psychométriques des tests et batteries (Gaul Bouchard *et al.* 2009).

Or, une étude sur les facteurs influençant le choix d'un outil standardisé plutôt qu'un autre a montré que seule l'année de publication du test était significativement corrélée avec sa fréquence d'utilisation. (Betz, Eickhoff et Sullivan, 2013).

Psychométrie et perspectives en orthophonie

Définitions

Pour l'étude des tests et batteries d'évaluation le domaine de la psychométrie offre une grille de lecture permettant l'analyse des outils :

La psychométrie est un domaine d'étude qui concerne la théorie et la méthodologie de construction et d'utilisation des échelles de mesure des caractéristiques mentales. Celles-ci peuvent être des connaissances, des traits de personnalité ou encore des composantes de la cognition, comme le langage, l'intelligence ou la mémoire (Grégoire pour Encyclopædia Universalis).

Lorsque l'on se penche sur l'étude des outils d'évaluation, des notions clés sont à définir :

- **Fidélité.** L'objectif d'un test va être de donner un score reflétant une compétence non-mesurable autrement. Mais le score obtenu n'est qu'un reflet de la réalité (score vrai) : il comporte une part d'erreur de mesure. La fidélité recouvre des notions de précision et de stabilité : le score obtenu est

proche du score vrai, le score obtenu par un individu est le même à un temps $t1$ et à un temps $t2$. Plus l'outil est fidèle, plus il reflète le score vrai.

- **Validité.** Il s'agit de la capacité de l'outil à mesurer ce qu'il est effectivement supposé mesurer.
- **Sensibilité et spécificité.** Ces deux notions vont de pair. La **sensibilité** correspond à la capacité d'un outil à détecter correctement les personnes porteuses de la caractéristique étudiée, il s'agit du taux de vrais positifs : les individus repérés sont bien porteurs de la caractéristique d'intérêt. La **spécificité** quant à elle correspond au taux de vrais négatifs : les personnes non-porteuses de la caractéristique étudiée sont bien identifiées comme ne portant pas cette caractéristique.

Le score seuil est lié à la sensibilité et à la spécificité (figure 1). Il s'agit d'une frontière établie afin de réduire au minimum le risque de négliger des individus ayant besoin d'être aidés (faux négatifs) et de prendre en charge des individus qui n'en ont pas besoin (faux positifs). « La notion de seuil diagnostique est d'ailleurs cruciale » (INSERM, 2007) : un seuil basé sur des données permet d'éviter le sur-diagnostic et le sous-diagnostic, à l'origine de conséquences sociales et économiques (Merrell et Plante, 1997).

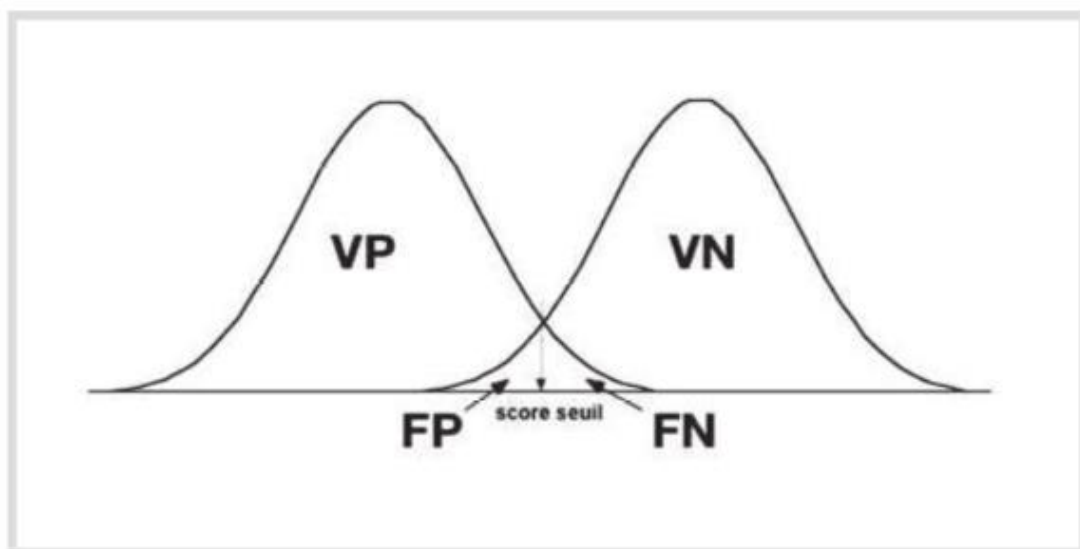


Figure 1 : Proportions de vrais positifs(VP), vrais négatifs(VN), faux positifs(FP) et faux négatifs(FN) en fonction du score seuil choisi

(Laveault et Grégoire p.275, 2014)

Cette notion de seuil revêt toute son importance dans l'étude des outils d'évaluation en orthophonie. En effet, elle est largement employée par la profession car c'est à partir de ce seuil que sera déterminé ce qui constitue un « score pathologique ». Par exemple pour une épreuve donnée ce seuil peut être fixé à -1,5 écart-type ; l'écart-type étant une mesure de dispersion des données autour de la moyenne. Tout patient ayant un score égal ou inférieur à ce seuil de -1,5 écart-type sera dans ce cas considéré comme présentant un « score pathologique ».

Les résultats obtenus au cours d'études en orthophonie sur ce sujet posent question. Un test est supposé être en capacité de discriminer les sujets avec un déficit des sujets qui n'en ont pas (Merrell et Plante, 1997). Or, la réalité est autre : il apparaît que le seuil offert par les manuels de tests n'est pas toujours le plus pertinent :

En étudiant les scores obtenus par des enfants et en y appliquant des seuils différents (-1 écart-type, -1,28 écart-type, -2 écarts-types) une étude a montré que le seuil de -2 écarts-types, bien que le plus utilisé, est le plus inapproprié. Lors de cette étude le seuil de -1 écart-type (ou équivalent du 16^{ème} percentile) apparaît comme la meilleure combinaison entre sensibilité et spécificité pour identifier un « primary language impairment », terme à rapprocher de celui de « dysphasie » (Thordardottir *et al.*, 2011).

De nombreux professionnels pensent qu'un même seuil convient à tous les tests, c'est faux : le score seuil doit être spécifique au test dont il est question (Greenslade, Plante et Vance, 2009). Certains états américains ont déterminé des seuils particuliers à adopter : il a été montré que ces seuils n'étaient pas en adéquation avec ceux donnés dans les manuels des tests (Spaulding, Swartout Szulga et Figueroa, 2012). Si des prises en charge, des aides ou des aménagements sont soumis à l'obtention d'un certain score à des tests précis cela pose des questions éthiques.

La sensibilité et la spécificité d'un outil vont dépendre du seuil donné, c'est pourquoi cette notion est importante (Shamahmood, Jalaie, Soleymani, Haresabadi et Nematti, 2016). Dans le champ de la recherche clinique en orthophonie, l'étude de la sensibilité et de la spécificité des outils est une « priorité » (Witko, 2014, p.14). Une étude réalisée sur 43 tests utilisés aux Etats-Unis pour diagnostiquer un trouble spécifique du langage (dysphasie) a montré que seuls 9 des 43 tests précisaient leurs sensibilité et spécificité au trouble (Spaulding, Plante,

Farinella, 2006). Lorsqu'un outil est présenté comme permettant d'établir un diagnostic il est indispensable que des informations concernant les capacités diagnostiques réelles de cet outil soient fournies (Merrell et Plante, 1997 ; Spaulding, *et al.*, 2012 ; Shamahmood *et al.*, 2016). Or, les études réalisées sur les tests et batteries en orthophonie vont dans le sens d'un manque à ce niveau là. Les notions de sensibilité et spécificité sont souvent absentes des manuels de tests (Leclercq et Veys, 2014). Le même constat est fait par ANAES (2002, p.52) en France : « des travaux restent à faire » sur l'évaluation de la sensibilité et de la spécificité.

Echantillon de normalisation

La constitution et les caractéristiques du groupe à partir duquel a été normée l'épreuve est liée à ces notions de sensibilité et spécificité. En effet, si le test a une visée diagnostique il est intéressant de savoir si des sujets présentant la pathologie que le test souhaite diagnostiquer ont été inclus, ou non, dans l'échantillon de normalisation. L'échantillon doit donc être en cohérence avec l'objectif du test et celui du praticien (Peña, Spaulding et Plante, 2006 ; Spaulding *et al.*, 2012). Ainsi, l'inclusion d'enfants avec un trouble spécifique du langage dans le groupe réduit la précision du diagnostic, or 19 tests sur 32 étudiés les ont inclus (Peña *et al.*, 2006). Cependant si l'objectif est de déterminer la sévérité plutôt que la présence d'un trouble, il peut être pertinent que l'échantillon de référence soit composé d'enfants présentant effectivement ce trouble. Les résultats devront être interprétés avec prudence, il est préférable de juger de la sévérité d'un trouble par l'impact de celui-ci dans la vie quotidienne que par le résultat à un test (Spaulding, 2011).

Disposer d'informations sur la constitution de l'échantillon de référence permet également de savoir si l'outil est adapté pour évaluer le patient. Les différences culturelles et linguistiques doivent être prises en compte, c'est particulièrement vrai dans l'évaluation du langage et donc en orthophonie. Au-delà d'une simple traduction des outils d'une langue à une autre, c'est l'adaptation de l'outil qui doit être réalisée (Laveault et Grégoire, 2014). La problématique est très présente au Canada francophone où peu d'outils spécifiques à cette population ont été développés et où la plupart des outils sont issus de maisons d'édition anglophones (Gaul Bouchard *et al.*, 2009 ; Thordardottir, Keheyia, Llessard, Sutton et Trudeau, 2010 ; Monetta *et al.* 2016).

Au bout de quelques années la question de la pertinence de l'emploi d'un test va se poser. Les langues et les populations évoluant, l'échantillon de l'époque reflète-t-il encore la norme actuelle ? Utiliser un test « trop » vieux pourrait constituer une violation du code d'éthique de l'American Speech-Language-Hearing Association (ASHA) pour Jakubowitz et Schill : le professionnel doit argumenter en quoi l'emploi d'une ancienne version d'un test reste pertinente (Jakubowitz et Schill, 2008). La notion de responsabilité du professionnel est à nouveau questionnée.

Etudes menées en orthophonie et constats

Plusieurs études sur les outils d'évaluation standardisés ont eu lieu dans le champ de l'orthophonie. L'une d'elles révèle par exemple qu'au niveau francophone du Canada les orthophonistes (n=282) sont plutôt insatisfaits quant aux éléments psychométriques des outils disponibles (Garcia, Paradis, Sénécal et Laroche, 2006).

Dès 1984 aux Etats-Unis McCauley et Swisher ont étudié 30 tests destinés à l'évaluation du langage et de l'articulation et les ont soumis à 10 critères psychométriques (McCauley et Swisher, 1984b). La conclusion était celle d'un échec à remplir les critères définis. Plusieurs manuels de tests ne comportaient aucune information sur la validité et la fidélité.

En 1994, les critères utilisés par McCauley et Swisher ont été repris et appliqués à 21 tests évaluant les compétences langagières (Plante et Vance, 1994). Dix ans après l'étude précédente, la conclusion faite fut la même : de nombreux tests échouent à l'obtention de critères satisfaisants.

Concernant l'évaluation francophone au Canada, une étude critique de 31 outils d'évaluation a été publiée en 2009 (Gaul Bouchard *et al.*, 2009). Les tests ont été soumis cette fois-ci à 16 critères psychométriques. La conclusion des auteurs met en évidence l'écart entre la qualité actuelle des outils d'évaluation et celle souhaitée. Les résultats sont même qualifiés « d'alarmants » (Gaul Bouchard *et al.*, 2009, p.136).

En 2010 et de nouveau aux Etats-Unis Friberg s'appuyant également sur la grille d'analyse proposée par McCauley et Swisher a étudié 9 tests évaluant le langage de l'enfant. Ces tests ont été sélectionnés en amont, avant d'être soumis aux critères, pour leurs niveaux

acceptables d'identification. En conclusion, tous les outils remplissent au moins 8 des 11 critères retenus (Friberg, 2010).

En 2014 et cette fois-ci en Belgique, Leclercq et Veys reprenant aussi les critères de McCauley et Swisher étudient 5 tests standardisés dans le cadre du diagnostic de dysphasie (Leclercq et Veys, 2014). A nouveau les résultats sont décevants, certains critères ne sont remplis par aucun test.

