Département d'Orthophonie
Gabriel DECROIX

MEMOIRE

En vue de l'obtention du

Certificat de Capacité d'Orthophoniste

présenté par :

Laura COLLI-VAAST

soutenu publiquement en juin 2018 :

Caractéristiques psychométriques des tests de langage écrit chez l'enfant

MEMOIRE dirigé par :

Perrine DANCHIN, orthophoniste et enseignante vacataire, Hôpital Saint-Vincent de Paul et département d'orthophonie, Lille

Lucie MACCHI, maître de conférences et orthophoniste, département d'orthophonie, laboratoire STL, Lille

Membres du jury :

Loïc GAMOT, orthophoniste et enseignant, département d'orthophonie, Lille

et les directrices du mémoire

Lille – **2018**

Remerciements

Je remercie mes directrices de mémoire Lucie Macchi et Perrine Danchin pour leurs conseils et leur disponibilité tout au long de ce travail.

Je tiens également à remercier ma famille pour son soutien et sa patience.

Résumé :

De nombreux outils ont été développés pour évaluer le langage oral et écrit chez l'enfant. Cependant peu de recherches se sont intéressées aux qualités psychométriques des outils francophones. L'objectif principal de notre étude est de réaliser l'inventaire des qualités psychométriques des tests francophones de langage écrit chez l'enfant. Notre objectif secondaire est de faire une étude exploratoire de la qualité de traitement de chacun des critères psychométriques examinés dans les tests. Un travail parallèle concernant le langage oral est effectué dans le mémoire d'Herman (2018). Vingt-six tests ou batteries actuellement diffusés ou commercialisés en France ont été analysés en fonction de quinze critères psychométriques sélectionnés à partir de la littérature. Nos principaux résultats suggèrent que peu de ces outils répondent aux standards psychométriques actuels et que la qualité des informations présentes dans les manuels des tests reste insuffisante.

Mots-clés :

Test, langage écrit, évaluation, psychométrie.

Abstract :

Many tools have been created to assess spoken and written language in children. Only few studies have been carried out on the psychometric characteristics of these francophone tools. The main objective of our study is to draw up the inventory of the psychometric characteristics of francophone written language tests in children. Our second objective is to conduct an exploratory study on the quality of each psychometric criteria examined in the tests. A similar work about spoken language is carried out in Herman's master thesis (2018). Twenty-six tests or batteries of tests currently available in France have been analyzed through fifteen psychometric criteria set out in existing studies. Our results show that only few tools meet the established psychometric standards. They also show the poor quality of the information contained in their respective manuals.

Keywords :

Tests, written language, assessment, psychometrics.

Table des matières

Introduction

La démarche clinique de l'orthophoniste nécessite une évaluation fiable et précise des compétences du patient afin de mettre en place une rééducation adaptée. De nombreux outils d'évaluation du langage ont été développés afin de répondre aux besoins de diagnostic des troubles du langage chez l'enfant. Plusieurs auteurs se sont intéressés à ces outils, en particulier aux Etats-Unis d'Amérique et au Canada. Ces différentes études ont montré la difficulté du choix des tests à employer en clinique et le peu d'influence des qualités des épreuves sur leur fréquence d'utilisation par les orthophonistes (Betz, Eickhoff & Sullivan, 2013 ; Kerr, Guildford, & Bird, 2003). Le choix éclairé d'un outil se fonde notamment sur la connaissance de son efficacité et de ses qualités psychométriques. L'impact de ces qualités sur l'interprétation des résultats et la pratique clinique guide également ce choix. Cette démarche s'inscrit dans le cadre de la pratique fondée sur les preuves (*Evidence-Based Practice* en anglais) et permet aux professionnels de prendre des décisions cliniques en tenant compte des données probantes de la recherche scientifique, complétée par l'expertise du thérapeute et les préférences du patient (Durieux, Palseau & Maillart, 2012).

Une des premières études s'intéressant aux qualités des tests, concernait trente tests américains de langage et d'articulation pour les enfants d'âge préscolaire. Elle portait sur dix critères psychométriques liés à la validité, la fidélité et aux caractéristiques de l'échantillon d'étalonnage. Seulement 20 % des tests remplissaient la moitié des critères et la plupart des informations nécessaires pour évaluer les qualités des tests étaient absentes des manuels (McCauley & Swisher, 1984a). Plus récemment, d'autres auteurs ont également mis en évidence l'absence de certains critères psychométriques dans les outils d'évaluation du langage chez l'enfant (Flipsen & Ogiela, 2015 ; Friberg, 2010 ; Kirk & Vigeland, 2014).

Peu de travaux ont été réalisés à propos des outils francophones, notamment au sujet de leurs qualités psychométriques (Bertrand, Fluss, Billard & Ziegler, 2010 ; Bouchard, Fitzpatrick & Olds, 2009 ; Leclercq & Veys, 2014). L'objectif principal de ce travail est de décrire les caractéristiques psychométriques des tests francophones de langage écrit chez l'enfant. Un travail similaire concernant le langage oral est réalisé dans le cadre du mémoire d'Herman (2018). Devant le manque d'informations nécessaires pour évaluer les qualités des tests dans les manuels évoqué par certains auteurs, nous avons pour objectif secondaire de réaliser une étude exploratoire de la qualité psychométrique de chaque critère psychométrique des tests de langage écrit étudiés.

Nous commencerons par présenter l'utilisation des tests de langage écrit chez l'enfant au sein de la démarche d'évaluation orthophonique. Puis nous présenterons les différents principes psychométriques des tests et les critères utilisés dans la littérature pour évaluer ces critères. Nous détaillerons ensuite notre méthodologie pour choisir les tests et les critères psychométriques d'une part, et pour les analyser d'autre part. Nous présenterons enfin nos résultats et les discuterons.

Contexte théorique, buts et hypothèses

1. Evaluation et tests

D'après les articles 1 et 2 du décret n°2002-721 du 2 mai 2002 relatif aux actes professionnels et à l'exercice de la profession d'orthophoniste : « L'orthophonie consiste à prévenir, à évaluer et à prendre en charge, aussi précocement que possible, par des actes de rééducation constituant un traitement, les troubles de la voix, de l'articulation, de la parole, ainsi que les troubles associés à la compréhension du langage oral et écrit et à son expression (…). Dans le cadre de la prescription médicale, l'orthophoniste établit un bilan qui comprend le diagnostic orthophonique, les objectifs et le plan de soins (…) ». Ainsi, la pratique orthophonique ne saurait se passer d'une évaluation diagnostique préalable à toute éventuelle intervention. Il paraît donc important de revenir sur les fondements d'une évaluation.

1.1. Evaluation : principes généraux et évaluation du langage écrit

1.1.1. Principes généraux d'une évaluation

Une procédure d'évaluation commence par une collecte de données fiables et valides. L'interprétation de ces données permet d'aboutir ou non à un diagnostic, c'est-à-dire à l'identification d'un trouble donné (Shipley & McAfee, 2009). Les orthophonistes se fondent sur ces informations recueillies pour formuler des hypothèses diagnostiques et rééducatives. Les décisions cliniques sont basées sur les résultats de l'évaluation. Pour que cette dernière soit pertinente, Shipley et McAfee (2009) proposent de suivre cinq principes généraux :

- Une évaluation est approfondie : le professionnel recueille un maximum d'informations pour poser un diagnostic précis et proposer des recommandations les plus appropriées possibles.
- Une évaluation vise à collecter des informations variées : elle combine les observations cliniques et les résultats aux différents tests utilisés, notamment la comparaison du résultat du patient à une norme.
- Une évaluation est valide : elle mesure ce qu'elle est censée mesurer.
- Une évaluation est fiable : elle reflète les compétences réelles du patient, ce qui implique idéalement que des mesures répétées des mêmes comportements aboutissent aux mêmes résultats.
- Une évaluation est adaptée au patient : à son âge, son sexe, son niveau et sa culture.

1.1.2. Evaluation du langage écrit

Les troubles spécifiques des apprentissages, en particulier ceux du langage écrit sont caractérisés par différents critères (Fletcher & al., 2004) :

- Le critère d'écart, de discordance : il existe un décalage entre les performances aux épreuves liées au trouble et celles liées au niveau cognitif global de l'enfant.
- Le critère d'exclusion : le trouble n'a pas comme cause primaire un retard global, un handicap sensoriel, un environnement défavorable ou un trouble mental.
- Le trouble est dû à des facteurs intrinsèques à l'enfant, son origine est neurobiologique.

Selon Casalis, Leloup et Bois Parriaud (2013), la dyslexie est un trouble persistant d'apprentissage de la lecture ne s'expliquant pas par des causes liées à l'environnement ou à des contextes de déficit cognitif ou sensoriel. Elle se manifeste par un défaut d'installation des mécanismes d'automatisation de l'identification des mots écrits.

Le diagnostic de dyslexie repose sur différents éléments : un retard en lecture d'au moins deux ans (ou un score inférieur ou égal au percentile 5), une intelligence normale, une absence de déficit sensoriel et des conditions environnementales favorables (Casalis & al., 2013). Les compétences de lecture, à savoir l'identification de mots écrits et la compréhension écrite, sont évaluées à l'aide de tests leximétriques et de tests déterminant les processus d'intégration syntaxique et sémantique (Ecalle & Magnan, 2006).

Pour évaluer le langage écrit, et ainsi recueillir les informations nécessaires au diagnostic de dyslexie, les orthophonistes utilisent des tests qu'il convient de définir.

1.2. Tests : définitions, types et utilisation en orthophonie

1.2.1. Définition

Selon Hogan (2012), le test est une méthode ou un outil standardisé qui fournit de l'information sur un échantillon de comportements ou de processus cognitifs sous une forme quantifiée.

1.2.2. Différents types de tests en orthophonie

Différents types d'outils sont à disposition des orthophonistes en fonction de l'objectif recherché (Coquet, 2013). En premier lieu, il existe des tests de dépistage, rapides et faciles d'utilisation qui visent la recherche de signes faisant suspecter une difficulté et le cas échéant, l'orientation vers une évaluation diagnostique. Il existe également des tests de première ligne situant les compétences de l'enfant au sein de sa classe d'âge ou de niveau scolaire, et permettant de décider d'arrêter le bilan de langage ou de le poursuivre en le complétant par des sous-épreuves plus précises et détaillées. Les tests ciblant un domaine spécifique, ou de deuxième ligne, proposent cette exploration plus fine et approfondie. Il existe aussi des batteries regroupant plusieurs épreuves afin d'établir un profil global de compétences. Enfin, les lignes de base sont l'un des principaux outils permettant de comparer les performances pré- et post-rééducation pour mesurer l'efficacité de la prise en charge.

1.2.3. Utilisation en orthophonie

Vrignaud, Castro et Mogenet (2003) ont élaboré pour la Société Française de Psychologie (SFP), la version française des recommandations internationales sur l'utilisation des tests (version originale : *International Test Commission*, 2000). Ces recommandations ont pour objectif de fournir des critères de compétences pour tout utilisateur de test : « Un utilisateur de tests compétent utilise les tests de manière appropriée, de manière professionnelle, et de manière éthique, en prenant en considération les besoins et les droits de ceux qui sont impliqués

dans le processus de passation des tests, les justifications de la passation, et le contexte, au sens large, dans lequel la passation se déroule » (Vrignaud & al., 2003, p.13).

Pour évaluer les troubles du langage, les orthophonistes ont à leur disposition deux grands types de tests : les tests étalonnés et les tests à critères (Lefebvre & Trudeau, 2005). Les tests étalonnés visent à vérifier la présence ou l'absence d'un trouble via une procédure formelle. Les tests à critères sont, quant à eux, fondés sur l'observation ou non d'habiletés spécifiques, sans recours à un groupe de référence. Dans le cadre de notre travail, nous nous intéresserons plus particulièrement aux tests étalonnés : « Ce sont des épreuves formelles administrées selon une procédure rigide visant à vérifier si un problème est présent ou non » (Lefebvre & al., 2005, p.17).

