Département d'Orthophonie
Gabriel DECROIX

MEMOIRE

En vue de l'obtention du

Certificat de Capacité d'Orthophoniste

présenté par :

Florence HERMAN

soutenu publiquement en juin 2018 :

Caractéristiques psychométriques des tests de langage oral chez l'enfant

MEMOIRE dirigé par :

Perrine DANCHIN, orthophoniste, Hôpital Saint Vincent de Paul,

et enseignante vacataire, département d'Orthophonie, Lille

Lucie MACCHI, maître de conférences au département d'Orthophonie,

STL, Université de Lille, Lille

Lille – **2018**

Remerciements

Je remercie en premier lieu Mesdames Danchin et Macchi pour leur implication, leur aide, leurs conseils et leur accompagnement stimulant tout au long de ce travail de recherche.

J'adresse mes chaleureux et amicaux remerciements à ma camarade Laura Colli-Vaast pour la richesse et la bienveillance de nos échanges, qui ont énormément contribué à l'approfondissement de ma réflexion.

Je tiens à remercier mes maîtres de stage, qui m'ont permis de découvrir la fabuleuse diversité de leur métier.

Je remercie particulièrement Yohan pour ses conseils techniques, son aide, ses relectures, ainsi que pour son écoute et son soutien renouvelés depuis cinq ans.

Enfin, un grand merci à ma famille et à mes amis, qui ont manifesté leur enthousiasme dès le début de ce projet de reconversion professionnelle et n'ont pas cessé depuis de me témoigner leurs encouragements.

Résumé :

L'évaluation du langage fait partie intégrante du travail de l'orthophoniste. Cette recherche s'intéresse aux outils francophones d'évaluation du langage oral de l'enfant. Elle a pour objectif de décrire et d'analyser les qualités psychométriques de 29 tests ou batteries de langage oral. Elle s'appuie sur quinze critères de validité, de fidélité, de sensibilité et de standardisation. L'état de la recherche soulignait des lacunes en ce domaine concernant les tests américains et canadiens. Notre hypothèse était que peu de tests francophones disponibles en France et évaluant le langage oral chez l'enfant répondaient aux standards psychométriques requis. Les résultats de notre étude confirment cette hypothèse : la plupart des tests n'apportent pas de niveau de preuves suffisant de validité, de fidélité, de sensibilité et de standardisation. Le manque de solidité psychométrique des outils d'évaluation orthophoniques nous conduit à interroger nos pratiques et notre utilisation clinique des outils d'évaluation disponibles. Il ouvre un champ de recherche conséquent afin d'améliorer ces outils.

Mots-clés :

langage oral – test – enfants – évaluation – psychométrie

Abstract :

Language assessment is an essential part of the speech therapist's work. This research focuses on french-language tools for assessing children's spoken language. It aims to describe and analyse psychometric qualities of 29 spoken language tests or batteries. It is based on fifteen criteria of validity, reliability, sensitivity and standardisation. Previous researchs pointed psychometric shortcomings concerning American and Canadian tests. We hypothesized that few French-language tests available in France for assessing children's spoken language satisfy psychometric norms. The results of our study confirm this hypothesis: most of tests do not give sufficient evidence of validity, reliability, sensitivity and standardisation. The lack of psychometric solidity of language assessment tools leads us to question our practices and our clinical use of available assessment tools. It opens a large research area to improve these tools.

Keywords :

spoken language – test – children – assessment – psychometrics

Table des matières

Introduction....1
Contexte théorique, buts et hypothèses....2
1.Définition et usage d'un test dans une démarche d'évaluation....2
1.1.Définition d'un test....2
1.2.Utilisation des tests en orthophonie et recommandations....2
2.Psychométrie....3
2.1.Validité....4
2.1.1.Validité de contenu....4
2.1.2.Validité théorique et validité de construit/de construction....4
2.1.3.Validité concourante....5
2.1.4.Validité prédictive....5
2.2.Fidélité....5
2.2.1.Fidélité test-retest....6
2.2.2.Fidélité inter-juges....6
2.2.3.Cohérence interne / homogénéité....6
2.2.4.Critères en lien avec la fidélité....6
2.3.Normes et standardisation....7
2.3.1.Description de l'échantillon d'étalonnage....7
2.3.2.Taille de l'échantillon....7
2.3.3.Mesure de la tendance centrale....8
2.4.Sensibilité et spécificité....8
3.État des lieux de la recherche sur les qualités psychométriques des tests de langage oral. 9
4.Objectif et hypothèses....10
Méthode....11
1.Choix des tests....11
2.Choix des critères psychométriques....12
3.Méthode de cotation pour chaque critère....13
4.Fiabilité de la méthode de cotation....15
Résultats....16
1.Analyse globale des résultats de cotation....16
2. Analyse des tests....17
2.1. Analyse globale....17
2.2.Analyse par test....18
3.Analyse des critères....20
Discussion....23
1.Discussion des résultats....23
1.1.Discussion sur les tests....23
1.2.Discussion sur les critères psychométriques....24
1.3.Mises en perspective avec les résultats d'autres recherches....26
2. Implications pratiques....26
3.Limites de l'étude....27
4.Pistes de recherches et perspectives....28
Conclusion....29
Bibliographie....31
Annexes....35
Annexe n°1 : Liste des tests ayant fait l'objet de mesures de fiabilités intra et inter-examinatrices. 35
Annexe n°2 : Tableau 4 : Cotation des critères psychométriques par test. 36