La même année, un inventaire critique des outils d'évaluation des habiletés mathématiques est réalisé (Lafay *et al.*, 2014). Cet inventaire, non orienté sur l'application de critères psychométriques, n'en est pas moins intéressant : les outils d'évaluation sont commentés de manière critique et détaillée. Ce travail permet de visualiser la complémentarité entre certains instruments d'évaluation : quel outil est disponible pour évaluer telle habileté mathématique. Il permet aussi de pointer des manques et besoins : notamment l'absence de données normatives pour les temps de réponse.

En effet, les mesures temporelles en orthophonie sont utiles : le temps d'exécution pour une tâche telle que la lecture est un indicateur précieux dont il faut tenir compte (Launay, 2015). Il arrive souvent que les épreuves ne soient pas chronométrées et cela peut empêcher l'orthophoniste de rendre objectifs des déficits (Casalis, Leloup et Bois Parriaud, 2013). Le développement d'épreuves informatisées est bénéfique en ce sens car celles-ci permettent une meilleure mesure du temps. (Marin-Curtoud, Rousseau et Gatignol, 2010).

Les différentes études réalisées, qu'elles soient francophones ou anglophones permettent de mieux connaître les outils d'évaluation ainsi que leurs limites.

L'évaluation en orthophonie est un domaine majeur relevant de multiples notions. Elle se situe au carrefour d'obligations, de questionnements sur les pratiques professionnelles et de la recherche d'une pratique fondée sur les preuves. La connaissance des instruments d'évaluation permet un recul nécessaire quant à leur utilisation et à l'interprétation des résultats émanant de ces tests. La psychométrie offre des notions utiles à l'étude des instruments d'évaluation de l'orthophoniste. Les études déjà menées sur les tests et batteries standardisés mettent en évidence les besoins de développer de nouveaux outils d'une part et de s'assurer de la qualité de ceux-ci d'autre part. C'est dans cette démarche d'évaluation des outils que s'inscrit le travail qui suit.

Méthode

Recensement des outils d'évaluation

Dans un premier temps il a fallu répertorier les outils d'évaluation à disposition des orthophonistes pour l'évaluation du langage écrit. Pour cela une recherche a été effectuée sur différentes plateformes dans un premier temps, puis, dans un second temps les données recueillies ont été croisées et les doublons ont été supprimés. L'inventaire des outils d'évaluation se voulant exhaustif une recherche a été menée :

- sur les sites internet des éditeurs d'outils d'évaluation
- sur le répertoire de la testothèque du CFUO de Nantes
- à partir de l'ouvrage « Tests en orthophonie : tome 1, langage oral-langage écrit, enfants adolescents » (Belot et Tricot, 2001).
- à partir de la thèse de Brin-Henry (2011) portant sur 435 comptes rendus de bilans orthophoniques
- à partir d'une liste de tests disponible sur le site de l'UNADREO (UNADREO, 2011)

Suite à la confrontation des données et à la suppression des doublons, une étape de tri a eu lieu. Parmi les outils recensés certains n'ont pas été retenus dans le cadre de ce travail :

- les outils d'approche uniquement qualitative
- les outils exclusivement utilisés dans le cadre scolaire (enseignant, médecin scolaire, RASED)
- les tests et batteries évaluant le langage écrit dans le cadre particulier des accidents vasculaires cérébraux, traumatismes crâniens, pathologies neurodégénératives

Cependant, les outils accessibles aux orthophonistes et effectivement utilisés par ceux-ci sur le terrain ont été conservés sur la liste; que ces outils soient accessibles à toute personne ou qu'ils soient partagés avec d'autres professions.

Les batteries évaluant plusieurs domaines et contenant des subtests visant le langage écrit ont été incluses.

Au final 42 outils (tests et batteries) ont été retenus (Annexe 2).

Matériel

Parmi les 42 outils répertoriés, 22 d'entre eux ont pu être étudiés et comparés. Les batteries étudiées par Leclercq et Veys et qui comportent des subtests relatifs au langage écrit ont été incluses dans la comparaison (Leclercq et Veys, 2014).

En raison de contraintes matérielles (outils peu ou plus utilisés, difficiles à se procurer), et temporelles (temps de recherche des outils, temps d'analyse des outils), l'ensemble des 42 outils répertoriés n'a pu être étudié. Cependant, la comparaison de 22 outils offre déjà certains constats intéressants.

L'étude a été menée à partir des manuels d'utilisation des outils. Les documents annexes ont été pris en compte : une analyse statistique du test est parfois présente dans le lot de documents fournis mais n'est pas insérée dans le manuel d'utilisation à proprement parler. Ainsi, tous les documents fournis au professionnel lors de l'achat ou du téléchargement gratuit de son outil ont été considérés.

Choix et définitions des critères psychométriques retenus pour l'analyse

Dans l'idée d'une uniformisation des procédures d'étude des outils d'évaluation en orthophonie, ce travail s'appuie sur les critères utilisés par McCauley et Swisher (1984b), Plante et Vance (1994), Gaul Bouchard *et al.* (2009), Friberg (2010) et Leclercq et Veys (2014). Aux 10 critères initialement établis par McCauley et Swisher Friberg a ajouté celui d' « objectif du test identifié ». Par la suite, Leclercq et Veys (2014) ont ajouté les critères de « validité théorique » et de « sensibilité et spécificité ». Il est intéressant de noter que dans l'étude de Friberg (2010), ce dernier point était un prérequis avant même de soumettre les outils aux autres critères.

Ainsi les 13 critères retenus sont les suivants :

- **1- Qualifications de l'évaluateur spécifiées**

Afin d'assurer la validité des résultats il est nécessaire que le manuel rende explicites les qualifications requises pour l'administrateur du test. L'utilisateur du test s'occupe de la passation, de la cotation et de l'interprétation du test et de ses résultats.

Pour valider ce critère le manuel doit indiquer le ou les professionnels à qui l'outil est destiné. En cas de mention absente ou imprécise (mention de plusieurs professions suivie de « etc. » ou « ... ») le critère n'est pas validé.

- **2- Consignes de passation et de cotation**

Des consignes de passation et de cotation doivent être indiquées. L'utilisateur doit effectuer la passation dans des conditions similaires à celles dans lesquelles l'étalonnage a été réalisé. Des consignes précises contribuent à la garantie d'une bonne fidélité du test. En effet, il est important que les consignes soient suffisamment claires pour que deux utilisateurs différents cotent les résultats de la même manière. La précision de ces consignes facilitera une bonne fidélité inter-juges.

Le critère est validé si des consignes sont présentes sur papier dans le manuel ou informatiquement.

- **3- Objectif du test précisé**

Le ou les objectifs du test doivent être indiqués. Idéalement, l'objectif devrait être clair et le test pensé et conçu pour y répondre. En consultant le manuel l'utilisateur doit alors savoir si l'objectif visé par l'outil est en adéquation ou non avec le sien pour l'évaluation du patient.

Ce critère est validé par la mention d'un ou plusieurs objectifs.

- **4- Sensibilité et spécificité**

Ces données sont particulièrement importantes dans le cas d'un objectif de diagnostic. Friberg recommande un taux de 80% ou plus d'identification précise avant d'envisager l'utilisation d'un outil dans un but diagnostique (Friberg, 2010).

Pour valider le critère le manuel doit apporter des données chiffrées suffisantes appuyant la recherche par les auteurs de la sensibilité et spécificité de l'outil.

○ 5- Taille de l'échantillon d'étalonnage

Une taille minimale de 100 sujets par sous-groupe est nécessaire pour valider ce critère.

Si un seul des sous-groupes (par exemple pour une classe ou un âge donné) n'atteint pas cette taille alors le critère n'est pas validé.

○ 6- Informations sur l'échantillon d'étalonnage

L'utilisateur de l'outil doit disposer de données concernant l'échantillon à partir duquel des normes ont été établies. Cela lui permet de s'assurer que le patient peut bien être comparé à l'échantillon de référence. Il est donc intéressant de connaître les données suivantes à propos de l'échantillon :

- Caractéristiques géographiques
- Caractéristiques socio-économiques
- Age
- Genre
- Caractéristiques linguistiques (langue parlée, bilinguisme)
- Informations sur la constitution de l'échantillon (critères d'inclusion/exclusion, sujets avec un diagnostic, sujets « sains »)

Le critère est validé en présence d'au moins trois informations parmi les six citées ci-dessus. L'objectif ici est d'estimer si l'outil procure assez d'informations ou non au professionnel pour juger l'échantillon. Il faut noter que ce n'est pas parce qu'un outil informe du détail de son échantillon que cela signifie que l'échantillon est bien représentatif de la population générale. De plus, un outil étalonné sur une population précise et non-représentative de manière générale peut très bien être utilisé si le patient y correspond et que cela est adapté à l'objectif d'évaluation défini.

○ 7- Mesures de tendance centrale

Les auteurs de l'outil doivent apporter des informations telles que les moyennes et écarts-types ou les centiles de l'échantillon de référence.

Un âge-lecture n'a pas été considéré comme permettant de valider le critère.

Validité :

○ 8- Validité théorique

L'outil d'évaluation doit être en lien avec des fondements théoriques.

Le critère est validé si le manuel présente des appuis sur la littérature, des références à des modèles théoriques.

○ 9- Validité concourante

La validité concourante (ou concomitante) correspond à la corrélation établie entre le score issu d'un outil d'évaluation et les résultats obtenus grâce à un autre outil.

Le critère est validé si le manuel mentionne la validité concourante et apporte des données chiffrées.

○ 10- Validité prédictive

La validité prédictive concerne le lien entre les résultats obtenus par l'outil dont il est question et une performance, un résultat futur dans le quotidien du sujet évalué.

Le critère est validé si le manuel mentionne une recherche de validité prédictive et apporte des données chiffrées.

Fidélité :

○ 11- Fidélité test-retest

La fidélité test-retest indique la stabilité des résultats : les scores obtenus sont similaires à un temps $t1$ et à un temps $t2$.

Le critère est validé si le coefficient de corrélation $r = .90$ est atteint.

○ 12- Fidélité inter-juges

Les résultats obtenus à un test ne doivent pas dépendre de la personne qui effectue la passation et la cotation du test. La subjectivité liée à l'utilisateur de l'outil doit être limitée. Des consignes de passation et de cotation précises et détaillées contribuent à limiter la variabilité inter-juges.

Le critère est validé si le coefficient de corrélation $r = .90$ est atteint.

○ 13- Analyse des items

Les items inclus dans le test ou les subtests doivent avoir été étudiés lors de la conception de l'outil : validité des items, difficulté des items, analyse de la cohérence interne. La cohérence (ou consistance) interne est relative au degré avec lequel les items d'un test mesurent bien la même caractéristique.

Des données chiffrées suffisantes doivent être présentes pour permettre à l'outil de remplir ce critère.

Les critères ont été considérés comme ayant tous le même poids : chacun d'eux rapporte un point s'il est validé. Ainsi chaque test peut atteindre un maximum de 13 points. Cependant dans certains cas il peut être plus difficile de trancher quant à la réussite ou non d'un critère : en effet un élément est parfois présent dans le manuel mais insuffisamment détaillé. Il est alors partiellement rempli. Dans ce cas un demi-point seulement est accordé. Cela permet d'affiner les résultats obtenus et de marquer des nuances entre les outils d'évaluation. Ce système de cotation permettant le demi-point a été utilisé dans l'étude de Leclercq et Veys, il semblait donc pertinent d'employer le même système afin de permettre des comparaisons entre les outils (Leclercq et Veys, 2014).

Chaque critère permet donc au test étudié d'obtenir de 0 à 1 point :

- 0 point : le critère n'est pas validé
- 1 point : le critère est validé
- 0,5 point : des données sont présentes mais n'ont pas été considérées suffisantes

A ces 13 critères deux éléments supplémentaires ont été ajoutés à titre indicatif :

- la tranche d'âge concernée par l'outil
- l'année de publication de l'outil ou, idéalement, l'année de standardisation de l'échantillon de normalisation

Ce critère de date fait partie des 16 critères retenus dans l'étude de Gaul Bouchard *et al.* (2009). Les auteurs indiquent d'ailleurs que les normes datant de plus de cinq ans sont « inappropriées et désuètes » : la population et la langue évoluent (Gaul Bouchard *et al.*, 2009, p.139). Il apparaît cependant difficile de trancher et de donner une date claire de

« péremption » de l'outil d'évaluation, date au-delà de laquelle l'utilisation de l'outil ne serait plus pertinente. En effet, Jakubowitz et Schill laissent entendre que selon la situation l'emploi d'un test « hors-date » peut être justifié (Jakubowitz et Schill, 2008). Cependant, ces auteurs expriment aussi que les compétences de l'orthophoniste utilisant un test trop vieux pourraient légitimement être remises en question, notamment lorsqu'une version plus récente d'un même outil est disponible. La question de la date des outils est délicate et dépend plus de ce qui est disponible sur le marché que du professionnel lui-même.