McCauley et Swisher (1984b) ont relevé différents types d'erreurs dans l'utilisation des tests étalonnés pour évaluer le langage : l'utilisation d'un score âge-équivalent comme score unique (ce score ne prend pas en compte l'étendue des performances normales), la comparaison de scores à des tests différents pour établir le profil d'un patient, l'échec à un seul item pour définir un déficit, les mesures répétées avec le même test pour évaluer les progrès du patient sans considérer l'effet test-retest. Selon Kerr, Guildford et Bird (2003), beaucoup d'orthophonistes ne sont pas conscients de l'impact des erreurs d'utilisation décrites par McCauley et Swisher (1984b) et continuent de les commettre, notamment parce que peu de cliniciens ont confiance en leurs connaissances psychométriques. Effectivement, ces dernières ont fait ou font probablement l'objet de peu d'enseignements en formation initiale ou continue, comparativement à d'autres cursus, comme celui de psychologie. Plus récemment, Betz, Eickhoff et Sullivan (2013) révèlent que dans le cadre du diagnostic de trouble spécifique du langage oral, les qualités psychométriques des tests n'influencent pas leur fréquence d'utilisation par les orthophonistes. C'est en effet la date de parution des tests qui est le seul facteur corrélé à leur fréquence d'utilisation.

Dans le but de guider les orthophonistes dans leurs choix de tests étalonnés les plus appropriés, plusieurs auteurs ont étudié les qualités des tests. Par ces travaux, ils tentent d'apporter des solutions permettant la sélection, l'administration et l'interprétation adéquates des tests étalonnés dans différents domaines. Bertrand, Fluss, Billard et Ziegler (2010) proposent de comparer différents tests de lecture et d'adopter une stratégie d'évaluation en fonction de leurs qualités psychométriques. Leclercq et Veys (2014) proposent une réflexion sur le choix de tests standardisés lors du diagnostic de dysphasie. D'autres auteurs proposent des revues psychométriques de tests ou des solutions plus globales. Ils proposent de suivre différentes étapes concernant la sélection, l'administration et l'interprétation des tests standardisés, ou de construire un arbre décisionnel aidant au choix des outils et basé sur leurs qualités psychométriques (Bouchard, Fitzpatrick & Olds, 2009 ; Flipsen & Ogiela, 2015 ; Friberg, 2010 ; Kirk & Vigeland, 2014 ; Lefebvre & Trudeau, 2005).

Pour aller plus loin, il est fondamental de définir les différents principes psychométriques.

2. Principes psychométriques

Selon Hogan (2012), les trois questions fondamentales de la psychométrie portent sur les normes (cadre d'interprétation des résultats du test), la fidélité (stabilité d'une mesure) et la validité (ce que le test mesure réellement).

2.1. De la sensibilité aux normes d'un test

Un test vise à mettre en évidence des différences inter ou intra-individuelles. Il est considéré sensible lorsqu'il permet de distinguer finement les individus relativement à ce qu'il est censé mesurer. Ainsi, la sensibilité ou le pouvoir séparateur d'un test est sa capacité à pouvoir discriminer les participants entre eux (Anceaux & Sockeel, 2006). L'examen de la sensibilité est une étape importante et sera en partie la base du choix du type d'étalonnage. L'évaluation de la sensibilité d'un instrument se fait sur plusieurs plans : au niveau global par l'examen de la distribution des scores bruts et plus précisément par l'étude de la sensibilité de chaque item. De plus, Bertrand, Fluss, Billard et Ziegler (2010) montrent que pour mesurer le degré d'efficacité d'un test diagnostique, les meilleurs indices sont la sensibilité et la spécificité. La sensibilité mesure la capacité d'un test à détecter les individus aux performances pathologiques et la spécificité est la probabilité que le test soit négatif sachant que l'individu est sain. Pour qu'un test soit considéré sensible et spécifique, les valeurs de ces indices doivent être supérieures à 0,80 (Bernaud, 2007).

Une note brute n'a pas de signification en soi. C'est la transformation, appelée étalonnage, de la distribution des scores bruts en une nouvelle distribution calquée sur une distribution théorique connue (ex. loi normale, figure 1) qui donne la possibilité de situer la personne dans son groupe de référence : l'échantillon d'étalonnage. Cet échantillon est supposé représentatif d'une population suivant différents critères (ex. âge, sexe, origine géographique, catégorie socio-professionnelle) à une certaine date. Ces informations, ainsi que la taille de l'échantillon, sont importantes à connaitre car elles permettent d'évaluer la représentativité d'un groupe de référence. Celui-ci doit correspondre le mieux possible à la population cible en ce qui concerne ces critères : c'est le degré de représentativité du groupe de référence (Hogan, 2012).

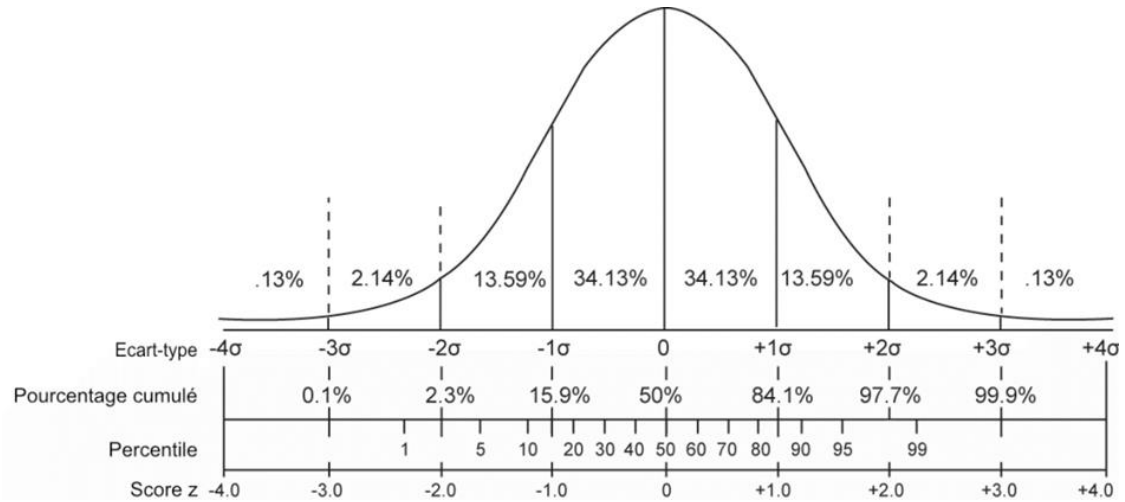

Figure 1: Courbe de distribution normale (adapté de Bartram, 1994, p. 22)

Selon Hogan (2012), il existe trois grandes catégories de normes de tests : les rangs percentiles, les scores pondérés et les normes de développement. Premièrement, le principe des rangs percentiles est de partager l'étendue des effectifs en 100 classes comportant un même pourcentage de sujets. Ils sont le plus souvent utilisés si la distribution des scores bruts ne suit pas une loi normale, par exemple en cas de petit échantillon. Deuxièmement, la conversion des scores bruts en scores pondérés s'effectue en plusieurs étapes et nécessite que la distribution des scores bruts soit normale. La première étape est de traduire le score brut en score *z* où la moyenne et l'écart-type (indice de dispersion autour de la moyenne) sont fixés respectivement à zéro et un. Cette étape permet une normalisation (centrée réduite ici) de la distribution des scores. La deuxième étape consiste à convertir les scores *z* en un nouveau système comportant une moyenne *M* et un écart-type *ET* choisis de façon arbitraire. Par exemple les scores *T* où *M* = 50 et *ET* = 10 ou les notes de QI où *M* = 100 et *ET* = 15. Troisièmement, les normes de développement sont utilisées lorsque le trait mesuré se développe systématiquement avec le temps. Les deux types de normes de développement le plus souvent utilisés sont les équivalents d'âge et les équivalents de niveau scolaire.

2.2. Fidélité

On considère qu'un outil de mesure est fidèle ou fiable lorsque son utilisation sur les mêmes participants entraîne des résultats identiques (Anceaux & Sockeel, 2006). La fidélité est d'autant plus importante que la proportion du score attribuable au hasard est faible. Une instabilité dans les résultats peut être due à deux types d'erreurs de mesure. La première est l'erreur systématique qui ne dépend pas du test (ex. le niveau d'intelligence d'un enfant hispanophone passant un test en français sera systématiquement sous-estimé). La deuxième est l'erreur aléatoire, appelée également erreur type de mesure (Grégoire, 2006). Les sources de cette erreur sont dues à la personne testée (fatigue, distraction, motivation) et à l'examinateur (conditions d'administration et de cotation du test). L'erreur type de mesure permet de déterminer une zone autour du score observé dans laquelle le score vrai (score sans erreur) a de fortes chances de se trouver. Cette zone correspond à un intervalle de confiance dont les bornes sont fixées en fonction du taux d'erreur accepté et d'un seuil de probabilité (α = 5 % le plus souvent). Il existe différentes méthodes permettant d'estimer la fidélité d'un test, chacune ciblant une source d'erreur de mesure différente (Hogan, 2012).

La **fidélité test-retest** s'intéresse à la stabilité des scores à un test dans le temps pour une même population. Cette mesure s'intéresse aux variations d'administration du test, par exemple les conditions physiques et mentales des participants (Hogan, 2012). Elle s'exprime par un coefficient de corrélation *r* entre les résultats obtenus aux deux passations. Selon Grégoire (2006), pour que les scores d'un test puissent avoir une réelle valeur pratique, la fidélité de ce test doit être au moins égale à 0,80 (*r* = 0,80). Il est également important de veiller à ce que le délai entre les deux passations soit suffisamment long (quelques jours à quelques semaines) pour limiter l'apprentissage du test (Lefebvre & Trudeau, 2005). Il est préférable de calculer ce coefficient sur un échantillon d'au moins cent sujets pour que les mesures statistiques soient fiables (Kline, 1994).

La **fidélité inter-juges** ou inter-observateurs est la constance des résultats obtenus par deux évaluateurs différents. Elle est mesurée par un coefficient de corrélation entre les scores obtenus : le coefficient de fidélité inter-observateurs. Ce coefficient de corrélation (*r* de Bravais-Pearson) doit être supérieur à 0,80 pour être satisfaisant (Bernaud, 2007). Cette mesure est meilleure si les qualités de standardisation du test, c'est-à-dire les procédures de passation, de cotation et d'interprétation sont d'une part indiquées dans le manuel et d'autre part respectées par l'examinateur.

La méthode des versions parallèles ou méthode des formes équivalentes s'intéresse au degré de concordance des résultats obtenus avec deux versions d'un même outil. Cette méthode permet de montrer l'équivalence entre deux versions d'un même instrument (Hogan, 2012).

Enfin, la **cohérence interne** ou homogénéité est le degré de constance des réponses d'un individu aux différents items d'un même test (Anceaux & Sockeel, 2006). On l'évalue le plus souvent par la méthode de bissection (appelée « split-half » en anglais) qui consiste à mesurer la corrélation entre les résultats de deux groupes d'items équivalents issus du même test. Si cette corrélation est forte, alors le test mesure bien ce qu'il doit mesurer par ses différents items. Une autre méthode visant à mesurer la cohérence interne est l'application de la formule de l'alpha de Cronbach (α). C'est une méthode d'analyse des variances qui renseigne sur le degré d'homogénéité des items (Hogan, 2012). Une valeur de 0,70 atteste d'un coefficient alpha correct, 0,80 d'un alpha satisfaisant et 0,90 d'un alpha élevé (Bernaud, 2007).

2.3. Validité

La validité d'un test est le plus souvent définie comme le degré avec lequel il mesure ce qu'il prétend mesurer (Hogan, 2012). Le concept de validité est central car si un test n'est pas valide, alors les interprétations des résultats seront erronées. Il existe cinq types de validité : la validité de surface ou d'apparence, la validité de contenu, la validité concordante, la validité prédictive et la validité de construit (Lefebvre & Trudeau, 2005).