Dans le cadre de ce travail la date a donc été traitée selon un gradient :

- $\text{Date} \leq 5 \text{ ans}$
- $5 \text{ ans} < \text{Date} \leq 10 \text{ ans}$
- $\text{Date} > 10 \text{ ans}$

Lorsque la date de publication d'un outil et la date à laquelle les sujets de l'échantillon ont été évalués ne se situent pas dans la même catégorie de ce gradient, nous rappelons que celle qui a le plus d'intérêt est celle de l'échantillon.

Traitement des données

Participants au traitement des données

Ce travail, réalisé dans le cadre du mémoire d'orthophonie, entraîne le fait qu'une seule personne a soumis les outils aux critères et a jugé si ceux-ci les remplissaient ou non. Or, dans le cas d'autres études du même type plusieurs personnes évaluaient un même test (Friberg, 2010). Si certains critères sont simples à coter (taille de l'échantillon) et entraînent dans les études déjà réalisées un accord de 100% entre les chercheurs, d'autres sont plus complexes (analyse des items notamment) (Friberg, 2010). En effet pour la taille de l'échantillon le critère n'appelle pas à discussion puisque la frontière est nette entre « critère validé » et « critère non validé ». En revanche, certains critères nécessitent un apport suffisant d'éléments pour être considérés comme satisfaits. La notion d'apport suffisant peut légèrement varier d'une personne à l'autre. Ainsi, la manière de coter pourrait être discutée. La possibilité de donner un demi-point aux critères contribue à limiter cette variabilité.

Procédure de traitement

Le traitement s'est effectué en deux étapes. Dans un premier temps les outils ont été étudiés indépendamment les uns des autres et un score a été attribué à chacun. Puis, l'analyse a été conduite une seconde fois à distance, en comparant la manière dont chaque critère a été validé, semi-validé ou invalidé entre les outils. L'intérêt de cette procédure est de minimiser la variabilité intra-individuelle dans la cotation des outils. Un critère tel que « Analyse des items » pourrait avoir été considéré comme suffisant pour un test A à un temps t mais comme insuffisant pour un test B à un autre moment ; alors que dans la réalité ces deux tests pourraient répondre au critère de manière équivalente. Suite à cette seconde étape de cotation des ajustements ont été effectués. Cette seconde étape de cotation permet également une relecture et sert de procédure de contrôle visant à limiter toute erreur humaine dans l'attribution des points.

Mode de présentation des résultats

Afin d’obtenir des résultats lisibles, de permettre plus aisément une comparaison entre les outils et d’en avoir une vue d’ensemble, il a été décidé de présenter les résultats sous forme d’un tableau (tableau 1). Ce système de tableau a été utilisé dans les études de Leclercq et Veys (2014) et de Lafay *et al.* (2014), ayant estimé ce mode de présentation très pertinent nous avons opté pour le même principe.

Les critères sont étudiés par ligne et chacun des tests possède sa colonne, son score total figurant en bas de celle-ci. Les tests ont été soumis à la grille suivante :

Tableau 1 : Grille d’analyse des tests

	Outil d’évaluation
Tranche d’âge	
Année : publication (échantillon)	
Qualifications de l’évaluateur	
Consignes d’administration et de cotation	
Objectif du test précisé	
Sensibilité et spécificité	
Taille de l’échantillon d’étalonnage	
Informations sur l’échantillon d’étalonnage	
Mesures de tendance centrale	
Validité	
Validité théorique	
Validité concourante	
Validité prédictive	
Fidélité	
Fidélité test-retest	
Fidélité inter-juges	
Analyse des items	
Score	

En haut de la grille se trouvent les deux informations ne contribuant pas au score final : la tranche d'âge concernée et l'année de publication de l'outil. La date à laquelle l'étalonnage a eu lieu, quand elle est disponible, est indiquée entre parenthèse à côté de l'année de publication de l'outil. En cas d'indisponibilité de la date un point d'interrogation (?) est mis.

Un code couleur (figure 2) suivant le gradient évoqué pour ce critère de date a été mis en place.

2000	2010	2015
------	------	------

Figure 2 : Illustration du gradient de date. Blanc : datation de plus de 10 ans, jaune : datation de 10 ans ou moins, orange : datation de 5 ans ou moins

En cas de dissociation de date entre la publication et la date correspondant à l'échantillon, les dates ont été surlignées.

L'ordre des critères dans le tableau a été déterminé en fonction de la facilité supposée avec laquelle ils seraient remplis par les manuels et de l'ordre dans lequel on pourrait les rencontrer dans ces derniers. Ils ont également été rassemblés par thématique. Le critère de sensibilité et spécificité a été placé sous celui de l'objectif du test pour la raison suivante : si nous savons qu'un outil a un objectif de diagnostic alors il est intéressant de savoir s'il remplit les conditions de sensibilité et spécificité. Visuellement, avoir ces critères l'un sous l'autre est intéressant. La taille de l'échantillon, ses caractéristiques et les mesures de tendance centrale sont des notions liées et souvent proches géographiquement dans les manuels. Les critères de validité théorique, validité concourante et validité prédictive ont été rassemblés. Il en va de même pour les critères de fidélité test-retest, fidélité inter-juges et analyse des items. La dernière ligne du tableau donne le nombre de points obtenus par l'outil d'évaluation.

Pour rendre les résultats plus aisés à lire les cases du tableau ont été colorées lorsque les critères concernés étaient remplis (figure 3).



Figure 3 : Illustration de la cotation par critère. Blanc : non validé, 0 point, gris : partiellement validé, 0,5 point, noir : validé, 1 point.

Le recensement et l'étude des outils d'évaluation constituent un moyen pour les professionnels de connaître les outils qu'ils peuvent se procurer. Lorsqu'un outil n'est plus disponible à l'achat l'information est indiquée par un astérisque (*) à côté de son nom.

Diffusion des résultats

Au-delà d'un état des lieux sur les outils d'évaluation en langage écrit, ce travail a d'autres objectifs :

- Sensibiliser les orthophonistes à la lecture critique des manuels de tests
- Affirmer l'importance de la présence de certains critères dans ces manuels
- Par conséquent, inciter les concepteurs de tests à remplir ces critères et, dans l'idéal, les rendre accessibles avant l'achat d'un outil

Afin de tenter de remplir ces différents objectifs il semble utile de diffuser les résultats de ce travail. Les mémoires d'orthophonie pouvant tout à fait être perçus comme rébarbatifs à lire, notamment en raison de leur longueur, il apparaît plus pertinent et efficace de diffuser les résultats autrement. Ainsi un article issu de ce travail sera proposé à la publication.

Dans un premier temps une liste d'outils utilisés dans l'évaluation du langage écrit a été composée. Suite à l'application de critères d'exclusion 42 outils ont été retenus. Sur ces 42 tests évaluant le langage écrit répertoriés, 22 ont pu être étudiés et soumis à 13 critères relevant du domaine de la psychométrie. Ces critères ont été déterminés en prenant appui sur des études déjà effectuées en orthophonie. Chacun des critères psychométriques permet à l'outil d'évaluation d'obtenir 0, 1 ou parfois 0.5 point. Un outil peut obtenir jusqu'à 13 points. La date de constitution de l'échantillon d'étalonnage est elle aussi indiquée par un gradient fonction de l'ancienneté. La tranche d'âge concernée par le test est également spécifiée.

Les scores obtenus par les tests ainsi que le nombre de tests répondant à chacun des critères sont présentés dans la partie Résultats qui suit.

Résultats

Les résultats complets sont à consulter en Annexe 3.

Résultats globaux

Tout d'abord, aucun des outils ne parvient à valider les 13 critères retenus dans cette analyse (figure 4).

Le score minimal obtenu est de 1 point.

Le score maximal obtenu est de 8 points.

Le score moyen obtenu par l'ensemble des outils est de 6 points.

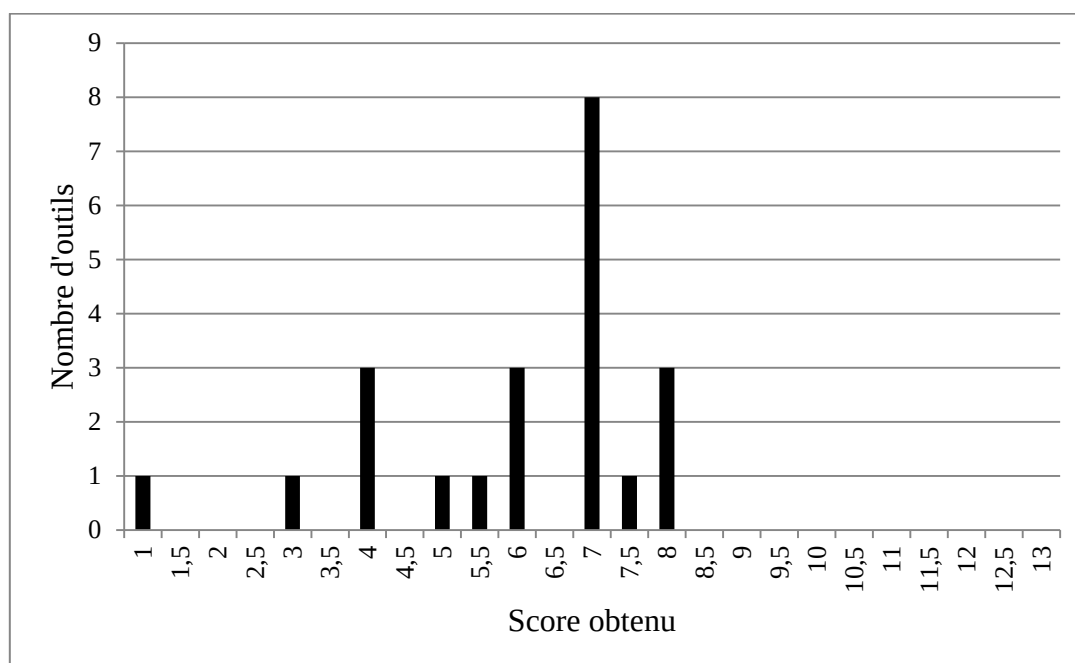


Figure 4 : Distribution des outils par score obtenu

Un seul outil obtient le score de 1 point, il s'agit de L'Alouette (Lefavrais, 1967). Trois outils atteignent le score de 8 points : le TCS, EXALang 8-11 et EXALang LyFac (Maeder, 2006 ; Thibault, Lenfant, Helloin, 2012 ; Thibault et Lenfant, 2014).

L'écart entre le plus petit score obtenu et le maximal atteint est donc de 7 points.

Au total, 13 critères ayant été retenus, un outil devrait avoir un score de 6,5 pour en valider théoriquement la moitié. Sur les 22 outils étudiés nous observons que 10 se situent sous ce seuil, soit 45% d'entre eux. Leurs scores vont de 1 à 6 points. Les 12 outils restants obtiennent tous un score compris entre 7 et 8 points : leur dispersion est moindre.

Ainsi, 55% des outils étudiés obtiennent un score allant de 7 points à 8 points.

Résultats par critère

Les résultats mettent en évidence des différences entre les différents critères : tous ne sont pas validés dans les mêmes proportions (figure 5).

Le critère le plus validé l'est par les 22 outils. A l'opposé, 4 critères sur les 13 étudiés ne sont validés par aucun outil.

Il y a donc une forte hétérogénéité dans la validation des critères par les outils.

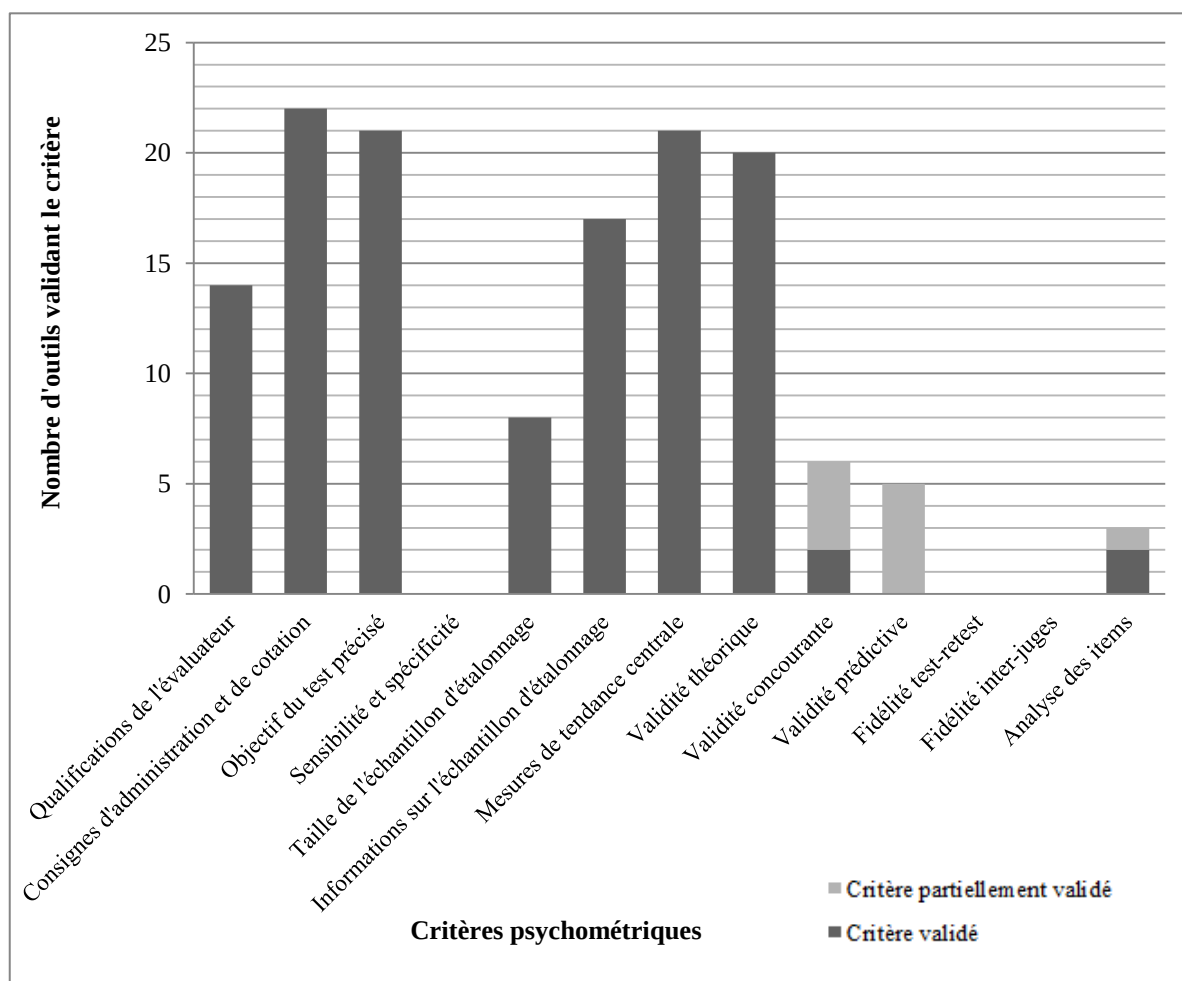


Figure 5 : Distribution du nombre d'outils par critère

Ces éléments sont majoritairement bien présents dans les manuels des outils d'évaluation. Il s'agit en effet de critères de base dont la présence est attendue. Ils sont simples à fournir et ne demandent aucune analyse statistique poussée aux concepteurs de l'outil. Leur absence pourrait être qualifiée d'étonnante (Leclercq et Veys, 2014). C'est le cas notamment du premier de ces critères :

- **Qualifications de l'évaluateur : 14 outils sur 22**

Il est assez surprenant de constater que certains outils ne comportent pas d'indication quant aux qualifications du professionnel à qui ils sont destinés. Il est possible d'imaginer différentes explications à ce phénomène :

- L'outil est en téléchargement libre et accessible à tous, quiconque pourrait être évaluateur : c'est le cas de l'outil E.L.FE. Il est simplement précisé que deux textes différents sont présents pour permettre « aux enseignants et aux chercheurs » de mesurer des progrès (Lequette, Pouget, Zorman, 2009, p.4)
- Cela pourrait-il être un oubli des auteurs ou simplement la non-considération de cet élément ?
- Les concepteurs avaient en tête certaines maisons d'édition spécialisées ou collections et il leur semblait tout à fait logique que le test soit réservé aux personnes qualifiées ?

Comme nous l'avons déjà évoqué, la précision des qualifications requises est importante et se rattache à une certaine notion de responsabilité.

Dans d'autres cas et bien que des professions soient citées, le critère n'a pas été validé car trop flou, laissant sous-entendre par des pointillés que d'autres professions pourraient être concernées : c'est le cas du test Le petit Poucet qui indique « orthophoniste, phoniatre... » (De Plazaola, Gauthier, Arsicaud, Pech-Georgel, 2003, p.7)

- **Consignes d'administration et de cotation : 22 outils sur 22**

Ce critère est validé pour tous les outils étudiés. Un test sans consignes n'aurait certainement pas grande utilité, ni grand succès auprès d'acquéreurs éventuels. Cependant

tous les outils ne répondent pas de la même manière à ce critère. Certains fournissent des indications très détaillées, d'autres moins et certains sont informatisés entraînant ainsi une cotation automatique des réponses.

- **Objectif du test précisé : 21 outils sur 22**

Un outil ne répond pas à ce critère, il s'agit de L'Alouette (Lefavrais, 1967). La question de la dyslexie est posée, la structure du test est présentée et ce dernier permet d'obtenir un niveau de lecture mais aucun objectif n'est formulé explicitement dans le manuel. Par conséquent le critère ne peut être validé.

En dehors de ce cas, l'objectif de l'outil est indiqué dans les manuels, ce qui semble assez logique puisqu'un outil doit présenter un intérêt pour être développé et donc prétendre répondre à un objectif fixé.

La capacité réelle du test à répondre à son ou ses objectifs n'a pas été jugée, il s'agit d'une question bien plus complexe. Nous laissons chacun libre de son jugement lors de la lecture des manuels.

Critères relatifs à l'échantillon

- **Taille de l'échantillon d'étalonnage : 8 outils sur 22**

Derrière un échec à ce critère se cachent de nombreuses disparités entre les outils. Par exemple, L'Alouette-R (Lefavrais, 2005) présente un nombre insuffisant de sujets sur de nombreux sous-groupes (de 18 à 49 individus par classe) tandis que l'EXALang 11-15 (Lenfant, Thibault, Helloin, 2009) échoue de peu à ce critère (85 à 97 individus), de même pour EMILIE (Duchêne, 2010). E.L.FE donne le nombre total d'enfants sur lesquels l'outil a été étalonné et, a priori, le seuil de n=100 par sous-groupe devrait être atteint (Lequette *et al.*, 2009). Mais en l'absence de précision il est impossible d'être certain du nombre d'individus par tranche. Le critère n'a donc pas pu être validé.

Au-delà du simple échec à ce critère, il reste intéressant pour le professionnel d'aller lire les manuels dans le détail. En effet dans le cas de l'EXALang 11-15 par exemple des études de normalité ont été effectuées sur les échantillons (Lenfant *et al.*, 2009).

- **Informations sur l'échantillon d'étalonnage : 17 outils sur 22**

Ce critère portant sur l'échantillon est mieux réussi que celui sur la taille. La plupart des outils fournit une description de la population ayant servi à l'étalonnage. Cette description comprend généralement les classes, l'âge, le sexe, le lieu (région, département, caractéristiques de l'école) dans lequel les passations ont été réalisées.

En fonction de l'objectif du test il est intéressant d'observer si l'échantillon comprend ou non des sujets diagnostiqués pour une pathologie et quels ont été les critères d'exclusion et d'inclusion lors de la constitution de l'échantillon. Une autre information, parfois présente, concerne les mois de l'année scolaire durant lesquels les passations ont été réalisées. En cas de doute sur le groupe auquel comparer le patient l'orthophoniste doit toujours se référer au manuel.

- **Mesures de tendance centrale : 21 outils sur 22**

La très grande majorité des outils procure à leurs utilisateurs des données quantitatives : moyenne, écart-type, rangs centiles. L'âge de lecture seul donné dans la version 1967 de L'Alouette (Lefavrais, 1967) n'est pas considéré comme permettant de valider ce critère.

Un élément complémentaire à vérifier est la distribution des données de l'échantillon. Si les données ne suivent pas une distribution normale, que leur distribution est asymétrique alors l'utilisation des rangs centiles est préférable.

Validité

- **Validité théorique : 20 outils sur 22**

Ce critère est bien validé par les différents outils étudiés. La majorité des manuels débute d'ailleurs par la présentation du cadre théorique dans lequel ils s'inscrivent et comportent des références à la littérature. Deux outils échouent à ce critère et ce pour des raisons différentes : d'une part L'Alouette (Lefavrais, 1967) et d'autre part E.L.FE (Lequette *et al*, 2009). Le manuel de L'Alouette apporte bien quelques références mais le critère ne peut être validé de nos jours (Lefavrais, 1967). Quant à E.L.FE, le critère est rejeté en raison d'une absence de données théoriques (Lequette *et al*, 2009).

- **Validité concourante : 2 outils sur 22**

Ce critère est très peu présent dans les manuels. Il est validé par la BHK et le TCS (Charles, Soppelsa, Albaret, 2004 ; Maeder, 2006). Quatre autres outils ont été considérés comme validant partiellement ce critère, il s'agit des EXALang : EXALang 5-8, EXALang 8-11, EXALang 11-15 et EXALang LyFac (Lenfant *et al.*, 2009 ; Thibault, Helloin, Croteau, 2010 ; Thibault, Lenfant, Helloin, 2012 ; Thibault et Lenfant, 2014). Il serait pourtant intéressant pour les professionnels d'avoir des données concernant les corrélations entre les tests. Si un outil existant est reconnu comme valide, une corrélation entre ses résultats et ceux d'un nouvel instrument constitue un argument en faveur de la validité du nouvel instrument.

- **Validité prédictive : 0 outils sur 22**

La validité prédictive est un point massivement échoué par les outils ayant été soumis à ce travail. Cependant 5 outils ont été considérés comme répondant en partie au critère : la BHK (Charles *et al.*, 2004) et les quatre batteries EXALang : 5-8, 8-11, 11-15 et LyFac (Lenfant *et al.*, 2009 ; Thibault *et al.*, 2010 ; Thibault *et al.*, 2012 ; Thibault et Lenfant, 2014).

La validité prédictive est une grande absente des manuels de tests. Elle est pourtant extrêmement intéressante car elle fait le lien entre le score obtenu à un test et la réalité du vécu quotidien.

Fidélité

- **Fidélité test-retest : 0 outil sur 22**

La grande majorité des manuels n'apporte aucune indication sur ce point. Il faut noter cependant que certains outils, comme les EXALang ou la BHK fournissent des données numériques pour ce critère, ce qui constitue un élément positif (Charles *et al.* 2004 ; Lenfant *et al.*, 2009 ; Thibault *et al.*, 2010 ; Thibault *et al.*, 2012 ; Thibault et Lenfant, 2014). Malheureusement les coefficients n'atteignent pas un seuil suffisant pour valider le critère.

- **Fidélité inter-juges : 0 outil sur 22**

Les constats pour ce résultat sont parfaitement identiques à ceux du point précédent. Les EXALang et la BHK apportent des éléments d'information mais le point ne peut être accordé (Charles *et al.* 2004 ; Lenfant *et al.*, 2009 ; Thibault *et al.*, 2010 ; Thibault *et al.*, 2012 ;

Thibault et Lenfant, 2014). Dans le cadre de la BHK les explications fournies sont intéressantes : la fidélité inter-juges atteint bien le coefficient de .90 lorsque l'étude de cette fidélité porte sur les auteurs du test. Cependant, la corrélation inter-juges entre des professionnels qui n'ont pas l'expérience de ce test est de .68. Les auteurs concluent à une bonne fidélité inter-juges lorsque l'outil est bien maîtrisé (Charles *et al.*, 2004).