La **validité de surface ou d'apparence** est le fait qu'un test semble valide. Elle correspond au « sentiment subjectif » qu'ont les évaluateurs et les personnes évaluées de la validité du test (Bernaud, 2007). Une bonne validité apparente peut engendrer chez les personnes évaluées des conduites positives telles qu'une plus grande implication dans la tâche et une plus grande sincérité des réponses. Kline (1994) préconise de choisir un test avec une bonne validité apparente pour établir une bonne relation avec les personnes testées.

La **validité de contenu** est le degré de représentativité des items en tant qu'échantillon du domaine de comportements à mesurer. Pour garantir une bonne validité de contenu, il est essentiel de s'assurer que le contenu soit défini clairement, que les items couvrent tous les aspects du contenu de manière proportionnelle, et que le test ne contienne que des items pertinents (Haynes, Richard & Kubany, 1995). Des experts évaluent le niveau de validité de contenu selon la pertinence, la représentativité, la spécificité et la clarté des items. La méthode classique d'analyse des items permet d'évaluer leur niveau de difficulté et leur validité par des indices (Bernaud, 2007). L'indice de difficulté *D*, est le ratio du nombre de personnes qui

réussissent l'item sur le nombre de personnes qui y ont répondu. Cet indice permet d'éliminer les items trop faciles (*D* faible) ou trop difficiles (*D* élevé), et de classer les items par difficulté croissante. L'indice de discrimination *R* permet d'évaluer la validité de l'item et se mesure par un coefficient de corrélation entre l'item et sa dimension d'appartenance ($0{,}30 < r < 0{,}80$). Une autre méthode pour analyser les items est l'analyse factorielle, qui a l'avantage de permettre l'organisation du test en différentes sous-échelles ou subtests.

La **validité concordante ou concourante** témoigne de la cohérence observée entre les résultats au test étudié et ceux d'un autre test considéré comme valide dans le domaine. Une bonne validité concordante pour un test à visée diagnostique, par exemple, démontre la qualité de ce test en tant qu'outil diagnostique. La validité concourante se mesure le plus souvent par une corrélation entre le test et un test déjà validé ou « gold standard » (Rondal, 1997). Ce coefficient de corrélation *r* est considéré suffisant à partir de 0,40 à condition qu'il soit calculé sur un échantillon d'au moins cent sujets (Kline, 1994).

La **validité prédictive** est la capacité du test à prédire un comportement futur du sujet à partir de son résultat. Par exemple, le résultat au test de conscience phonologique en maternelle prédit le futur résultat en lecture au CP (Lefebvre & Trudeau, 2005). Une bonne validité prédictive montre la capacité pronostique du test. Elle est mesurée par un coefficient de corrélation entre les scores obtenus au test administré à un échantillon représentatif, à deux moments différents allant de quelques semaines à quelques années (Bartram, 1994). Cette mesure dépendant d'un nombre important de variables, Kline (1994) propose une valeur seuil de 0,30 pour le coefficient de corrélation *r*.

La **validité de construit ou validité théorique** vise à vérifier l'acceptation théorique d'un instrument de mesure (Smith, 2005). Elle repose sur le lien entre le test et le modèle théorique du construit (définition conceptuelle ou théorique de ce qui est mesuré dans le test). Selon Bernaud (2007), la validité de construit repose sur deux composantes indissociables : « une procédure déductive », qui consiste à vérifier si les observations sont conformes au modèle théorique; et une « procédure inductive », qui permet de faire des inférences sur les traits mesurés à partir des scores au test. Bernaud (2007) cite deux méthodes d'analyse de variances pour évaluer la validité de construit : l'analyse multitraits multiméthodes (MTMM) et l'analyse factorielle. Selon Kline (1994), on peut considérer qu'un test a une bonne validité de construit s'il présente de bonnes validités concourantes et prédictives.

2.4. Critères psychométriques des tests

Dans le but de construire notre propre liste de critères, nous répertorions ici les qualités psychométriques les plus souvent prises en compte dans la littérature pour analyser les tests. Nous décrivons ensuite brièvement les résultats des principales revues psychométriques des tests.

Dans leur étude, McCauley et Swisher (1984a) ont établi une liste de dix critères de qualité pour les tests étalonnés :

1) La description de l'échantillon d'étalonnage.
2) La taille de l'échantillon : un minimum de cent sujets par sous-groupe est recommandé.
3) L'analyse des différents items dans leur construction et sélection.
4) L'indication de la tendance centrale et de la variabilité des scores pour chaque sous-groupe.
5) La validité concourante indique la capacité diagnostique du test.
6) La validité prédictive indique la capacité pronostique du test.
7) La fidélité test-retest montre la stabilité des résultats dans le temps et doit être présentée pour chaque sous-groupe.
8) La fidélité inter-juges montre la constance des résultats quel que soit l'examinateur.
9) La description précise des procédures d'administration, de cotation et d'interprétation du test. Si elle est absente, la comparaison des performances du sujet aux normes proposées est contestable.
10) L'indication des qualifications requises pour l'examinateur afin de garantir la qualité des résultats obtenus.

Sur la base de ces critères, plusieurs auteurs ont réalisé d'autres études sur les qualités psychométriques des tests : celle de Hutchinson (1996) qui propose un guide de lecture des manuels des tests, celle de Bouchard et al. (2009) avec seize critères, puis celle de Friberg (2010) avec onze critères. Plus récemment, Flipsen et Ogiela (2015) ont étudié dix tests publiés depuis 1990 sur la base des dix critères de McCauley et Swisher (1984a) et sept critères additionnels (tableau 1).

Tableau 1 : Critères psychométriques étudiés dans la littérature.

Critères/Auteurs	McCauley & Swisher (1984a)	Hutchinson (1996)	Bouchard & al. (2009)	Friberg (2010)	Flipsen & Ogiela (2015)	Nombre total d'études retenant le critère
Sources (critères communs avec ceux de McCauley et Swisher / critères ajoutés)	10 / 0	7 / 7	7 / 9	10 / 1	9 / 6	
1) Description de l'échantillon d'étalonnage	O	O	O	O	O	5
2) Validité concourante	O	O	O	O	O	5
3) Fidélité test / re-test	O	O	O	O	O	5
4) Fidélité inter-juges	O	O	O	O	O	5
5) Taille de l'échantillon	O	N	O	O	O	4
6) Tendance centrale et variabilité des scores	O	O	N	O	O	4
7) Validité prédictive	O	N	O	O	O	4
8) Procédures d'administration, de cotation et d'interprétation du test	O	N	O	O	O	4
9) Qualifications requises par l'examinateur	O	O	N	O	O	4
10) Objectifs du test	N	O	O	O	N	3
11) Analyse statistique des items	O	O	N	O	N	3
12) Erreur standard de mesure	N	O	O	N	O	3
13) Représentation des sujets dans les extrêmes	N	O	O	N	N	2
14) Validité de contenu	N	O	O	N	N	2
15) Validité de construit / Validité théorique	N	O	O	N	N	2
16) Précision diagnostique	N	O	N	N	O	2
17) Effets planchers / plafonds	N	N	O	N	N	1
18) Différence de genre discutée	N	N	N	N	O	1
19) Définitions des construits mesurés par le test	N	N	O	N	N	1
20) Différence de dialectes discutée	N	N	N	N	O	1
21) Analyse des voyelles	N	N	N	N	O	1
22) Analyse des processus phonologiques	N	N	N	N	O	1
23) Cohérence interne	N	N	O	N	N	1
24) Année de publication et de standardisation de l'échantillon	N	N	O	N	N	1
25) Niveaux conjoints de sensibilité et spécificité	N	O	N	N	N	1
Nombre total de critères par étude	10	14	16	11	15	

Note. O : Le critère est présent ; N : Le critère est absent

Les critères psychométriques étudiés dans tous les articles sont la description de l'échantillon d'étalonnage, la validité concourante, la fidélité test-retest et la fidélité inter-juges. Les critères retenus dans quatre études sur cinq sont : la taille de l'échantillon, la tendance centrale, la validité prédictive, les procédures d'administration, de cotation et d'interprétation du test et les qualifications requises par l'examinateur. Les auteurs de ces revues psychométriques ont étudié en moyenne treize critères.

Nous décrivons maintenant les résultats principaux de ces différentes études concernant les tests anglophones puis francophones de langage oral. Une des premières revues psychométriques portait sur trente tests anglophones de langage oral chez l'enfant d'âge préscolaire (McCauley & Swisher, 1984a). La moitié de ces tests répond à moins de deux critères psychométriques sur dix, et seulement trois tests possèdent quatre critères. Les qualités

les plus fréquemment absentes sont la validité et la fidélité. Dans son article, Friberg (2010) étudie neuf tests diagnostiques anglophones évaluant le langage oral d'enfants d'âge pré- et scolaire. Chaque test remplit au moins huit critères sur onze, et cinq tests en ont dix. Cette analyse est la seule portant sur les tests diagnostiques, ce qui explique probablement ses meilleurs résultats. Une autre étude concerne dix tests anglophones publiés depuis 1990, et évaluant la production de mots (phonologie et articulation) d'enfants de 1 an 6 mois à 21 ans 11 mois (Flipsen & Ogiela, 2015). Tous les tests répondent à au moins trois critères sur les dix proposés par McCauley et Swisher (1984a), et neuf des dix tests en possèdent au moins cinq (M = 5,6 ; ET = 1,17). La plupart des tests étudiés ont quelques critères additionnels sur les sept ajoutés.

Concernant les tests francophones, l'analyse de Bouchard, Fitzpatrick et Olds (2009) porte sur 31 tests utilisés au Québec pour évaluer le langage d'enfants d'âge préscolaire. En moyenne, chaque test remplit 6,81 critères sur 16 (ET = 2,91 ; Min = 2 ; Max = 14). La majorité des tests ne répond pas à plusieurs des critères définissant un test standardisé de qualité. Enfin, une dernière étude considère cinq batteries francophones évaluant le langage oral d'enfants d'âge scolaire (Leclercq & Veys, 2014). Les batteries remplissent en moyenne 5,9 critères sur 13 (ET = 1,88 ; Min = 3 ; Max = 8). Seules deux batteries possèdent plus de la moitié des critères psychométriques étudiés.

Les résultats de ces différentes études montrent que peu de tests ou batteries de langage oral détiennent les qualités psychométriques attendues pour un test standardisé. A notre connaissance, aucun article n'a été publié sur les valeurs psychométriques des tests de langage écrit chez l'enfant.

3. Buts et hypothèses

L'objectif principal de ce travail est de réaliser l'inventaire des caractéristiques psychométriques des tests de langage écrit utilisés chez l'enfant, en Europe francophone. En effet, comme précisé précédemment, ce travail n'a jamais été réalisé, à notre connaissance, et peut s'avérer utile. Notre objectif secondaire est de faire une étude exploratoire de la qualité de traitement de chacun des critères psychométriques dans les tests et batteries de langage écrit, afin de connaître les critères les moins et ceux les plus satisfaisants dans ces tests et batteries.

Nous supposons que les résultats de notre étude iront dans le même sens que ceux des revues psychométriques des tests décrits dans la partie 2.4. Notre première hypothèse est que peu d'outils évaluant le langage écrit chez l'enfant répondent aux principaux standards psychométriques actuels. Notre deuxième hypothèse est que la qualité des informations présentes dans les manuels des tests ou batteries de langage écrit est insuffisante. Concernant l'étude de la qualité des critères psychométriques, nous nous plaçons dans une perspective exploratoire. En effet, le manque de données dans la littérature nous conduit à ne pas formuler d'hypothèse précise.

Méthode

Dans cette partie, nous présenterons notre méthodologie de sélection des tests et des critères psychométriques, puis nous décrirons notre échelle de cotation de ces critères. Enfin, nous exposerons comment nous mesurerons les fidélités intra et inter-juges.

1. Tests

Nous avons sélectionné les tests et batteries de langage écrit figurant sur au moins une des deux listes qui répertorient les tests ou batteries actuellement utilisés par les orthophonistes et les chercheurs francophones d'Europe (France, région francophone de la Belgique, Suisse romande, Luxembourg, Monaco). La première liste a été établie en 2011 par l'Union Nationale pour le Développement de la Recherche et de l'Evaluation en Orthophonie (UNADREO), une société savante, principale interlocutrice de la Haute Autorité de Santé en matière d'orthophonie. Cette liste concerne des tests orthophoniques de langage oral et écrit. Deuxièmement, nous avons utilisé une liste de tests et batteries employés par les logopèdes exerçant en Belgique francophone. Cette liste est disponible sur le site de l'Institut National d'Assurance Maladie-Invalidité (INAMI, 2018). Au sein de ces deux listes, nous n'étudions que les tests de langage écrit actuellement diffusés ou commercialisés. Ce critère nous a conduit à écarter de notre analyse quatre tests sur les trente de départ : le Test de Niveau d'Orthographe (Doutriaux & Lepez, 1980), la NBNO (Ravard & Ravard, 1991), l'Evaluation des compétences de lecteur (Aubret & Blanchard, 1991) et la batterie de langage oral et écrit ORLEC L3 (Lobrot, 2004).

Au total, vingt-six tests ou batteries de langage écrit sont donc étudiés (tableau 2) du point de vue des critères psychométriques décrits dans la partie 2 (choix des critères psychométriques).

Nous nous sommes procuré les différents manuels des tests ou batteries soit auprès d'orthophonistes les possédant, soit à la testothèque du département d'orthophonie de Lille. Nous avons également recherché des articles pouvant apporter des informations complémentaires aux manuels. Les dates de parution des tests ou batteries s'étendent de 1995 à 2012 (M = 2005 ; ET = 5,01 ans).

Tableau 2 : Les tests et batteries de langage écrit étudiés dans notre mémoire.

Modalités langagières	Nom du test ou batterie	Auteur(s)	Date de parution	Tranche d'âge évaluée en années
Langage écrit	Alouette-R	Lefavrais	2005	[6 ; 0 – 16 ; 0]
	Analec	Inizan	1998	[8 ; 0 – 15 ; 0]
	BALE	Jacquier-Roux, Lequette, Pouget, Valdois et Zorman	2010	[8 ; 0 – 11 ; 0]
	Batelem-R	Savigny, Barbier, Coupey Le Roy, Girard et Roussel	2001	[5 ; 0 – 8 ; 0]
	BELEC	Mousty, Leybaert, Alégria, Content et Morais	1995	[8 ; 0 – 10 ; 0]
		Mousty et Leybaert	1999	
	BELO	George et Pech-Georgel	2008	[5 ; 8 – 8 ; 6]
	BLI	Khomsi	2003	[7 ; 3 – 10 ; 3]
	Chronosdictées	Baneath, Boutard et Alberti	2006	[7 ; 6 – 14 ; 7}
	ECHAS-C	Simonart	1998	[9 ; 0 – 12 ; 0]
	ECL-Collège	Khomsi, Nanty, Pasquet et Parbeau-Guéno	2005	[11 ; 0 – 15 ; 0]
	Forme noire	Maeder	2010	[9 ; 0 – 12 ; 0]
	LMC-R	Khomsi	1999	[7 ; 4 – 12 ; 5]
	ODEDYS	Jacquier-Roux, Valdois, Zorman, Lequette et Pouget	2005	[7 ; 0 – 10 ; 0]
	Orthographe au collège	Thibault	2008	[12 ; 0 – 15 ; 0]
	PEDA 1C	Simonart	1998	[7 ; 0 – 8 ; 0]
	Phonolec collège	Plaza, Gatignol, Oudry et Robert-Jahier	2011	[12 ; 0 – 15 ; 0]
	Timé-2	Ecalle	2003	[6 ; 2 – 8 ; 3]
	Timé-3	Ecalle	2006	[7 ; 4 – 14 ; 3]
	Vol du P.C.	Boutard, Claire et Gretchanovsky	2006	[11 ; 0 – 20 ; 0]
Langage oral et écrit	CLéA	Pasquet, Parbeau-Guéno et Bourg	2014	[2 ; 5 – 15 ; 0]
	E.CO.S.SE.	Lecocq	1996	[4 ; 0 – 14 ; 4]
	EXALang 5-8	Thibault, Helloin et Croteau	2010	[11 ; 5 – 15 ; 0]
	EXALang 8-11	Thibault, Helloin et Lenfant	2012	[6 ; 1 – 8 ; 2]
	EXALang 11-15	Helloin, Lenfant et Thibault	2009	[8 ; 8 – 10 ; 8]
	L2MA-2	Chevrie-Muller, Maillart, Simon et Fournier	2010	[7 ; 0 – 13 ; 0]
	TCS	Maeder	2006	[8 ; 5 – 16 ; 0]

2. Critères psychométriques

2.1. Choix des critères psychométriques

Sur la base des revues psychométriques présentées dans la partie 2.4 du contexte théorique, nous conservons les critères les plus fréquemment étudiés. Ainsi les douze critères psychométriques présents dans trois à cinq articles sur cinq (tableau 1) ont été retenus.

Nous en ajoutons trois autres. Premièrement, le praticien doit pouvoir démontrer que son interprétation des résultats est soutenue par des arguments théoriques et empiriques suffisants (Grégoire, 2006). La validité théorique ou de construit semble donc un critère fondamental. Deuxièmement, dans le but d'être fidèle, un test doit présenter une bonne cohérence interne et ainsi permettre de savoir si le test mesure bien ce qu'il est sensé mesurer (Bartram, 1994). Troisièmement, les meilleurs indices pour mesurer le degré d'efficacité d'un test sont la sensibilité et la spécificité (Bertrand, Fluss, Billard & Ziegler, 2010). C'est pourquoi nous

incluons les trois critères suivants dans notre liste (tableau 3) : validité théorique ou de construit, cohérence interne, sensibilité et spécificité. Notre liste comporte donc quinze critères psychométriques au total.

Tableau 3 : Critères psychométriques étudiés dans notre mémoire.

Normes	1) Description de l'échantillon d'étalonnage
	2) Taille de l'échantillon
	3) Mesure de tendance centrale
	4) Sensibilité / spécificité
Fidélité	5) Fidélité test-retest
	6) Fidélité inter-juges
	7) Cohérence interne
En lien avec la fidélité	8) Objectifs du test
	9) Instructions d'administration / cotation / interprétation
	10) Qualifications de l'examinateur
	11) Erreur standard de mesure
Validité	12) Validité de contenu / analyse formelle des items
	13) Validité théorique / de construit
	14) Validité concourante
	15) Validité prédictive

2.2. Cotation des critères psychométriques

Le but de notre travail est de réaliser l'inventaire des critères psychométriques des tests et batteries de langage écrit. Nous souhaitons également coter les critères de la façon la plus homogène et précise possible, ce qui nous oblige à recourir à une même échelle de cotation pour chaque critère. Nous nous sommes fondé sur une échelle de cotation à quatre niveaux (tableau 4). Cette cotation permet d'obtenir une note pour chaque test en fonction du degré de précision des informations apportées par les auteurs.

Tableau 4 : Echelle de cotation des critères psychométriques.

Critère / note	0	1	2	3
Description de l'échantillon d'étalonnage[1]	Non indiquée	1 ou 2 critère(s) principaux présent(s) et/ou pas de preuve de représentativité de l'échantillon	3 ou 4 critères principaux présents et/ou pas de preuve de représentativité de l'échantillon	4 critères principaux présents et/ou Preuve de représentativité de l'échantillon
Taille échantillon	Non indiquée	Moins de 50 sujets par sous-groupe	Entre 50 et 99 sujets par sous-groupe	100 sujets ou plus par sous-groupe
Mesure de tendance centrale	Non indiquée	Mesure de tendance centrale indiquée mais non adaptée à la distribution de la population contrôle[2]	Mesure de tendance centrale adaptée à la distribution de la population contrôle, mais indication imprécise ou confuse ou mal placée dans le manuel ou le logiciel.	Mesure de tendance centrale adaptée à la distribution de la population contrôle, et indication précise, claire et bien placée dans le manuel ou le logiciel.
Sensibilité / Spécificité (seuil = 0,80)	Non indiquées	Evocation de l'un ou l'autre, sans indication du seuil.	Indication du taux de sensibilité et/ou de spécificité mais < seuil	Indication du taux de sensibilité et/ou de spécificité ≥ seuil
Fidélité test-retest (seuil = 0,80)	Non indiquée	Indiquée mais incomplète (ex. pas de chiffres)	Indiquée mais < seuil, ou moins de 100 sujets par sous-groupe	Indiquée et ≥ seuil
Fidélité inter-juges (seuil = 0,80)				
Cohérence interne (seuil = 0,80)				
Objectifs du test	Non indiqués	*Pour les batteries* : Objectif général + précision des objectifs pour moins de 50 % des domaines étudiés et des épreuves *Pour les tests à épreuve unique* : objectif très général, flou et peu spécifique	*Pour les batteries* : Objectif général + objectifs pour 50 à 100 % des domaines étudiés et des épreuves	*Pour les batteries* : Objectif général + objectifs pour tous les domaines étudiés et toutes les épreuves *Pour les tests à épreuve unique* : objectif clair, précis et spécifique

[1] Les critères principaux de l'échantillon d'étalonnage considérés dans notre étude sont : l'âge, le sexe, l'origine géographique et la catégorie socio-professionnelle des parents.

[2] Ex. Les moyennes et écarts-types sont indiqués sans analyse de la distribution des données contrôle et/ou des éléments montrent que la distribution n'est pas gaussienne pour au moins une note et/ou pas de recommandation d'utilisation des percentiles en cas de distribution non gaussienne.

Instructions d'administration / cotation / interprétation	Non indiquées	Clair et précis sur seulement un type d'instructions[3]	Clair et précis sur 2 types d'instructions	Clair et précis sur les 3 types d'instructions
Qualifications de l'examinateur	Non indiquées	Terme générique et flou (ex. praticien, professionnel, clinicien)	Liste lacunaire de professionnels, sans lien avec l'objectif ou les conditions de passation du test[4]	Indication précise en lien avec l'objectif du test : orthophoniste, ou liste précise de professionnels (dont l'orthophoniste)
Erreur type de mesure	Non indiquée	Pas de tableau des erreurs de mesure (ETM) et pas de tableau des intervalles de confiance (IC) mais présence de données[5] permettant à l'utilisateur de calculer les ETM et les IC	Présence d'un tableau des ETM mais pas de tableau des IC[6]	Présence de tableaux des IC
Validité de contenu / analyse formelle des items	Non indiquée	Analyse uniquement qualitative[7] des items, quel que soit le nombre d'items.	Analyse qualitative et partiellement quantitative[8] des items. Nombre d'items par épreuve < 20.	Analyse qualitative et quantitative des items. Nombre d'items par épreuve ≥ 20.
Validité théorique ou de construit	Non indiquée	Indication de références ou de modèles théoriques, sans analyse statistique sur la validité du construit de l'épreuve.	Indication de références ou de modèles théoriques, avec une analyse statistique incomplète, ne portant que sur une partie des résultats ou moins de 100 sujets par sous-groupe.	Indication de références ou de modèles théoriques, avec une analyse statistique complète[9] et plus de 100 sujets par sous-groupe.
Validité concourante (seuil r = 0,40) **Validité prédictive (seuil : *r* = 0,30)**	Non indiquée	Indiquée mais incomplète (ex. pas de chiffres)	Indiquée mais < seuil, ou moins de 100 sujets par sous-groupe	Indiquée et ≥ seuil

[3] Nous considérons trois types d'instructions : l'administration, la cotation et l'interprétation.
[4] Ex. Le test peut être utilisé dans une école, mais les enseignants ne sont pas mentionnés comme potentiels utilisateurs.
[5] La connaissance du coefficient de fidélité test-retest et des écarts-types permettent de calculer les ETM et donc les IC.
[6] Les IC peuvent être calculés à partir des ETM selon la formule : IC = [SB – ETM seuil ; SB + ETM seuil]
[7] Nous appelons analyse « qualitative » des indications sur les critères de choix des items (ex. fréquence, complexité) ou des indications des % de réussite des items, sans analyse statistique plus approfondie.
[8] Nous appelons analyse « quantitative » l'analyse statistique effectuée. Ex. analyse classique des items avec un seul indice renseigné : indice de difficulté ou de discrimination.
[9] Ex. Analyse factorielle, analyse multitrait-multiméthode

3. Mesure des fidélités intra et inter-juges

La fidélité intra-juge a été évaluée sur les mesures psychométriques de 5 tests et batteries de langage écrit sélectionnés aléatoirement parmi les 26 étudiés dans notre mémoire. Ceci représente 19 % des tests. Il y a 92 % de mesures identiques entre la première et la deuxième cotation, trois semaines à un mois plus tard, selon les tests. De plus, une corrélation positive significative a été trouvée entre ces deux cotations ($r = 0,97$; $p < 0,0001$).