Nous pouvons nous questionner sur l'impact de la clarté des consignes de cotation sur ce critère.

- **Analyse des items : 2 outils sur 22**

Les deux outils validant ce critère sont CLéA (Pasquet, Parbeau-Gueno, Bourg, 2014) et le TCS (Maeder, 2006). Un autre outil, la L2MA-2 le valide partiellement (Chevrie-Muller, Maillart, Simon, Fournier, 2010).

Tout comme les critères précédents, cet élément est largement absent des manuels étudiés. Une analyse qualitative et quantitative de la pertinence des items inclus dans les épreuves permet pourtant d'assurer une meilleure validité de celles-ci.

Sensibilité et spécificité

- **Sensibilité et spécificité : 0 outil sur 22**

Les résultats obtenus pour ce critère sont clairs : aucun test parmi ceux étudiés n'apporte assez de preuves d'une sensibilité et d'une spécificité suffisantes. Pourtant ces éléments sont directement liés à la notion de seuil dit « pathologique », très présente dans la profession. Certains outils fournissent dans leur manuel une mention sur la sensibilité mais il faudrait que de plus nombreux détails soient donnés au lecteur pour juger de la réussite ou non à ce critère. En l'absence d'informations quantitatives suffisamment développées le point n'a donc pu être accordé à aucun outil.

Synthèse des résultats

Les critères les plus échoués sont donc ceux de fidélité test-retest, fidélité inter-juges, et de sensibilité et spécificité. Ces critères ont le point commun de n'être validés par aucun

des outils étudiés. L'analyse des items n'est validée que par deux outils. La validité prédictive n'est partiellement validée que par cinq d'entre eux.

En l'absence de données l'utilisateur du test ne peut savoir quel degré de confiance accorder à son outil. Dans ce cas, comment peut-il s'assurer que son outil est fidèle, valide et donc reflète bien le score vrai du patient ?

Données complémentaires

Outils et date

Parmi les manuels, 9 ne fournissent pas d'indication sur l'année de standardisation de l'échantillon. Dans ces cas-là, à défaut d'avoir cette date, l'année de publication de l'outil a été prise en compte. Il s'agit pourtant d'une donnée qu'il est utile de connaître. Connaître la date de standardisation de l'outil présente plus d'intérêt que de connaître l'année de sa parution car c'est à cet échantillon que le patient est comparé. Comment savoir si un outil n'a pas été publié de nombreuses années après sa standardisation ? Cette question pourrait surtout se poser dans le cas d'un outil réédité. D'autre part, il arrive que des batteries soient composées de différents tests ayant déjà été publiés. Il serait intéressant d'avoir des informations claires et précises à ce sujet : les normes de l'échantillon sont-elles celles du test déjà existant ou viennent-elles d'être établies et ré-étalonnées ?

Sur les 22 outils et en référence à l'année en cours (2018) :

- Deux sont datés de moins de 5 ans :
 - CLéA, dont l'étalonnage arrive au seuil des 5 ans (Pasquet *et al.*, 2014)
 - EXALang LyFac, dont l'année d'étalonnage n'est pas précisée. Nous avons tout de même l'indication d'un étalonnage « en juillet de l'année de référence » (référence à la classe du patient) (Thibault et Lenfant, 2014, p. 17)
- Neuf ont entre 5 ans et 10 ans : parmi eux certains sont de nouvelles versions d'un outil qui existait antérieurement. Il convient donc de se référer à l'année d'étalonnage, qui par ailleurs peut varier d'un sous-groupe à l'autre. Par exemple l'EXALang 5-8 a ajouté un niveau mi-CP dont l'étalonnage date de 2010 ; les autres groupes sont datées de 2002 (Thibault *et al.*, 2010).

- Onze ont plus de 10 ans. Cela représente exactement la moitié des outils étudiés ; sans compter les outils du gradient précédent dont la date de standardisation de l'échantillon diffère de celle de publication ou n'est pas fournie.

La majorité des outils étudiés comporte donc un échantillon de normalisation de référence datant d'il y a plus de dix ans. Etant conscients, en tant qu'orthophonistes, de l'évolution de la langue et de la population il convient donc de garder un certain recul dans l'interprétation des résultats obtenus grâce à nos outils (Gaul Bouchard *et al.*, 2009).

Population visée

Concernant les tranches d'âges ciblées par ces outils les résultats n'ont rien d'étonnant : les outils développés pour évaluer le langage écrit visent majoritairement les enfants du CP à la 3^{ème}. Certains, moins nombreux puisqu'ils ne sont que cinq permettent une évaluation après le collège : Alouette-R, ECLA 16+, EXALang LyFac, Le vol du PC et Vitesse en lecture (Boutard, Claire, Gretchanovsky, 2004 ; Lefavrais, 2005 ; Khomsi, Pasquet, Nanty, Parbeau-Gueno, 2005 ; Gola-Asmussen, Lequette, Pouget, Rouyer, Zorman, 2010 ; Thibault et Lenfant, 2014). Chronosdictées a été complété de normes pour cette population qui n'ont pas pu être étudiées ici (Baneath, Boutard, Alberti, 2006).

Il reste intéressant de savoir que des outils ont été développés spécifiquement pour les lycéens et jeunes adultes en réponse à un besoin d'évaluation pour cette population : ECLA 16+ et Exalang LyFac (Gola-Asmussen *et al.*, 2010 ; Thibault et Lenfant, 2014).

Disponibilité des outils

Seuls L'alouette (Lefavrais, 1967) et Evalec (Sprender-Charolles, Colé, Piquard-Kipffer, Leloup, 2010) ne sont plus disponibles à l'achat.

Les résultats des 22 outils soumis aux 13 critères psychométriques définis mettent essentiellement en avant les manques de preuves apportées par les manuels de tests sur ces aspects : validité, fidélité ainsi que sensibilité et spécificité. Ces éléments importants dans le jugement des qualités métrologiques d'un outil apparaissent comme les parents pauvres de l'évaluation du langage écrit en orthophonie. Il s'avère aussi qu'une large partie des outils étudiés possède des normes qui commencent à vieillir.

Face à ces constats, que pouvons-nous dire ? Quelles pistes de réflexion, quelles perspectives ?

Discussion

Le but de ce travail était d'étudier les outils disponibles dans l'évaluation du langage écrit pour la population francophone. Les résultats obtenus permettent de remplir cet objectif d'un état des lieux sur les qualités psychométriques d'outils utilisés dans l'évaluation orthophonique du langage écrit.

Constats

Un des premiers points, et sans doute le plus frappant, concerne l'hétérogénéité avec laquelle les différents critères sont remplis ou non. Si certains sont validés par la totalité des outils ou presque, d'autres sont largement absents des manuels.

Cette absence presque totale de données concernant notamment la validité, la fidélité, la sensibilité et la spécificité fait écho aux résultats obtenus au cours des études précédentes (McCauley et Swisher, 1984b ; Plante et Vance, 1994 ; Gaul Bouchard *et al.*, 2009 ; Friberg, 2010 ; Leclercq et Veys, 2014). Certains auteurs ont émis l'idée que les critères psychométriques pouvaient ne pas être remplis par les outils en raison de la sévérité de la cotation des critères. Mais cette possibilité a cependant été rapidement balayée par le constat commun suivant : dans la très grande majorité des cas, lorsqu'un critère n'est pas rempli, c'est bien par absence de sa mention dans le manuel de l'outil et non en raison d'une cotation sévère (telle qu'un seuil élevé à atteindre). Une analyse plus souple n'aurait ainsi rien changé aux résultats, sauf exceptions (Gaul Bouchard *et al.*, 2009). Au cours de ce travail notre constat a été le même : c'est souvent l'absence d'information sur un critère qui empêche sa validation. Comme le formulaient déjà McCauley et Swisher en 1984 à ce propos « no news is bad news » : « pas de nouvelles, mauvaises nouvelles » (McCauley et Swisher, 1984b, p. 41).

Ainsi, de manière générale les concepteurs de tests ne font pas état de certaines informations. Comment expliquer cela ?

L'hypothèse principale reste celle de l'investissement nécessaire pour pouvoir fournir ces données : plus elles sont contraignantes à obtenir, moins elles sont présentes dans les manuels de tests. Gaul Bouchard *et al.* (2009) mentionnent la nature de ces manques d'investissement : financier, temporel mais aussi lié aux connaissances psychométriques. A

l'opposé, les critères les plus aisément validés sont ceux qui ne demandent aucun investissement supplémentaire, ou un moindre. Le critère relatif aux consignes d'administration et de cotation est à titre d'exemple bien réussi par les outils dans les différentes études menées.

Certaines informations sont absentes alors qu'elles ne nécessitent, a priori, aucun investissement particulier. Nous pensons en particulier au critère « Qualifications de l'évaluateur ». Dans notre étude ce critère n'a été validé que pour 14 des 22 outils, soit une validation dans 64% des cas. Nous nous attendions à un résultat supérieur sur ce point. Par ailleurs, tout comme Leclercq et Veys (2014) nous observons que les qualifications requises ne sont pas données assez explicitement. Avoir en début de manuel une formulation de type « Outil à destination des orthophonistes » serait appréciable et plus clair. Ce critère des qualifications nécessaires à l'utilisation de l'outil d'évaluation a été validé par l'ensemble des outils dans l'étude de Friberg (2010), soit 9 tests, et par environ la moitié des outils dans les études de McCauley et Swisher (1984b) et de Plante et Vance (1994).

Les résultats de cette étude doivent nous rappeler toutes les précautions dont nous devons faire preuve lors de l'évaluation normée d'un patient et de l'interprétation de ses résultats. Les outils étudiés présentent des lacunes sur certains aspects, particulièrement sur ces données importantes que sont la validité, la fidélité, la sensibilité et la spécificité. Si la validité d'un test n'est plus garantie, comment juger du crédit à accorder aux résultats ? Un test non-sensible aurait pour conséquence de ne pas repérer un patient qui devrait l'être, un test non-spécifique aurait pour conséquence de repérer, à tort, trop de patients. Ces éléments dont l'importance a été présentée précédemment devraient figurer dans les manuels des outils d'évaluation.

Ainsi, Friberg (2010) évoque la nécessité d'un examen attentif des propriétés psychométriques des outils d'évaluation et ajoute que, heureusement, un manuel est fourni avec l'outil.

Mais encore faut-il le lire.

La lecture, mais surtout la compréhension de la construction d'un outil demande parfois un certain temps et donc un certain investissement. Tout comme la notion d'une responsabilité partagée entre auteur, éditeur et utilisateur du test avait été développée, la notion d'un investissement de part et d'autre est présentée ici. D'un côté, les auteurs d'outils doivent mettre en œuvre les moyens nécessaires pour procurer au lecteur du manuel des informations claires et détaillées. De l'autre côté, les orthophonistes doivent porter un regard attentif aux données présentes dans ce manuel.

Par ce travail sur les outils d'évaluation c'est aussi un double message que nous souhaitons faire entendre : aux orthophonistes et aux auteurs et éditeurs.

Aux orthophonistes

L'importance du manuel d'utilisation

La lecture du manuel d'utilisation d'un outil est une étape importante et indispensable dans la pratique clinique. C'est par cette lecture que le professionnel pourra garantir les trois points suivants :

- Avoir une connaissance des limites de l'outil
- Veiller à la fidélité et à la validité des résultats: il est impossible de garantir que les conditions de passation et cotation seront les mêmes que celles utilisées lors de l'étalonnage en l'absence d'une prise de connaissances complète de ces éléments
- S'assurer que l'outil est en adéquation avec l'objectif de l'orthophoniste

Connaître les limites d'un outil permet de pondérer l'importance qui pourrait être donnée à un score. Il s'agit d'une prise de recul vis-à-vis des résultats obtenus. Quel degré de confiance leur accorder? Une piste intéressante est la présence d'intervalles de confiance dans les outils. L'intervalle de confiance permet d'estimer la marge d'erreur possible autour du score observé. Si le test présente de bonnes qualités métrologiques, que les procédures de passation et cotation ont été respectées et qu'un intervalle de confiance est présent alors il est plus aisé pour le professionnel de porter un jugement sur le score obtenu par le patient.

Concernant ce score, les manuels fournissent à leurs utilisateurs diverses mesures de tendance centrale : moyenne, écart-type, rangs centiles, z-score. Au-delà du constat de la présence de ces mesures, il est important pour le professionnel d'aller lire le manuel et de voir ce qui se cache derrière ces données chiffrées. L'utilisation de l'écart-type est-elle bien adaptée ? Quelles sont les caractéristiques relatives aux centiles : y a-t-il un effet de plafond et une épreuve vite saturée ? L'interprétation clinique des résultats ne sera pas la même selon la situation. Etudier ces données permet de limiter les erreurs d'interprétation ou de surinterprétation du score d'un patient. Un score situé à par exemple -1 écart-type de la moyenne de l'échantillon de référence peut très bien refléter une difficulté fonctionnelle réelle dans le quotidien tandis qu'un score situé à -2 écart-types dans une autre épreuve peut n'avoir aucun lien avec la vie du patient. Les scores seuils restent des limites arbitrairement fixées. Le seuil fixé devrait avoir sa pertinence justifiée par des données. La manière dont l'épreuve a été construite, son échantillon d'étalonnage mais aussi, et surtout, les répercussions fonctionnelles dans le quotidien du patient sont des éléments qu'il est indispensable de prendre en compte (Grégoire, 2014).

Si le professionnel peut limiter la part de biais provenant de son utilisation de l'outil d'évaluation, il n'est pas responsable des manques ou défauts présents dans les outils.

Une fois que le clinicien est conscient des limites des outils d'évaluation standardisés et normalisés il pourrait être tenté de douter de leur utilité et préférer ne plus s'en servir (McCauley et Swisher, 1984b). Mais rejeter ces outils serait oublier leur intérêt et la raison de leur existence : celle de limiter les biais du seul jugement subjectif du clinicien (Grégoire, 2014). Ainsi pour l'orthophoniste toute la subtilité de l'évaluation d'un patient va être de savoir doser entre son observation qualitative et les mesures chiffrées obtenues ; l'objectif étant de limiter les biais de part et d'autre.

Il apparaît intéressant de rappeler qu'un test normalisé et standardisé n'est pas adapté à tout objectif d'évaluation et que d'autres méthodes d'évaluation, comme l'évaluation dynamique, peuvent être utilisées. L'évaluation dynamique est liée à la notion de Zone Proximale de Développement développée par Vygotsky : il s'agit d'évaluer le potentiel d'apprentissage du patient et de repérer les étayages nécessaires et efficaces dans l'acquisition

d'une compétence (Cattini et Clair-Bonaimé, 2017). Elle est donc, par exemple, bien plus adaptée à la mesure de l'efficacité d'une thérapie que ne peut l'être un test normalisé.

Des revendications de la profession ?

L'influence de l'orthophoniste, son rôle en tant qu'acheteur d'outils constitue un potentiel levier d'action dans l'amélioration des outils d'évaluation. McCauley et Swisher (1984b) ont conclu leur étude par l'idée que les qualités psychométriques doivent faire partie intégrante du travail de création d'un test, et non constituer un élément facultatif. Si dans son étude en 2010 Friberg notait une amélioration quant aux informations présentes dans les manuels des tests, un constat reste identique : les auteurs et éditeurs d'outils doivent encore faire l'effort d'inclure plus de données. Nous pouvons imaginer que si de nombreux orthophonistes faisaient part de leur envie (ou de la nécessité) aux auteurs et éditeurs de voir figurer plus d'informations dans les manuels des tests alors les choses évolueraient.

Le critère « qualifications de l'évaluateur » est un point sur lequel les orthophonistes peuvent, et ont tout intérêt à se pencher. Cet élément va de pair avec le moyen d'obtention de l'outil. Un test ou une batterie accessible au grand public ou à des nombreux professionnels peut constituer une menace à la bonne évaluation d'un patient. Le fait qu'un patient ait déjà été confronté, peut-être même peu de temps avant l'évaluation réalisée par l'orthophoniste, au même test pourrait fausser les résultats, par effet de mémorisation de l'épreuve potentiellement. L'orthophoniste doit alors avoir à sa disposition plusieurs outils permettant de répondre au même objectif d'évaluation afin d'en changer si le cas se présente.

Plusieurs outils mentionnent d'autres professionnels que l'orthophoniste en tant qu'évaluateur qualifié. Ainsi Vitesse en lecture indique « enseignants, psychologues et rééducateurs scolaires, conseillers d'orientation psychologues, orthophonistes ou bien encore médecin » (Khomsy *et al.*, 2005, p.V), Le petit poucet « orthophoniste, phoniatre.. » (De Plazaola *et al.*, 2003, p.7). ECLA 16+, BALE et ODEDYS indiquent quant à eux « orthophoniste, psychologue, médecin » (Jacquier-Roux, Valdois, Zorman, Lequette, Pouget, 2005 ; Jacquier-Roux *et al.* ; Gola-Asmussen *et al.*, 2010). En tant qu'orthophonistes il reste intéressant de savoir avec quels professionnels certains outils sont partagés.

S'intéresser aux outils d'évaluation permet au professionnel de prendre des décisions éclairées dans sa pratique clinique, que ce soit dans le choix de la méthode d'évaluation, celui du test ou dans le savant mélange des données quantitatives et qualitatives.

Aux concepteurs et éditeurs d'outils d'évaluation

Les tests et batteries présentent l'intérêt d'apporter des données objectives au clinicien. Il est important pour lui de savoir à quel point il peut se fier à ces données car, au-delà d'un score, c'est bien par le patient que ce clinicien est concerné. La présence de toutes les informations requises à une bonne utilisation de l'outil est nécessaire, c'est à nouveau une question d'éthique et de responsabilité. Malgré toutes les contraintes temporelles ou financières inhérentes à la création d'un outil d'évaluation, il reste primordial que les concepteurs fassent leur possible pour répondre aux qualités psychométriques de validité, fidélité, sensibilité et spécificité.

Nous invitons tout créateur potentiel d'outil à consulter les recommandations concernant les outils d'évaluations établies par Leclercq et Veys (2014) (Annexe 4).

Il existe un autre élément que nous trouverions intéressant de voir figurer dans les manuels d'utilisation : les limites de l'outil. En effet, si le concepteur est conscient des limites de son épreuve et les partage avec l'utilisateur cela faciliterait grandement le travail de ce dernier dans l'estimation du degré de confiance à accorder aux résultats.

Comme l'ont fait remarquer Cattini et Clair-Bonaimé (2017), un souci rencontré par les professionnels est celui de l'absence de données disponibles sur les qualités psychométriques avant l'achat d'un outil. Il est souvent nécessaire pour le professionnel de devoir investir avant de savoir si les caractéristiques psychométriques lui conviennent. Nous appelons les concepteurs et éditeurs à rendre ces informations disponibles avant l'achat.

Le point suivant concerne plutôt les éditeurs et vendeurs d'outils : si les auteurs précisent les qualifications nécessaires à la bonne utilisation d'un test, il faut alors garantir l'accès à celui-ci exclusivement aux professionnels concernés.

Limites du travail effectué

Une des premières limites à ce travail a été l'impossibilité d'analyser tous les outils d'évaluation du langage écrit recensés. Il aurait en effet été préférable d'avoir une vue d'ensemble de l'éventail des outils existants et de leurs qualités et limites.

Une seconde limite que nous pouvons constater est liée à la procédure de cotation des outils. En effet, l'analyse n'a pu être effectuée que par une personne. Une procédure avec deux examinateurs ou plus aurait permis une confrontation dans la manière d'évaluer les outils et des ajustements. Dans l'étude de Friberg (2010) six personnes différentes ont étudié les 9 outils sur la base des 10 critères psychométriques retenus. Le taux d'accord entre ces six personnes dans la cotation variait de 79% (critère : analyse des items) à 100% (critères : objectif du test, taille de l'échantillon, description de l'échantillon). McCauley et Swisher (1984b) ont soumis les tests à la cotation de deux examinateurs, l'accord entre eux allait de 67% (critère : analyse des items) à 100% (critère : fidélité inter-juges). Cette procédure a l'avantage d'offrir une discussion puis une convergence dans la manière de coter et permet de limiter les erreurs. Il serait pertinent de pouvoir appliquer cette procédure à toute analyse de ce type.

Perspectives

Ce travail tout comme les autres déjà effectués met en avant des manques quant aux qualités psychométriques des outils utilisés en orthophonie. Quelles pistes s'offrent à la profession pour changer les choses ?

Tout d'abord une solution proposée par Leclercq et Veys (2014) consiste à élaborer des standards de qualité. Elles invitent pour cela les orthophonistes et les concepteurs d'outils à réfléchir à la qualité des outils. En effet, nous pensons aussi que ce n'est que par un élan de la profession que les choses pourront évoluer : par une prise de conscience à la fois des

cliniciens et des créateurs d'outils d'évaluation. Les concepteurs sont d'ailleurs bien souvent des orthophonistes cliniciens eux-mêmes.

D'autre part, l'inscription des études d'orthophonie dans le parcours Licence-Master-Doctorat pourrait se révéler être un levier dans la création de nouveaux outils avec la possibilité pour les orthophonistes d'intégrer le domaine de la recherche. C'est d'ailleurs par une unité d'enseignement appartenant au point « 7. Recherche en orthophonie » du référentiel de formation que l'idée de ce mémoire a commencé à germer : il s'agit de l'unité d'enseignement « 7.4 - Méthodologie d'analyse d'article ». En effet, s'il est possible de lire des articles en y portant un regard critique, ne pourrions-nous pas appliquer cette même démarche lors de la lecture des manuels des outils d'évaluation ? Petit à petit, il est envisageable que de plus en plus d'étudiants s'intéressent tôt aux tests et batteries utilisés par la profession.

Enfin, sans aller immédiatement à la création de nouveaux outils, une étape plus facilement envisageable consiste à ré-étalonner ceux existants. Des travaux en ce sens permettraient d'obtenir des normes plus récentes.

Concernant la poursuite de travaux dans le domaine de l'étude des outils de tests, celle-ci est et sera toujours possible. En effet, un travail similaire à celui-ci pourrait être conduit dans d'autres champs de l'évaluation orthophonique. Un catalogue complet de tous les outils et de leurs caractéristiques psychométriques serait un outil extrêmement intéressant pour les professionnels.

La reproduction à intervalles réguliers d'études sur les outils d'évaluation permettrait de constater s'il y a une évolution dans la conception des tests au fil du temps. Les critères seront-ils mieux remplis dans dix ans ?

Le présent travail pourrait également être complété par une étude se rapprochant de celle effectuée par Lafay *et al.* (2014) : les outils seraient étudiés par sous-domaines d'évaluation. Cela produirait un répertoire des tests disponibles et étalonnés pour une tranche d'âge donnée, évaluant une compétence bien précise.

Les possibilités de travaux et d'avancée sur les outils d'évaluation ne manquent pas.

Conclusion

Le travail effectué avait pour objectif de proposer, comme un état de lieux, une analyse psychométrique de différents outils utilisés dans l'évaluation du langage écrit. L'utilisation d'outils standardisés et étalonnés est très présente chez les orthophonistes mais que pouvons-nous dire de ces outils ?

Après nous être intéressés à la pratique professionnelle de l'évaluation en orthophonie ainsi qu'au domaine de la psychométrie et son lien avec la profession, nous avons développé le choix d'une méthode d'analyse. Cette dernière a pris appui sur plusieurs études déjà réalisées sur les outils d'évaluation.

Au total ce sont 22 outils qui ont été comparés et soumis à une grille d'analyse. Ces outils ont été étudiés selon 13 critères psychométriques que nous avons rassemblés de cette manière : critères généraux, critères relatifs à l'échantillon, validité, fidélité, sensibilité et spécificité. Un score a été attribué à chaque outil qui pouvait obtenir un maximum de 13 points, correspondant à une validation des 13 critères. A cela nous avons ajouté des informations concernant l'année de publication ou d'étalonnage et les tranches d'âge concernées par les outils étudiés. Les résultats ont été présentés dans un tableau permettant d'effectuer rapidement une comparaison entre les différents outils sur la base des critères retenus.

Les résultats ont mis en avant de grands manques concernant les critères de validité, fidélité et de sensibilité et spécificité. L'absence de ces qualités psychométriques doit conduire les professionnels à la prudence lors de l'utilisation d'un outil standardisé et normalisé. En effet, 45% des outils ont obtenu un score inférieur à 6,5/13 et aucun outil ne dépasse le score de 8/13. Les conclusions de cette étude sont semblables à celles des travaux précédemment réalisés :

- Les outils manquent de preuves sur leur validité et fidélité
- La sensibilité et la spécificité sont grandement absentes des manuels

Il est donc important que les orthophonistes et les concepteurs d'outils d'évaluation aillent ensemble vers une amélioration des outils d'évaluation.