La fidélité inter-juge a été réalisée en utilisant les mesures déterminées par une autre examinatrice, étudiante en 5e année d'orthophonie (Herman) et nous-même, concernant 3 tests et batteries de langage oral et écrit sélectionnés aléatoirement parmi les 26 étudiés. Les deux examinatrices ont fourni 84 % de scores identiques pour la cotation des critères des tests de langage écrit. Une corrélation positive significative a été trouvée entre les deux cotations ($r = 0,94$; $p < 0,0001$).

L'ensemble de ces résultats est satisfaisant et tend à valider la méthodologie de cotation.

Résultats

La cotation des critères psychométriques pour chaque test ou batterie de langage écrit ainsi que deux exemples détaillés sont proposés en annexe concernant les tests Alouette-R et Timé-2 (cf. annexes A1, A2 et A3). Les différents résultats sont présentés selon trois analyses. Nous commencerons par examiner la qualité des informations psychométriques, puis les tests, et enfin les critères psychométriques.

1. Analyse de la qualité des informations psychométriques

L'objectif de cette analyse est de connaître le degré de précision des informations psychométriques indiquées par les auteurs des tests de langage écrit. Pour cela, nous avons réalisé le calcul suivant : le nombre total de scores 0 divisé par le nombre total de cotations de critères, tous tests et critères confondus, c'est-à-dire 390 cotations. Nous avons procédé de même pour chacune des scores 1, 2 et 3. Chacun de ces quatre ratios a été rapporté en pourcentage pour plus de lisibilité des résultats (figure 2).

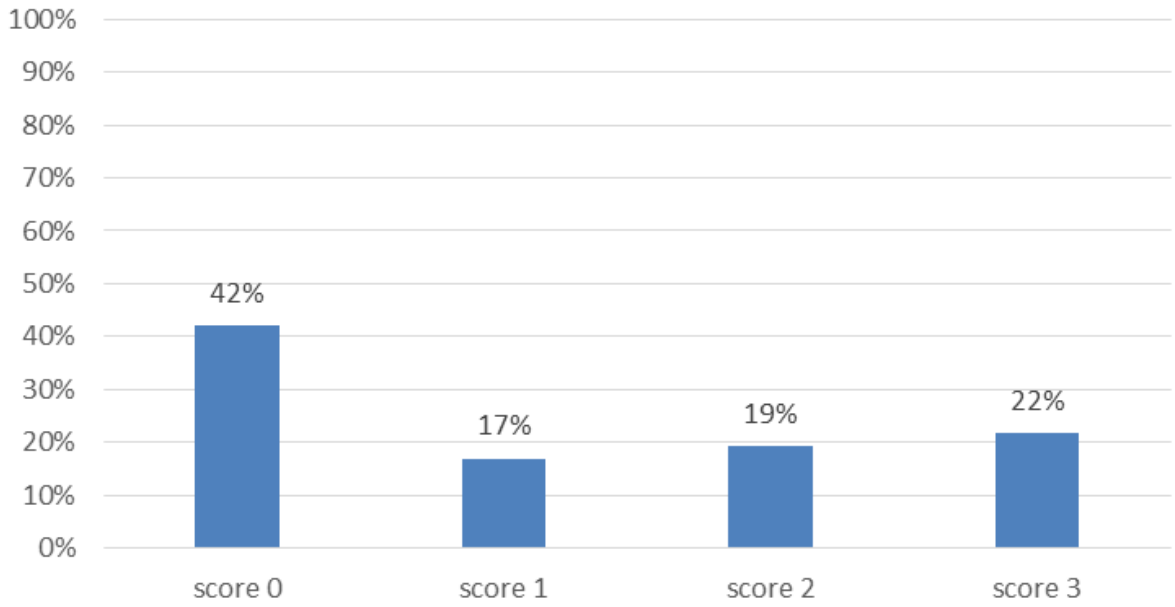

Figure 2 : Pourcentages des scores 0, 1, 2 et 3 pour l'ensemble des 15 critères et des 26 tests ou batteries de langage écrit.

Nous observons une majorité de critères cotés 0 (critère absent) et un faible nombre de critères cotés 3 (critère satisfaisant). Le pourcentage de critères non satisfaisants, c'est-à-dire dont les scores sont 0, 1 et 2, est de 78 %.

2. Analyse par test

Notre objectif dans cette partie est d'analyser la qualité psychométrique des tests et batteries de langage écrit. Pour cela, nous mesurons le nombre moyen de critères psychométriques par test ou batterie et le score total moyen des tests. Puis nous considérons le nombre de critères pour chaque test et batterie.

Afin d'analyser le nombre moyen de critères psychométriques par test, nous mesurons deux types de variables. La première concerne le nombre de critères mentionnés dans les manuels, elle correspond à considérer les critères partiellement ou totalement satisfaisants (score du critère > 0), sans tenir compte du degré de satisfaction des informations. La deuxième variable concerne le nombre de critères totalement satisfaisants (score du critère = 3) dans les manuels des tests ou batteries explorés.

Nous avons réalisé une analyse statistique descriptive pour ces deux variables ainsi que pour le score total des tests. Les moyennes, écarts-types et étendues des données sont présentées dans le tableau 5 et illustrées par des boîtes à moustaches (figures 3, 4 et 5).

Tableau 5 : Nombres moyens de critères psychométriques par test ou batterie et scores totaux moyens

	Nombre de critères cotés 1 ou + /3 (max : 15 critères)	**Nombre de critères cotés 3/3 (max : 15 critères)**	**Score total (max : 45 points)**
***M* (*ET*)**	8,69 (2,41)	3,27 (1,43)	18,12 (4,89)
Med	8	3	17,5
Etendue (min-max)	6 - 14	1 - 6	10 - 28

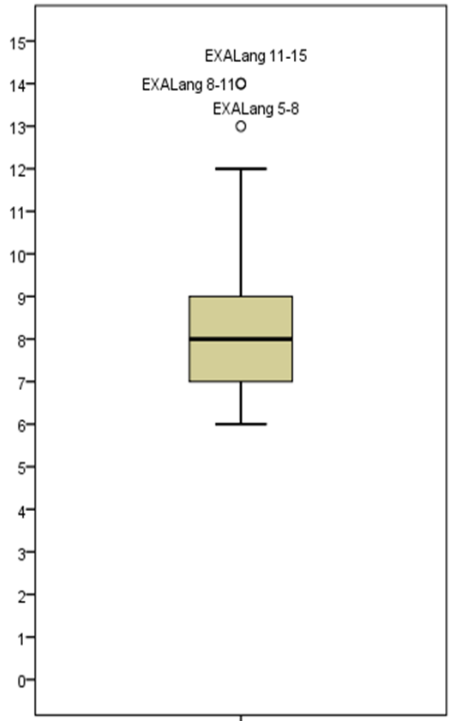

Figure 3 : Nombre de critères cotés 1 ou + /3 (max : 15 critères).

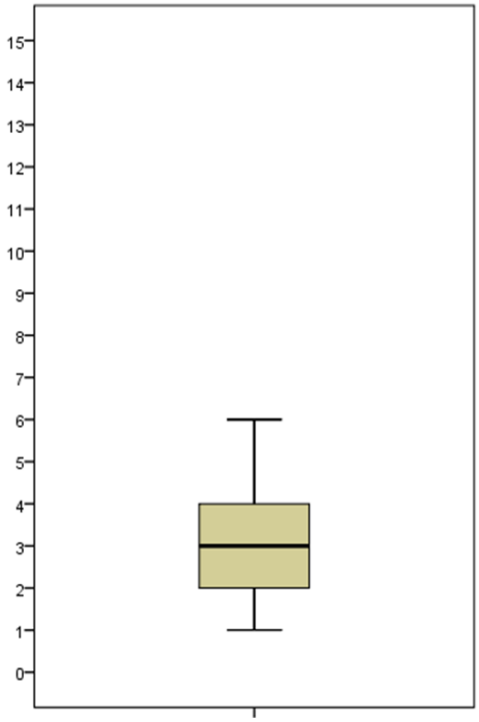
Figure 4 : Nombre de critères cotés 3 / 3 (max : 15 critères).

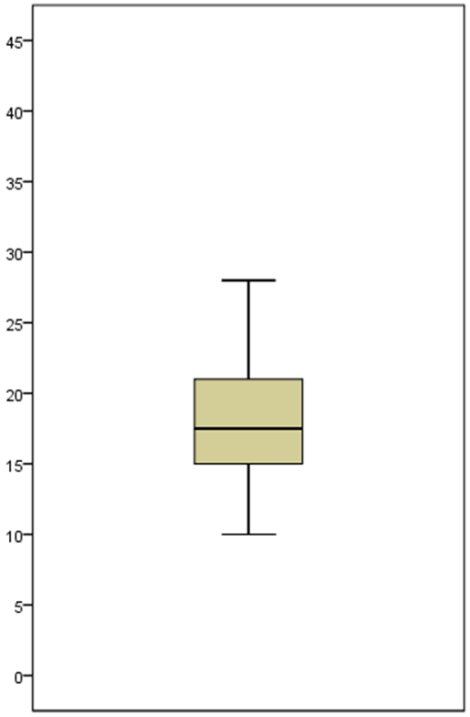
Figure 5 : Score total (max : 45 points).

Les résultats montrent que le nombre moyen de critères satisfaisants par test ou batterie de langage écrit est très faible. Notre analyse révèle également un faible score psychométrique total moyen des tests de langage écrit. Par ailleurs, les mesures indiquent un nombre moyen de critères satisfaisants par test très inférieur au nombre moyen de critères mentionnés. Cette différence correspond au manque de preuves apportées par les auteurs au sujet des critères psychométriques. Nous remarquons que les tests EXALang 5-8, 8-11 et 11-15 ont un nombre moyen de critères mentionnés très supérieur à celui des autres tests (figure 3). Cela signifie que leurs auteurs ont évoqué plus de critères dans le manuel que les auteurs des autres tests.

Pour rendre compte de la qualité psychométrique de chaque test ou batterie, nous voulons identifier les tests dont les critères psychométriques sont les mieux renseignés par les auteurs. Pour cela, nous mesurons le nombre de critères mentionnés et le nombre de critères satisfaisants par test ou batterie de langage écrit. Leur distribution est présentée dans la figure 6.

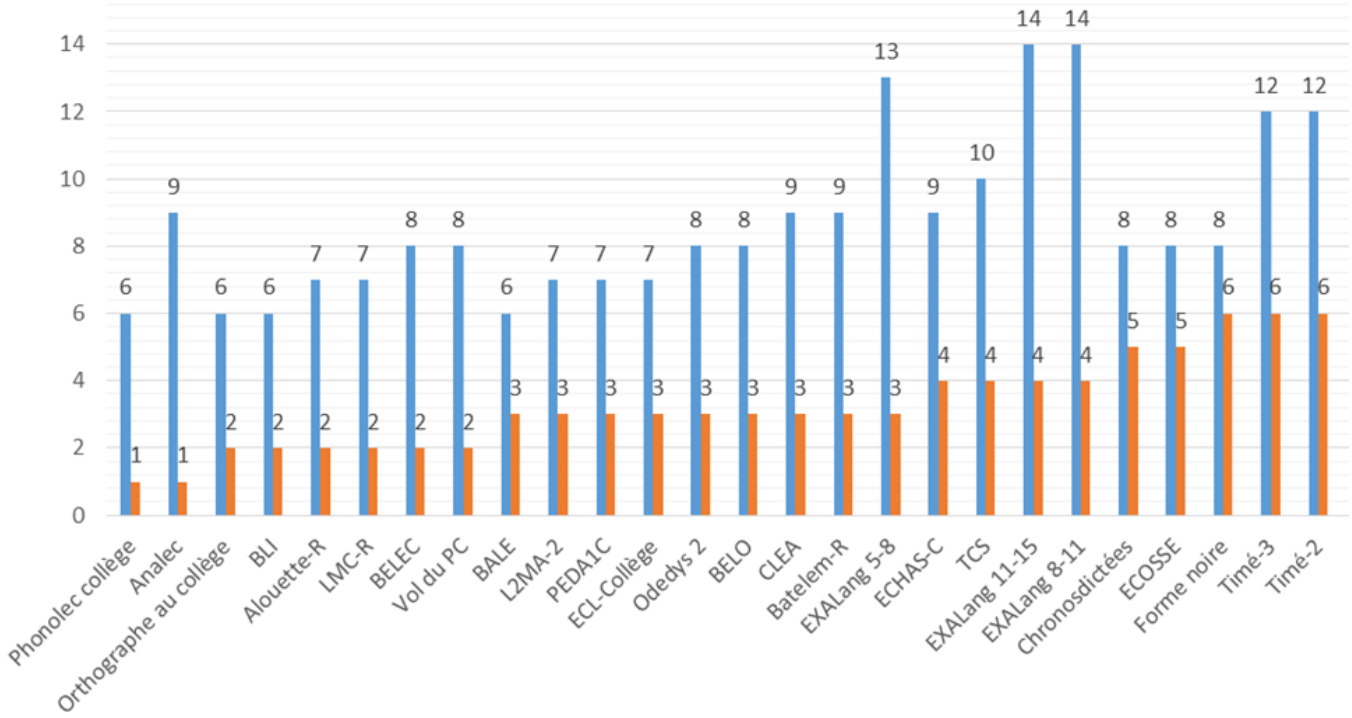

Figure 6 : Nombre de critères mentionnés et satisfaisants par test ou batterie.

Nombre de critères mentionnés (score > 0)
Nombre de critères satisfaisants (score = 3)

Le nombre de critères mentionnés varie de 6 à 14 sur 15, et 65,4 % des tests évoquent plus de la moitié des critères psychométriques. Seuls 6 critères sur 15 sont cités par les quatre tests les moins satisfaisants : la BALE, la BLI, l'Orthographe au collège et le Phonolec collège. Les deux outils les plus satisfaisants évoquent 14 critères sur 15 : les EXALang 8-11 et EXALang 11-15.

Le nombre de critères satisfaisants varie de 1 à 6 sur 15. Aucun test ou batterie ne possède la moitié des critères psychométriques. Les deux tests les moins satisfaisants ne présentent qu'un critère : le Phonolec collège et l'Analec. Les trois tests ou batteries les plus satisfaisants possèdent 6 critères, soit un taux de 40 % de critères satisfaisants : la Forme noire, le Timé-2, le Timé-3.

3. Analyse par critères

Notre objectif dans cette partie est d'évaluer la qualité de traitement des critères dans les tests et batteries de langage écrit en moyenne, puis pour chaque critère.

Dans ce but, nous avons déterminé le nombre moyen de tests et batteries de langage écrit qui mentionnent les critères psychométriques (score au critère > 0), puis le nombre moyen de tests et batteries qui satisfont ces critères (score au critère = 3). L'analyse statistique descriptive, avec les moyennes, écarts-types et étendues des données, est présentée dans le tableau 6. Une illustration est proposée par des boîtes à moustaches (figures 7 et 8).

Tableau 6 : Nombre de tests ou batteries par critère

	Nombre de tests ou batteries par critère coté 1 ou + /3 (max : 26 tests)	**Nombre de tests ou batteries par critère coté 3 (max : 26 tests)**
***M* (*ET*)**	15,07 (9,70)	5,67 (7,58)
Med	9	2
Etendue (min-max)	3 - 26	0 - 26

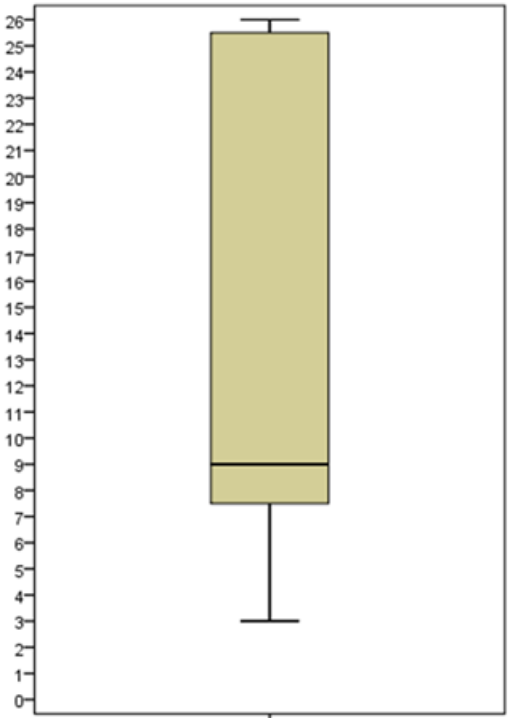

Figure 7 : Nombre de tests ou batteries par critère coté 1 ou + /3
(max : 26 tests ou batteries).

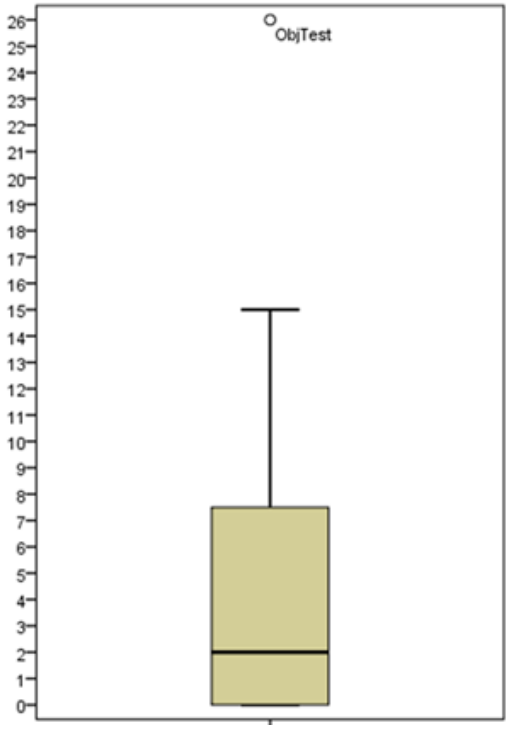

Figure 8 : Nombre de tests ou batteries par critère coté 3 / 3
(max : 26 tests ou batteries).

Nous remarquons qu'une majorité de tests ou batteries de langage écrit ne fait que mentionner les critères psychométriques. De plus, nos résultats montrent une grande variabilité de traitement des critères par les auteurs des tests et batteries.

Pour rendre compte de la qualité de traitement de chaque critère psychométrique dans les tests ou batteries de langage écrit, nous souhaitons identifier les critères les mieux renseignés par les auteurs. Pour cela, nous mesurons le nombre de tests et batteries pour chaque critère mentionné (score au critère > 0), puis pour chaque critère satisfaisant (score au critère = 3). Nos analyses sont illustrées dans la figure 9.

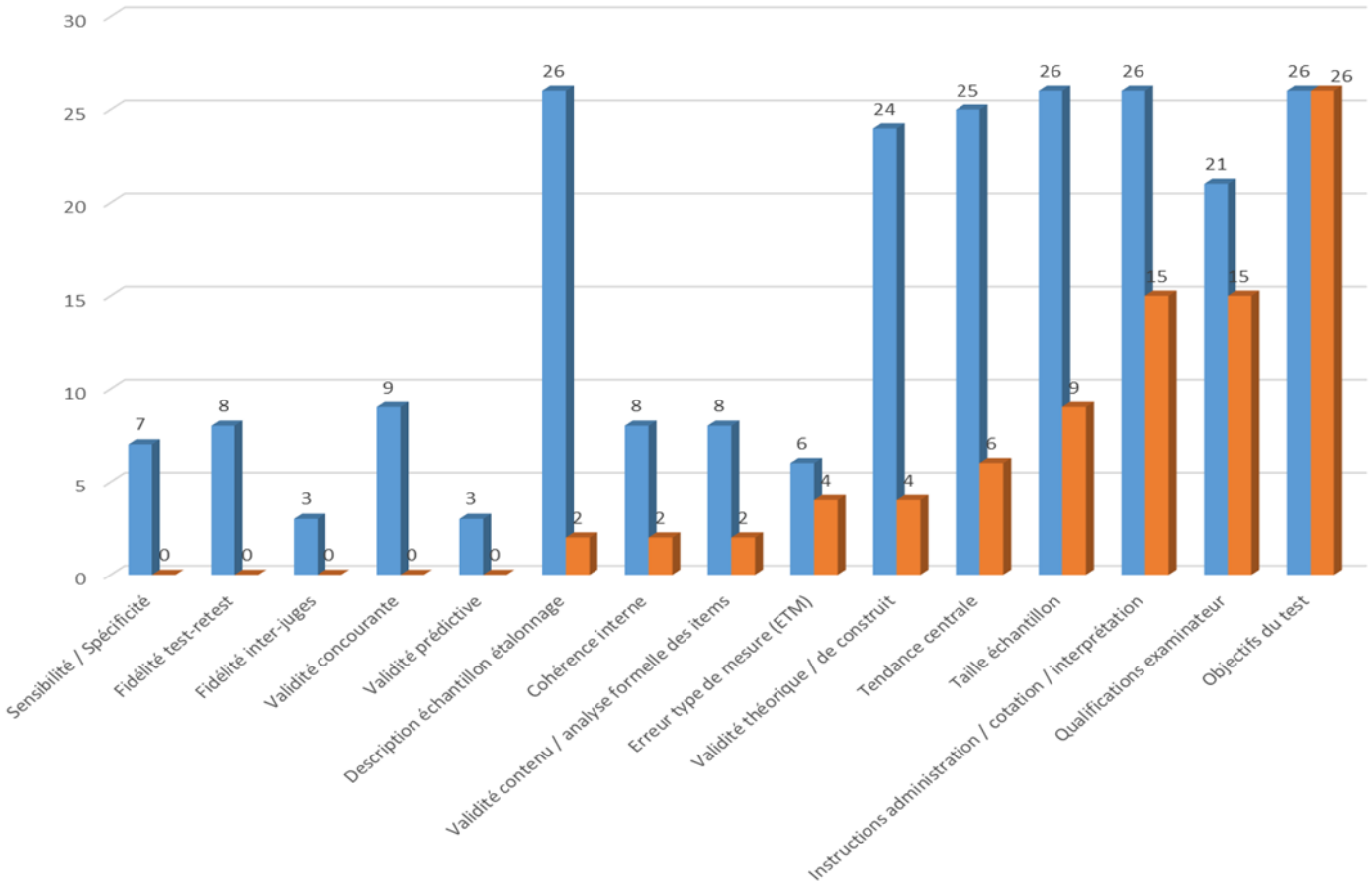

Figure 9 : Nombre de tests ou batteries par critère psychométrique

Nombre de tests ou batteries par critère coté 1 ou + /3 (max : 26)
Nombre de tests ou batteries par critère coté 3/3 (max : 26)

Les deux critères les moins présents sont la fidélité inter-juges et la validité prédictive. Ces deux critères ne sont mentionnés que dans trois tests : les EXALang 5-8, 8-11 et 11-15. En revanche, quatre critères sont toujours présentés dans les manuels : la description et la taille de l'échantillon d'étalonnage, les objectifs du test et les instructions d'administration, de cotation et d'interprétation du test.