Bibliographie

ANAES. (2002). L'orthophonie dans les troubles spécifiques du développement du langage oral chez l'enfant de 3 à 6 ans. *Rééducation orthophonique*, 209.

Auzias, S., Le Garsmeur, A. (2012). L'évolution du rôle social du test dans la pratique orthophonique (Université Claude Bernard Lyon1 – ISTR, Lyon).

Baneath, B., Boutard, C., Alberti, C. (2006). Chronosdictées, outils d'évaluation des performances orthographiques avec et sans contrainte temporelle. Isbergues : Ortho Edition.

Belot, C., Tricot, M. (2001). *Les tests en orthophonie Tome 1 : Langage oral- Langage écrit Enfants - Adolescents*. Isbergues : Ortho Edition

Bétrancourt, P., Lederlé, E. (2017). Analyse qualitative d'une expérimentation d'Evaluation des Pratiques Professionnelles. *Rééducation orthophonique*, 272, 5-107.

Betz, S. K., Eickhoff, J. R., Sullivan, S. F. (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.

Boudreault, M-C., Cabirol, E-A., Trudeau, N., Poulin-Dubois, D., Sutton, A. (2007). Les inventaires MacArthur du développement et de la communication : validité et données normatives préliminaires. *Revue canadienne d'orthophonie et d'audiologie*, 31(1), 27-37.

Boutard C., Claire I., Gretchanovsky, L. (2004). *Le vol du PC : évaluation fonctionnelle de la lecture chez les sujets de 11 à 18 ans*. Isbergues : Ortho Editions.

Brin-Henry, F. (2011). La terminologie crée-t-elle la pathologie ? Le cas de la pratique clinique de la pose du diagnostic orthophonique. (Thèse, Université Nancy 2)

Brin-Henry, F. (2015). Réflexion sur la terminologie des troubles du langage écrit en orthophonie. *Rééducation orthophonique*, 262, 175-192.

Brin-Henry, F. (2016). Epreuves, tests et batteries d'évaluation utilisés par les orthophonistes en France. *Guide de l'orthophoniste : Intervention dans les troubles du langage oral et de la fluence*, II, 63-77.

Calvarin, M. (2013). *Les tests en orthophonie : évaluation des troubles d'origine neurologique de l'adulte*. Isbergues : Ortho Edition

Casalis, S., Leloup, G., Bois Parriaud, F. (2013). *Prise en charge des troubles du langage écrit chez l'enfant*. Issy-Les-Moulineaux : Elsevier Masson.

Cattini, J., Clair-Bonaimé, M. (2017). Les apports de l'Evidence-Based Practice et de la Practice-Based Evidence : du bilan initial à l'auto-évaluation du clinicien. *Rééducation orthophonique*, 272, 109-138.

Charles, M., Soppelsa, R., Albaret, J.-M. (2004). *BHK, échelle d'évaluation rapide de l'écriture chez l'enfant*. Paris : ECPA.

Chevrie-Muller, C., Maillart, C., Simon, A.-M., Fournier, S. (2010). *L2MA-2, batterie langage oral, langage écrit, mémoire, attention - 2^{ème} édition*. Paris : ECPA.

De Plazaola, C., Gauthier, F., Arsicaud, M.F., Pech-Georgel, C. (2003). *Test d'orthographe Le petit Poucet*. Marseille : Solal.

Duchene, A. (2010). *EMILIE, protocole de compréhension d'un texte écrit chez les collégiens*. Grenade : Orthomotus.

Durieux, N., Pasleau, F., Vandenput, S., Maillart, C. (2013). Les orthophonistes utilisent-ils des données issues de la recherche scientifique? Analyse des pratiques afin d'ajuster la formation universitaire en Belgique. *Glossa*, 113, 110-118.

Ferrand, P. (2007). Du rôle fondamental de la démarche clinique dans l'évaluation des troubles du langage. *Rééducation orthophonique*, 231, 25-33.

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.

Garcia, L. J., Paradis, J., Sénécal, I., Laroche, C. (2006). Utilisation et satisfaction à l'égard des outils en français évaluant les troubles de communication. *Revue d'orthophonie et d'audiologie*, 30(4), 239-249.

Gaul Bouchard, M-E., Fitzpatrick, E. M., Olds, J. (2009) Analyse psychométrique d'outils d'évaluation utilisés auprès des enfants francophones *Revue canadienne d'orthophonie et d'audiologie*, 33(3), 129-139.

Gola-Asmussen, C., Lequette, C., Pouget, G., Rouyer, C., Zorman, M. (2010). *ECLA-16+, Outil d'évaluation de compétences de lecture chez l'adulte de plus de 16 ans*. Grenoble : Cogni-sciences.

Greenslade, K., J., Plante, E., Vance, R. (2009). The Diagnostic Accuracy and Construct Validity of the Structured Photographic Expressive Language Test – Preschool : Second Edition. *Language, Speech, and Hearing Services in Schools*, 40, 150-160.

Grégoire, J. (2014, mai). *L'examen diagnostic est-il normatif ?* Conférence lors de la journée "Actualités de l'examen psychologique de l'enfant : l'apport de la neuropsychologie" à l'espace Reuilly, Paris .Repéré à : <https://www.youtube.com/watch?v=6Rw7SilGrqg&t=1767s>

Grégoire, J., Psychométrie. Dans Encyclopædia Universalis. Repéré à : <http://www.universalis.fr/encyclopedie/psychometrie/>

Hernandez, E., Ribaga, S. (2010). La construction et l'évolution du cadre théorique de l'orthophonie en France : Lecture par une sélection de batteries et tests (1967-2005) (Université Claude Bernard Lyon1 – ISTR, Lyon).

INSERM. (2007). IV Prévention et prise en charge. Dans INSERM, *Dyslexie, dysorthographe, dyscalculie : Bilan des données scientifiques*. Paris : Les éditions Inserm.

Jakubowitz, M., Schill, M. J. (2008). Ethical Implications of Using Outdated Standardized Tests. *Perspectives on School-Based Issues*, 9, 79-83.

Jacquier-Roux, M., Lequette, C., Pouget, G., Valdois, S., Zoman, M. (2010). *BALE, batterie analytique du langage écrit*. Grenoble : Cogni-sciences.

Jacquier-Roux, M., Valdois, S., Zorman, M., Lequette, C., Pouget, G. (2005). *ODEDYS, outil de dépistage des dyslexies version 2*. Grenoble : Cogni-sciences.

Khomsy, A. (1999). *LMC-R, épreuve d'évaluation de la compétence en lecture*. Paris : ECPA.

Khomsy, A., Pasquet, F., Nanty, I., Parbeau-Gueno, A. (2005). *Vitesse en lecture, dépistage du 'risque' en lecture et première caractérisation du niveau de performance*. Paris : ECPA.

Lafay, A., Saint-Pierre, M-C., Macoir, J. (2014). L'évaluation des habiletés mathématiques de l'enfant : inventaire critique des outils disponibles, *Glossa*, n° 116, 33-58.

Launay, L. (2015). Vers un diagnostic argumenté de la dyslexie et/ou dysorthographe développementale. *Rééducation orthophonique*, 262, 83-119.

Laveault, D., Grégoire, J. (2014). *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (3e éd.). Bruxelles: De Boeck.

Leclercq, A-L., Veys, E., (2014). Réflexions sur le choix de tests standardisés lors du diagnostic de dysphasie *A.N.A.E* n° 131, 26(IV), 374-382.

Lederlé, E., Maeder, C. (2016). Le bilan orthophonique. *Guide de l'orthophoniste : Intervention dans les troubles du langage oral et de la fluence*, II, 19-29.

Lefavrais, P. (1967). *L'Alouette*. Paris : ECPA.

Lefavrais, P. (2005). *Alouette-R*. Paris : ECPA.

Lenfant, M., Thibault, M. -P., Helloin, M. -C. (2009). *EXALang 11-15-, batterie informatisée d'examen du langage oral, langage écrit, compétences transversales*. Mont-Saint-Aignan : Motus.

Lequette, C., Pouget, G., Zorman, M. (2009). *E.L.FE, Evaluation de la Lecture en Fluence*. Grenoble : Cogni-sciences.

Maeder, C. (2006). *TCS Test de Compréhension Syntaxique*. Isbergues : Ortho Edition.

Marin-Curtoud, S., Rousseau, T., Gatignol, P. (2010). Etat des lieux sur « le test ». *L'orthophoniste* n°296, 19-26.

McCauley, R.J., Swisher, L. (1984a). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *The Journal of speech and hearing disorders*. 49. 338-348.

McCauley, R. J., Swisher, L. (1984b). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders*, 49(1), 34-42.

Merrell, A. W., Plante, E. (1997). Norm-referenced Test Interpretation in the Diagnostic Process. *Language, Speech, and Hearing Services in Schools*, 28, 50-58.

Ministère de l'enseignement supérieur et de la recherche. (2013) : Certificat de capacité d'orthophoniste : Annexe 1 Référentiel d'activités (Bulletin officiel n° 32 du 5 septembre 2013).

Monetta, L., Desmarais, C., MacLeod, A. A., St-Pierre, M-C., Bourgeois-Marcotte, J., Perron, M., (2016). Recension des outils franco-qubécois pour l'évaluation des troubles du langage et de la parole . *Canadian Journal of Speech-Language Pathology and Audiology*, 40(2), 165-175.

Monetta, L. et tous les membres de l'axe 4 du REPAR (2016). Fiches descriptives des outils validés et/ou normés en franco- québécois pour l'évaluation du langage et de la parole, de 1980 à 2014. Repéré à <http://www.repar.veille.qc.ca/fr/outils/Orthophonie.html>

Observatoire National de la Démographie des Professions de Santé. (2012). Tome 3 : Des professions de santé en évolution pharmaciens, orthophonistes, médecins en formation. Repéré à :

<http://solidarites-sante.gouv.fr/ministere/acteurs/instances-rattachees/article/rapports-ondps>

Pasquet, F., Parbeau-Gueno, A., Bourg, E. (2014). *CLéA, batterie de langage nouvelle génération*. Paris : ECPA.

Peña, E. D., 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, 247–254, doi : 1058-0360/06/1503-0247

Perdrix, R. (2015). L’entretien d’anamnèse comme dispositif organisateur premier du raisonnement clinique dans le diagnostic de dyslexie développementale. *Rééducation orthophonique*, 262, 53-81.

Perdrix, R. (2016). Dyslexie développementale : méthodologie diagnostique, clinique orthophonique. *Guide de l’orthophoniste : Intervention dans les troubles du langage écrit et raisonnement*, III, 73-107.

Plante, E., Vance, R. (1994). Selection of preschool language tests : a data-based approach. *Language, Speech, and Hearing Services in Schools*, 45, 15-24

Shahmahmood, T. M., Jalaie, S., Soleymani, Z., Haresabadi, F., & Nemati, P. (2016). A systematic review on diagnostic procedures for specific language impairment: The sensitivity and specificity issues. *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, 21, 67. doi : 10.4103/1735-1995.189648

Skarakis-Doyle, E., Miller, L. T., & Reichheld, M. (2000). Construct validity as a foundation of evidence-based practice: The case of the Preschool Language Assessment Instrument. *Journal of Speech-Language Pathology and Audiology*, 24(4), 180-191.

Spaulding, T. J., Plante, E., Farinella, K.A. (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. (2011). Comparison of severity ratings on norm-referenced tests for children with specific language impairment. *Journal of Communication Disorders*, 45(1), 59-68.

Spaulding, T. J., Swartout Szulga, M., Figueroa, C. (2012). Using norm-referenced tests to determine severity of language impairment in children: disconnect between U.S. policy makers and test developers. *Language, Speech, and Hearing Services in Schools*, 43(2), 176-90. doi: 10.1044/01611461(2011/10-0103)

Sprenger-Charolles, L., Colé, P., Piquard-Kipffer, A., Leloup, G. (2010). *EVALEC, Batterie informatisée d'évaluation diagnostique des troubles spécifiques d'apprentissage de la lecture*. Isbergues : Ortho Edition.

Thibault, M. -P., Helloin, M.-C., Croteau, B. (2010). *EXALang 5-8, batterie informatisée d'examen du langage oral et écrit chez des enfants de 5 à 8 ans*. Mont-Saint-Aignan : Motus.