Nous observons cinq critères jamais démontrés : la sensibilité et spécificité, la fidélité test-retest, la fidélité inter-juges, la validité concourante et la validité prédictive. Un seul critère est toujours satisfaisant : les objectifs du test.

Discussion

Dans cette partie, nous allons interpréter les résultats puis discuter des limites et implications de notre étude.

1. Discussion des résultats

L'objectif principal de ce travail était de réaliser l'inventaire des caractéristiques psychométriques des tests de langage écrit chez l'enfant, en Europe francophone. Notre objectif secondaire était de faire une étude exploratoire de la qualité des critères psychométriques dans ces tests.

1.1. Analyse de la qualité des informations psychométriques

Nos résultats concernant le degré de précision des informations montrent que 42 % des critères psychométriques ne sont même pas mentionnés par les auteurs des tests ou batteries de langage écrit. Le pourcentage de critères non satisfaisants est de 78. Ce résultat est élevé et suggère que notre hypothèse selon laquelle la qualité des informations présentes dans les manuels des tests ou batteries de langage écrit est insuffisante, est vérifiée.

Une explication pourrait être le parcours de formation de beaucoup d'auteurs de tests actuels ne permettant pas d'atteindre un niveau de connaissances suffisant en psychométrie pour mettre en place des tests aux qualités psychométriques avérées. Une explication complémentaire pourrait être le coût de mise en place de ces qualités.

1.2. Analyse par test

Les résultats de notre étude indiquent que le nombre moyen de critères satisfaisants par test ou batterie de langage écrit n'est que de 3,27 sur les 15 attendus (*ET* = 1,43). De plus, aucun test ou batterie ne satisfait la moitié des critères psychométriques étudiés. Cette valeur très faible montre que la qualité psychométrique globale des tests de langage écrit est insuffisante pour une pratique d'évaluation fondée sur les preuves.

Cependant, ces résultats généraux cachent des variations importantes entre les tests ou batteries. En effet, les tests Analec et Phonolec collège ne remplissent qu'un critère psychométrique alors que les tests Timé-2, Timé-3 et la Forme noire en satisfont six sur les quinze étudiés. Pour expliquer ces différences, nous proposons plusieurs explications. Pour le test Analec, la date de parution relativement ancienne (1998) pourrait éventuellement expliquer la moindre exigence de qualité psychométrique de son auteur. De plus, son objectif principal semble plus fondé sur une volonté d'approfondir la recherche et la prévention que sur une volonté diagnostique. En ce qui concerne le test Phonolec collège, nous supposons que leurs auteurs, orthophonistes, n'ont pu bénéficier d'une formation initiale leur permettant d'accéder aux connaissances psychométriques nécessaires contrairement aux auteurs des tests Timé-2, Timé-3 et la Forme noire.

Nos résultats indiquent également un nombre moyen de critères psychométriques satisfaisants par test ou batterie (M = 3,27 ; ET = 1,43) très inférieur au nombre moyen de critères simplement mentionnés (M = 8,69 ; ET = 2,41). Ces résultats laissent supposer que les informations contenues dans les manuels des tests ou batteries ne sont pas suffisamment pertinentes pour satisfaire les qualités psychométriques étudiées. Certains auteurs citent ces qualités sans en apporter la preuve (ex. les EXALang 5-8 ; 8-11 et 11-15).

L'ensemble de ces résultats est cohérent avec ceux de la plupart des revues psychométriques de la littérature (Bouchard & al., 2009 ; Flipsen & Ogiela, 2015 ; McCauley & Swisher, 1984a), bien que leurs auteurs évaluent des tests de langage oral et que les critères étudiés ne sont pas exactement identiques aux nôtres. Nous constatons une différence uniquement avec l'étude de Friberg (2010) dans laquelle chacun des neuf tests anglophones étudiés remplit au moins huit critères psychométriques sur onze. Nous supposons que cette différence est due au fait que les tests étudiés sont des tests diagnostiques donc d'une qualité psychométrique supérieure.

Notre hypothèse selon laquelle peu d'outils évaluant le langage écrit chez l'enfant répondent aux principaux standards psychométriques actuels semble donc vérifiée.

1.3. Analyse par critère

Concernant les critères psychométriques, les résultats indiquent un nombre moyen de tests ou batteries de langage écrit par critère psychométrique jugé satisfaisant de 5,67 (ET = 7,58) sur les 26 tests ou batteries étudiés. Ce score moyen est très faible. Il semble confirmer notre hypothèse d'une qualité psychométrique faible des tests de langage écrit.

De plus, cinq critères ne sont jamais jugés satisfaisants : la sensibilité et spécificité, la fidélité test-retest, la fidélité inter-juges, la validité concourante et la validité prédictive. Un seul critère est toujours rempli : les objectifs du test. Ces résultats sont en accord avec ceux des revues psychométriques de la littérature, notamment celle de McCauley et Swisher (1984a). En effet, elles indiquent que les critères les plus fréquemment absents sont la validité et la fidélité.

En ce qui concerne l'absence du critère de sensibilité et spécificité, notre résultat est à nuancer. Il ne signifie pas que les tests ou batteries étudiés ont une mauvaise sensibilité ou spécificité mais que les informations concernant ces indices sont absentes des manuels. Par exemple, le manuel du test l'Alouette-R (Lefavrais, 2005) ne contient aucune information sur la sensibilité ou la spécificité. Néanmoins, dans un article récent, ces indices de sensibilité et de spécificité ont été déterminés respectivement à 83,1 % et 100 %, dans une population dyslexique universitaire (Cavalli, Colé, Leloup, Poracchia-George, Sprenger-Charolles & El Ahmadi, 2017). Ces taux sont supérieurs au seuil de 0,80 recommandé par Bernaud (2007), ce qui montre qu'un test ne renseignant pas une qualité psychométrique dans son manuel n'en est pas forcément exempt.

Là encore, une explication possible de l'absence ou la présence d'informations au sujet d'un critère psychométrique est la qualité de la formation en psychométrie du parcours professionnel de l'auteur. Ceci peut entraîner un manque de connaissances sur l'importance de mesurer les qualités psychométriques d'un test comme la fidélité ou la validité (ex. le Phonolec collège). En effet, en prenant l'exemple de l'orthophonie, sa formation est en évolution et a été réformée en 2013, accordant aux nouveaux diplômés le grade master. Cette réforme a induit un changement des programmes d'enseignement, notamment sur le plan de la psychométrie.

Une autre explication possible est le coût que la mise en place de certains critères peut amener. En effet, un critère psychométrique comme les objectifs est peu onéreux, alors que l'organisation d'une double passation, nécessaire à la mesure de la fidélité test-retest, pour plus de cent sujets par sous-groupe, est plus coûteuse en temps et en argent.

2. Limites de l'étude

Notre étude comporte certaines limites qu'il convient de prendre en compte dans l'interprétation des résultats. La première limite concerne le choix des tests. En effet, il aurait été intéressant de choisir les tests étudiés sur le critère de leur fréquence d'utilisation par les orthophonistes, pour être au plus près des pratiques du terrain. Notre expérience clinique nous suggère que certains tests de notre liste ne sont que peu usités (ex. ECHAS-C, BLI). A notre connaissance, aucune recherche récente ne recense cette fréquence d'utilisation. De plus, le choix d'une liste belge (INAMI, 2018) ne reflète peut-être pas l'usage des tests de langage écrit par les orthophonistes de France.

La deuxième limite concerne la sélection des critères psychométriques. Il aurait été intéressant de prendre en compte d'autres qualités des tests. Par exemple, concernant la date de parution, nous savons que les normes ont une durée de vie limitée (Grégoire, 2011). En effet, la récence de l'étalonnage du test est un critère qui semble important à prendre en compte. De plus, pour être plus complète, notre étude aurait peut-être pu considérer d'autres qualités intéressantes, bien que non psychométriques, par exemple l'aspect ludique, la praticité des conditions d'administration, la durée de passation.

Pour finir, la cotation de critères psychométriques peut sembler objective et aisée. Mais l'évaluation qualitative de certains critères demande souvent de déterminer une limite arbitraire, lorsque les informations sont parcellaires ou manquent de clarté. Par exemple, il est parfois difficile de comprendre l'objectif d'une corrélation qui peut être utilisée pour prouver différents critères psychométriques du test ou de la batterie (ex. la validité théorique et la validité de contenu). Les auteurs ne précisent pas toujours le but de leurs analyses. De plus, la cotation d'une batterie est moins aisée que celle d'un test « simple ». Par exemple, sur certains critères psychométriques comme la validité de contenu, tous les subtests ne sont pas forcément étudiés. Par ailleurs, la cotation de la validité concourante, qui suppose une corrélation avec un test considéré comme un gold standard n'est pas évidente. En effet, la plupart du temps, il n'existe pas de gold standard et les corrélations sont réalisées avec d'autres tests (ex. dans le Timé-2, corrélation avec « la pipe et le rat »). Enfin, notre tableau de cotation est sans doute perfectible

étant donné les choix arbitraires que nous avons dû faire pour délimiter les différents niveaux de notre échelle de cotation. Par exemple, le nombre de sujets inférieur à 50 pour coter 1 le critère de taille de l'échantillon.

3. Implications pratiques

Notre étude porte sur les tests de langage écrit chez l'enfant. Un travail parallèle sur le langage oral est effectué par Herman (2018). Nos deux études combinées constituent un point de départ pour élargir le champ d'investigation des caractéristiques psychométriques à d'autres domaines pour établir un état des lieux de la psychométrie dans les tests en orthophonie.

Notre étude pourrait également constituer un point de départ pour évaluer individuellement chaque test, et établir une liste d'améliorations à apporter pour obtenir des qualités psychométriques satisfaisantes. Les auteurs auraient ainsi une base sur laquelle se fonder pour améliorer les tests existant.

Dans sa démarche diagnostique, l'orthophoniste prend la responsabilité du choix de ses outils (Grégoire, 2011 ; Perdrix, 2018). Il est supposé en connaître les qualités, les propriétés et les limites et les utiliser dans un but précis en minimisant le risque de se tromper. Il utilise aussi ses compétences cliniques pour interpréter les scores et poser son diagnostic puis les expliquer au patient. Or, Betz et ses collègues (2013) ont montré que les orthophonistes ne fondaient pas leur choix de tests sur leurs qualités psychométriques, mais plutôt sur leur date de parution. De plus, Kerr et al. (2003) soulignent que les orthophonistes n'ont pas confiance en leurs connaissances psychométriques et commettent des erreurs d'utilisation des tests. Néanmoins, depuis la réforme des études en orthophonie en France évoquée précédemment, la formation intègre un enseignement en psychométrie. Notre étude est donc susceptible de répondre à ce besoin croissant d'information sur les qualités psychométriques et sur la fiabilité des outils à notre disposition. Elle permet également d'améliorer la transparence de nos outils pour les professionnels avec lesquels nous sommes amenés à communiquer en tant que praticiens ou créateurs d'outils.

Conclusion

Notre principal objectif dans cette étude était de réaliser l'inventaire des caractéristiques psychométriques des tests francophones de langage écrit chez l'enfant. Notre objectif secondaire était de faire une étude exploratoire de la qualité de traitement des critères psychométriques dans les tests étudiés. Pour atteindre ces objectifs, nous avons sélectionné 26 tests ou batteries sur la base de deux listes répertoriant les tests ou batteries actuellement utilisés par les cliniciens et les chercheurs francophones d'Europe (INAMI, 2018 ; UNADREO, 2011). Nous avons sélectionné quinze critères psychométriques à analyser en nous fondant sur les revues psychométriques de la littérature. Nous avons établi une échelle de cotation en quatre niveaux la plus fiable et précise possible. Enfin, nous avons réalisé la cotation des 15 critères psychométriques des 26 tests ou batterie de langage écrit.

Les principaux résultats de notre étude suggèrent une qualité psychométrique insatisfaisante des tests ou batteries de langage écrit. Ces résultats vont dans le sens de ceux des principales revues psychométriques de la littérature (Bouchard & al., 2009 ; Flipsen & Ogiela, 2015 ; Friberg, 2010 ; McCauley & Swisher, 1984a). Un autre point important soulevé dans notre étude est que les informations contenues dans les manuels des tests ou batteries sont souvent incomplètes voire absentes. Il est cependant important de nuancer cette observation. Elle signifie que les auteurs n'apportent pas de preuves de validité, de fidélité ou de standardisation correcte. Il est possible que ces tests possèdent les qualités psychométriques attendues.

Dans la pratique orthophonique, notamment la démarche diagnostique, l'utilisation d'outils standardisés est fréquente. Il est donc important de connaître les qualités et les limites de ces outils, pour les choisir et les employer de la manière la plus appropriée possible, le premier outil de l'orthophoniste restant son sens clinique, guidé par son savoir, son savoir-faire et son savoir-être.

Pour aller plus loin, il serait intéressant d'élargir ce type d'étude à d'autres domaines de l'orthophonie et ainsi établir un état des lieux psychométrique des tests orthophoniques francophones. Il semble également nécessaire de développer des outils plus fiables sur le plan psychométrique afin de poursuivre le développement de la pratique fondée sur les preuves. Le but de cette démarche est de répondre aux attentes des patients et celles des professionnels, qu'on espère mieux formés à la psychométrie. Ce travail pourrait constituer une base de données utile à l'amélioration des tests orthophoniques existants d'une part, et à la création de nouveaux tests ou batteries d'évaluation d'autre part.

Bibliographie

Anceaux, F. & Sockeel, P. (2006). Mise en place d'une méthodologie expérimentale : hypothèses et variables. *Recherche en Soins Infirmiers, 84*, 66-83.

Baneath, B., Boutard, C., & Alberti, C. (2006). *Chronosdictées. Outils d'évaluation des performances orthographiques avec et sans contrainte temporelle, du CE1 à la 3e*. Isbergues : OrthoÉdition.

Bartram, D. (1994). Mesurer les différences entre les personnes. Fidélité et validité. In J.R. Beech et L. Harding (traduit de l'anglais sous la direction de J.P. Rolland et J.L. Mogenet). *Tests, mode d'emploi... Guide de psychométrie*. Paris : ECPA, 65-100.

Bernaud, J.-L. (2007). *Introduction à la psychométrie*. Paris : Dunod.

Bertrand, D., Fluss, J., Billard, C. & Ziegler, J.C. (2010). Efficacité, sensibilité, spécificité : comparaison de différents tests de lecture. *L'Année Psychologique, 110*, 299-320.

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.

Bouchard, M-E.G., 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*, 129-139.

Boutard, C., Claire, I., & Gretchanovsky, L. (2006). *Le vol du P.C. Évaluation fonctionnelle de la lecture chez les sujets de 11 à 18 ans*. Isbergues : OrthoÉdition.

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

Cavalli, E., Colé, P., Leloup, G., Poracchia-George, F., Sprenger-Charolles, L. & El Ahmadi, A. (2017). Screening for Dyslexia in French-Speaking University Students : An Evaluation of the Detection Accuracy of the Alouette Test. *Journal of Learning Disabilities*, *51*, 268-282.

Chevrie-Muller, C., Maillart, C., Simon, A.-M., & Fournier, S. (2010). *L2MA-2. Langage oral, Langage écrit, Mémoire, Attention. 2ème édition*. Paris : ECPA

Coquet, F. (2013). *Troubles du langage oral chez l'enfant et l'adolescent : Pistes pour l'évaluation*. Isbergues : OrthoEdition.

Durieux, N., Palseau, F. & Maillart, C. (2012). Sensibilisation à l'Evidence-Based Practice en logopédie. *Les Cahiers de l'ASELF, 9*, 7-15.

Ecalle, J. (2003). *Timé-2. Test d'identification de mots écrits de 6 à 8 ans*. Paris: ECPA. N'est plus édité, est disponible auprès de l'auteur.

Ecalle, J. (2006). *Timé-3. Test d'identification de mots écrits pour enfants de 7 à 15 ans*. Paris : Mot à Mot.

Ecalle, J. & Magnan, A. (2006). Des difficultés en lecture à la dyslexie : problème d'évaluation et de diagnostic. *Glossa, 97*, 4-19.

Fletcher, J.M., Coulter, W.A., Reschly, D.J. & Vaughn, S. (2004). Alternative Approaches to the Definition and Identification of Learning Disabilities : Some Questions and Answers. *Annals of Dyslexia, 54(2)*, 304-331.

Flipsen, P.J. & Ogiela, D.A. (2015). Psychometric Characteristics of Single-Word Tests of Children's Speech Sound Production. *Language, Speech, and Hearing Services in Schools, 46*, 166–178.

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.

George, F., & Pech-Georgel, C. (2008). *BELO. Batterie d'Evaluation de la Lecture et de l'Orthographe*. Marseille : De Boeck Solal.

Gregoire, J. (2006). *L'examen clinique de l'intelligence de l'enfant : Fondements et pratique du WISC-IV.* Sprimont, Belgique : Mardaga.

Gregoire, J. (2011). La psychométrie est-elle compatible avec l'éthique ? *Rééducation Orthophonique, 247*, 33-43.

Grégoire, J. (2014). L'examen diagnostique est-il normatif ? *A.N.A.E., 132-133,* 459-465.

Haynes, S.N., Richard, D.C.S. & Kubany,E.S. (1995). Content validity in psychological assessment : A functional approach to concepts and methods. *Psychological Assessment, 7(3)*, 238-247.

Herman, F. (2018). *Caractéristiques psychométriques des tests de langage oral chez l'enfant* (Master's thesis). Université de Lille, Lille.

Hogan, T.P. (2012). *Introduction à la psychométrie*. Traduction et adaptation française par R. Stephenson & N. Parent, Montréal, Canada : Chenelière Education.

Hutchinson, T.A. (1996). What to look for in the technical manual : twenty questions for users. Language, *Speech and Hearing Services in Schools, 27*, 109-121.

INAMI. (2018). Liste limitative des tests de logopédie. Lecture. Consulté à l'adresse http://www.riziv.fgov.be/SiteCollectionDocuments/liste-logopede-tests-dyslexie-2018.pdf

INAMI. (2018). Liste limitative des tests en logopédie. Orthographe. Consulté à l'adresse http://www.riziv.fgov.be/SiteCollectionDocuments/liste-logopede-tests-dysorthographie-2018.pdf

Inizan, A. (1998). *ANALEC. Analyse du savoir lire de 8 ans à l'âge adulte*. Paris : ECPA.

Institut national de la santé et de la recherche médicale. (2007). Dyslexie Dysorthographie Dyscalculie : Bilan des données scientifiques. Paris, France : Les éditions Inserm.

Jacquier-Roux, M., Lequette, C., Pouget, G., Valdois, S., & Zorman, M. (2010). *BALE. Batterie Analytique du Langage Ecrit*. Grenoble : Cogni-sciences. UPMF de Grenoble.

Jacquier-Roux, M., Valdois, S. & Zorman, M. (2002). *ODEDYS. Outil de dépistage des dyslexies. Grenoble* : Cogni-sciences. UPMF de Grenoble.

Kerr, M.A., Guildford, S. & Bird E.K-R. (2003). Standardized Language Test Use : A Canadian Survey. *Journal of Speech-Language Pathology and Audiology, 27(1)*, 10-28.

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

Khomsi, A. (2003). *BLI. Bilan de Lecture Informatisé*. Paris : ECPA.

Khomsi, A., Nanty, I., Pasquet, F. & Parbeau-Guéno, A. (2005). *ECL-Collège. Evaluation des Compétences Linguistiques au Collège*. Paris : ECPA.

Kirk, C. & Vigeland, L. (2014). A Psychometric Review of Norm-Referenced Tests Used to Assess Phonological Error Patterns. *Language, Speech, and Hearing Services in Schools, 45*, 365-377.

Kline, P. (1994). Choisir le meilleur test. In J.R. Beech et L. Harding (traduit de l'anglais sous la direction de J.P. Rolland et J.L. Mogenet). *Tests, mode d'emploi... Guide de psychométrie*. Paris : ECPA, 101-137.

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

Lecocq, P. (1996). *L'E.CO.S.SE. Une Épreuve de COmpréhension Syntaxico-SEmantique*. Villeneuve d'Ascq : Presses Universitaires du Septentrion.

Lefebvre, P. & Trudeau, N. (2005). L'orthophoniste et les tests normalisés. *Frequences, 17(2)*, 17-20.

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

Lenfant, M., Thibault, M.-P., & Helloin, M.-C. (2009). *EXALang 11-15. Batterie informatisée du langage oral, langage écrit, compétences transversales. Collégiens - Adolescents*. Grenade (France) : Orthomotus.

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

Maeder, C. (2010). *La forme noire*. Isbergues : OrthoÉdition.

McCauley, R.J. & Swisher, L. (1984b). Use and misuse of norm-referenced tests in clinical assessment: a hypothetical case. *Journal of Speech and Hearing Disorders, 49*, 338-348.

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

Mousty, P., Alegria, J., Leybaert, J., Morais, J., & Content, A. (1995). *BELEC. Batterie d'évaluation du langage écrit*. Bruxelles : Université Libre de Bruxelles, Laboratoire Cognition, Langage, Développement.

Ministère de l'emploi et de la solidarité. Décret n°2002-721 du 2 mai 2002 relatif aux actes professionnels et à l'exercice de la profession d'orthophoniste abrogé par décret 2004-802 2004-07-29 article 5 [en ligne]. Journal Officiel de la République Française, n°183 du 8 août 2004. Consulté à l'adresse https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000413069

Mousty, P., & Leybaert, J. (1999). Évaluation des habiletés de lecture et d'orthographe au moyen de BELEC : données longitudinales auprès d'enfants francophones testés en 2e et 4e années. *Revue Européenne de Psychologie Appliquée*, *49*(4), 325-342.

Pasquet, F., Parbeau-Guéno, A., & Bourg, E. (2014). *CLéA. Communiquer, Lire et Écrire pour Apprendre*. Paris : ECPA.

Perdrix, R. (2018). Pour une contribution éclairée et raisonnée de l'évaluation standardisée et normalisée aux diagnostics de troubles développementaux du langage. Quelques éléments pour choisir, comprendre et exploiter les tests en orthophonie. *Rééducation Orthophonique, 273*, 47-70.

Plaza, M., Oudry, M., Gatignol, P., & Robert, A.-M. (2011). *Phonolec Collège*. Gisors : Adeprio.

Rondal, J.A. (1997). *L'évaluation du langage*. Bruxelles : Mardaga.

Savigny, M., Barbier, C., Coupey Le Roy, R., Girard, J., & Roussel, G. (2001). *Batelem Révisée. Batterie d'épreuves pour l'école élémentaire. Cycle II et première année du cycle III*. Paris : ECPA

Shipley, K.G. & McAfee, J. (2009). *Assessment in Speech-Language Pathology: A Resource Manual Fourth Edition.* Clifton Park, United States of America : Delmar Cengage Learning.

Simonart, G. (1998a). *ECHAS-C : Echelle d'apprentissages scolaires primaires*. Braine-le-Château (Belgique) : Application des techniques modernes.

Simonart, G. (1998b). *Peda 1C : Tests pédagogiques de premier cycle primaire*. Braine-le-Château (Belgique) : Application des techniques modernes.

Smith, G.T. (2005). On construct validity : Issues of method and measurement. *Psychological Assessment, 17(4)*, 396-408.

Thibault, M.-P. (2008). *Orthographe au collège*. Grenade (France) : Orthomotus.

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

UNADREO (2011) Liste des tests orthophoniques utilisés pour l'évaluation du langage oral et langage écrit. Consulté à l'adresse http://www.unadreo.org/assets/medias/fichiers/2014-11-04-16-46-39_9422956.pdf

Vrignaud, P., Castro, D., & Mogenet, J.-L. (2003). Recommandations internationales sur l'utilisation des tests : version 2000. *Pratiques Psychologiques* [Numéro spécial hors-série].

Liste des annexes