Thibault, M.-P., Lenfant, M. (2014). *EXALang LyFac, batterie informatisée d'examen du langage oral, langage écrit, mémoire pour les jeunes adultes*. Grenade : Orthomotus

Thibault, M.-P., Lenfant, M., Helloin, M.-C. (2012). *EXALang 8-11, batterie informatisée d'examen du langage oral, langage écrit, mémoire, attention, compétences transversales*. Mont-Saint-Aignan : Motus.

Thordardottir, E., Kehayia, E., Mazer, B., Lessard, N., Majnemer, A., Sutton, A., Trudeau, N., 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, 580-597. doi:10.1044/1092-4388(2010/09-0196)

Thordardottir, E., Keheyia, E., Lessard, N., Sutton, A., Trudeau, N. (2010). Typical Performance on Tests of Language Knowledge and Language Processing of French-Speaking 5-Year-Olds. *Revue canadienne d'orthophonie et d'audiologie*, 34(1), 5-15.

UNADREO. (2011). Tests et modules. Repéré à
<http://www.unadreo.org/articles/getArticle/59/353>

Witko, A. (2013). Pour une recherche en orthophonie-logopédie au service de la santé, du handicap et du langage. *Glossa*, 113, 27-44.

Witko, A. (2014). Recherche en orthophonie-logopédie et identité professionnelle. *Rééducation orthophonique*, 257, 3-14.

Annexes

Annexe 1 - Message d'Anne Dehêtre et de Michel Barbier communiqué sur la page Facebook de la Fédération Nationale des Orthophonistes le 9 février 2018

Vous trouverez ci-dessous un message de la part d'Anne Dehêtre et de Michel Barbier, concernant la diffusion du matériel Ortho Edition.

Ortho Edition a été créée par des orthophonistes pour des orthophonistes et depuis cette création, est gérée par des orthophonistes en exercice. A l'initiative de la FNO, cette création partait du constat que les orthophonistes n'avaient que peu de matériel adapté à leurs besoins et que la profession devait se doter d'une maison d'édition qui promeuve les outils nécessaires à l'exercice professionnel, qui incite et facilite les orthophonistes à créer et à publier du matériel, des ouvrages et des outils de bilans.

Les objectifs : l'autonomie d'une profession dans ses choix de matériels, d'outils et de références théoriques, du matériel correspondant à une réalité quotidienne d'exercice et enfin, des outils et des ouvrages de qualité à des prix abordables.

D'autres professionnels que les orthophonistes ont placé leur confiance dans cette société et ont choisi d'éditer des ouvrages et des tests co-créés avec des orthophonistes chez Ortho-Edition.

Récemment, Ortho Edition a investi dans la recherche, en lien étroit avec la clinique pour créer, valider, étalonner et enfin éditer les bilans de langage oral - Evalo BB et Evalo - et poursuit avec EVALEO qui sort cette année. Plus récemment encore, l'UNADREO s'est vue confier une mission d'évaluation de la validité des tests proposés à la vente. Nous rappelons ici que, hormis pour les outils de dépistage qui peuvent être destinés à d'autres professionnels, seuls des orthophonistes et logopèdes diplômés ont toutes les compétences nécessaires pour pouvoir exploiter et utiliser le matériel proposé à la vente. Seules les connaissances et les compétences spécifiques des orthophonistes permettent une utilisation de ce même matériel dans un but diagnostique et/ou thérapeutique.

Il nous est impossible de penser que d'autres personnes non orthophonistes pourraient utiliser ce même matériel pour des résultats similaires. Un diagnostic orthophonique ne dépend pas de la simple lecture d'un livret de passation, une rééducation ne dépend pas de la lecture d'une règle d'utilisation. L'intervention orthophonique relève d'une démarche de clinicien, expert des troubles relevant de sa compétence. Le diagnostic orthophonique est défini de par la loi et ne relève que des orthophonistes diplômés.

La vente de matériel Ortho Edition dans une enseigne « tout public » est jusqu'à présent restée anecdotique, et ne représente qu'une dizaine d'exemplaires vendus en 2017.

Cependant, la FNO et Ortho Edition entendent et comprennent les craintes exprimées par les orthophonistes quant à l'exploitation de l'image qui pourrait être véhiculée par cette possibilité d'achat sur des sites tout public et sur la non possibilité de vérifier la qualité des acheteurs.

La FNO et Ortho Edition vont mettre en place dès à présent des procédures plus strictes de contrôle des potentiels acquéreurs de matériels et de tests pour encore intensifier la protection des outils et matériels spécifiques à notre profession.

Anne Dehêtre
Présidente de la FNO

Michel Barbier
Gérant d'Ortho-Edition

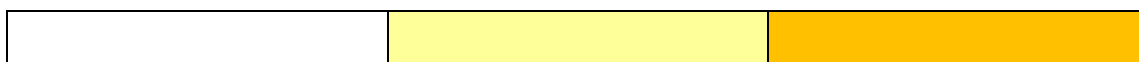
Annexe 2 - Liste des outils évaluant le langage écrit répertoriés

- | | |
|-------------------------------------|------------------------|
| 1. Alouette-r | 22. EXALang LyFac |
| 2. Analec | 23. EXALng 11-15 |
| 3. BALE | 24. EXALng 8-11 |
| 4. Batterie ALOE | 25. INZEC |
| 5. BELEC | 26. L'Alouette |
| 6. BELO | 27. L2MA-2 |
| 7. BHK | 28. La forme noire |
| 8. BLI Bilan de lecture informatisé | 29. Le petit poucet |
| 9. BP/BL | 30. Le vol du PC |
| 10. Chronosdictées | 31. LMC-R |
| 11. CLéA | 32. Logator |
| 12. D-OR-LEC | 33. Maxence |
| 13. ECLA 16+ | 34. Morphorem |
| 14. ECL-Collège | 35. Odedys v2 |
| 15. ECS II | 36. Phonolec |
| 16. ECS III | 37. Read play |
| 17. E.L.FE | 38. TCS |
| 18. EMILIE | 39. TCT |
| 19. EVALAD | 40. Timé-3 |
| 20. Evalec | 41. Tinfolec |
| 21. EXALang 5-8 | 42. Vitesse en lecture |

Annexe 3 - Résultats de l'analyse des outils d'évaluation en fonction des critères psychométriques établis

Légende des tableaux :

- (*) : Outil indisponible à l'achat
- (**): Outils également étudiés par Leclercq et Veys (2014) :
 - EXALang 5-8
 - EXALang 8-11
 - L2MA-2
- (?) : date inconnue
- Gradient de date :



- Blanc : date > 10 ans
- Jaune : 10 ans $\geq$ date > 5 ans
- Orange : date $\leq$ 5 ans

- Validation du critère :



- Blanc (-) : critère non validé
- Gris (+ -) : critère partiellement validé
- Noir (+) : critère validé

	Alouette-R	BALE	BELO	BHK	Chronosdictées	CLéA	ECLA 16+	E.L.FE	Emilie	Evalec (*)	EXALang 5-8 (**)
Tranche d'âge	cp-terminale	ce1-cm2	cp-ce1	cp-cm2	ce1 -3ème	ps - 3 ^{ème}	16-25 ans	ce1-5ème	6 ème-3ème	cp-cm1	gs-ce1
Année : publication (échantillon)	2005 (?)	2010 (1999-2000)	2006 (?)	2004 (?)	2006 (?)	2014 (2012-2013)	2010 (2007-2008)	2009 (?)	2010 (2007)	2010 (?)	2010 (2002 +2010)
Qualifications de l'évaluateur	-	+	+	-	+	+	+	-	-	+	+
Consignes d'administration et de cotation	+	+	+	+	+	+	+	+	+	+	+
Objectif du test précisé	+	+	+	+	+	+	+	+	+	+	+
Sensibilité et spécificité	-	-	-	-	-	-	-	-	-	-	-
Taille de l'échantillon d'étalonnage	-	+	-	+	+	-	+	-	-	-	-
Informations sur l'échantillon d'étalonnage	-	+	+	+	+	+	+	-	+	+	+
Mesures de tendance centrale	+	+	+	+	+	+	+	+	+	+	+
Validité											
Validité théorique	+	+	+	+	+	+	+	-	+	+	+
Validité concourante	-	-	-	+	-	-	-	-	-	-	+ -
Validité prédictive	-	-	-	+ -	-	-	-	-	-	-	+ -
Fidélité											
Fidélité test-retest	-	-	-	-	-	-	-	-	-	-	-
Fidélité inter-juges	-	-	-	-	-	-	-	-	-	-	-
Analyse des items	-	-	-	-	-	+	-	-	-	-	-
Score	4/13	7/13	6/13	7,5/13	7/13	7/13	7/13	3/13	5/13	6/13	7/13

	EXALang 8-11 (**)	Exalang 11-15	EXALang LyFac	L2MA- 2 (**)	L'Alouette (*)	Le petit poucet	Le vol du PC	LMC-R	Odedys V2	TCS	Vitesse en lecture
Tranche d'âge	ce2-cm2	6 ^{ème} - 3ème	jeunes adultes, 16ans+	ce1- 6ème	gs-3ème	ce1- 3ème	11- 18ans	ce1-cm2 5ème	ce1-cm2 5ème	cm1- 3ème	ce1- terminale
Année : publication (échantillon)	2012 (2012)	2009 (2009)	2014 (?)	2010 (2009)	1967 (?)	2003 (?)	2004 (1999)	1999(1998)	2005(1999- 2000+2005)	2006 (2004- 2005)	2005 (2004)
Qualifications de l'évaluateur	+	+	+	-	-	-	+	-	+	+	+
Consignes d'administration et de cotation	+	+	+	+	+	+	+	+	+	+	+
Objectif du test précisé	+	+	+	+	-	+	+	+	+	+	+
Sensibilité et spécificité	-	-	-	-	-	-	-	-	-	-	-
Taille de l'échantillon d'étalonnage	+	-	+	-	-	-	-	-	+	-	+
Informations sur l'échantillon d'étalonnage	+	+	+	+	-	-	+	-	+	+	+
Mesures de tendance centrale	+	+	+	+	-	+	+	+	+	+	+
Validité											
Validité théorique	+	+	+	+	-	+	+	+	+	+	+
Validité concourante	+ -	+ -	+ -	-	-	-	-	-	-	+	-
Validité prédictive	+ -	+ -	+ -	-	-	-	-	-	-	-	-
Fidélité											
Fidélité test-retest	-	-	-	-	-	-	-	-	-	-	-
Fidélité inter-juges	-	-	-	-	-	-	-	-	-	-	-
Analyse des items	-	-	-	+ -	-	-	-	-	-	+	-
Score	8/13	7/13	8/13	5,5/13	1/13	4/13	6/13	4/13	7/13	8/13	7/13

Annexe 4 - Synthèse des recommandations concernant les outils d'évaluation par Leclercq et Veys (2014)

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 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 concurrente	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.

Titre du Mémoire : Analyse psychométrique d'outils utilisés dans l'évaluation orthophonique du langage écrit

RESUME

Ce travail a pour but d'étudier les outils d'évaluation normalisés utilisés par les orthophonistes dans l'évaluation du langage écrit. Les outils ont été recensés et 22 ont été passés en revue sur la base de leurs propriétés psychométriques : 13 critères psychométriques ont été utilisés. Les résultats montrent que la majorité des outils n'apporte pas de preuves suffisantes de leur validité, fidélité, sensibilité et spécificité. Ces données sont fréquemment absentes des manuels d'utilisations. Les orthophonistes devraient donc se montrer prudents dans l'utilisation des outils normalisés et l'interprétation des résultats obtenus grâce à eux. Les concepteurs d'outils devraient fournir des données psychométriques plus complètes.

MOTS-CLES

évaluation, fidélité, langage écrit, psychométrie, test, validité

ABSTRACT

The aim of this work was to study standardized assessment tools used by speech-language pathologists in the evaluation of written language. Tools were listed and 22 of this tools were reviewed based on their psychometric characteristics : 13 psychometric criteria have been used. Results show that the majority of tools don't provide sufficient evidence of their validity, reliability, sensibility and specificity. This characteristics are often missing in manuals. Speech-language pathologists should be cautious when using standardized assessment tools and when interpreting the scores obtained by individuals. Tests developers should give more complete information about psychometrics properties.

KEY WORDS

assesment, psychometrics, reliability, test, validity, written language