Annexe n°3 : Etude de l'ELO (Khomsi, 2001).....37
Annexe n°4 : Etude de l'EVIP (Dunn et al., 1993). 38

Introduction

Ainsi que le précise le décret de compétences relatif aux actes professionnels de l'orthophoniste (2002), tout acte de soin nécessite une évaluation préalable des compétences et des difficultés du patient. Celle-ci doit être fiable et précise, afin d'identifier l'éventuelle présence d'un trouble, d'en préciser la nature, les caractéristiques et la sévérité, puis de décider d'une potentielle prise en charge, la plus adaptée possible au patient. Pour ce faire, le professionnel dispose d'un choix conséquent de tests, notamment pour évaluer le langage oral chez l'enfant. Dans cet objectif d'évaluation, le choix des épreuves utilisées s'avère primordial. L'utilisateur d'un test doit connaître la façon dont le test a été conçu et les méthodes psychométriques qui ont permis l'analyse des résultats, afin d'interpréter correctement les scores des épreuves et d'établir un diagnostic fiable. Cette démarche est celle d'une pratique fondée sur les preuves (*Evidence-Based Practice* en anglais) : celle-ci permet au clinicien d'orienter sa pratique en se basant notamment sur les résultats de la recherche scientifique et sur son expertise clinique.

Plusieurs auteurs se sont intéressés aux outils d'évaluation du langage chez l'enfant, en particulier aux États-Unis d'Amérique et au Canada. Trente tests américains de langage et d'articulation pour les enfants d'âge préscolaire ont été analysés en s'appuyant sur dix critères psychométriques : seulement 20 % des tests remplissaient la moitié des critères et les informations nécessaires pour évaluer les qualités du test étaient pour la plupart absentes des manuels (McCauley & Swisher, 1984). A propos des qualités psychométriques des outils francophones, peu de travaux ont été réalisés (Bouchard, Fitzpatrick & Olds, 2009). Seuls quelques outils d'évaluation du langage chez l'enfant utilisés par les professionnels francophones canadiens ou belges ont été étudiés : ils offrent peu de garanties concernant leurs qualités psychométriques (Bouchard et al, 2009 ; Leclercq & Veys, 2014). Les études ou les synthèses d'informations disponibles sur les qualités psychométriques de beaucoup de tests utilisés en France sont donc rares ou incomplètes.

L'objectif de ce présent travail est de décrire les caractéristiques psychométriques des tests francophones de langage oral chez l'enfant. Un travail similaire concernant les tests de langage écrit est réalisé par Colli-Vaast (2018). Dans un premier temps, nous définirons ce qu'est un test et l'usage qui en est fait dans une démarche d'évaluation, avant de définir la psychométrie et les différents critères sur lesquels appuyer notre analyse psychométrique. Nous établirons par la suite un état des lieux de la recherche concernant la description des caractéristiques psychométriques des tests orthophoniques. Suite à cette revue de littérature, nous présenterons notre hypothèse : peu de tests évaluant le langage oral chez l'enfant, disponibles et utilisés en France, répondent aux standards psychométriques attendus. Nous expliquerons ensuite notre méthodologie (choix des tests et batteries d'épreuves étudiés dans ce mémoire ; choix des critères psychométriques retenus ; méthode de cotation de chaque critère), avant de présenter nos résultats et les éléments de la discussion.

Contexte théorique, buts et hypothèses

Nous présenterons en premier lieu la définition d'un test et son utilisation par les professionnels de soin. Puis nous définirons les critères de qualité psychométrique d'un test et synthétiserons la revue de la littérature traitant de notre sujet. Enfin, nous établirons les hypothèses relatives à notre étude.

1. Définition et usage d'un test dans une démarche d'évaluation

Nous aborderons dans cette partie ce que recouvre la notion de test, ainsi que la manière dont il est utilisé par les cliniciens.

1.1. Définition d'un test

Un test est un outil ou une méthode standardisés, servant de stimulus à un comportement, observé chez un individu par un examinateur. Il existe plusieurs catégories de tests, notamment les tests d'aptitudes intellectuelles, les tests de rendement, les tests de personnalité et les mesures des intérêts et des attitudes (Hogan, 2012). Les observations de l'examinateur se résument généralement sous forme de scores chiffrés.

L'objectif d'un test est de classer un individu par rapport à un groupe de référence, en comparant statistiquement ses résultats à ceux d'autres individus placés dans la même situation (Pichot, 1999). Un test doit présenter quatre propriétés (Grégoire, 2014 ; Huteau & Lautrey, 1997 ; Rondal, 1997) : standardisation ; présence de normes pour situer le sujet par rapport au groupe de référence ; fidélité ; validité.

Nous reviendrons sur la définition précise de ces critères dans la suite de cette partie (paragraphe 2). Ces critères font partie intégrante de la définition du mot « test » et sont garants de la qualité de celui-ci.

1.2. Utilisation des tests en orthophonie et recommandations

Les orthophonistes ont à leur disposition trois types d'outils différents : les tests de dépistage, les tests de première ligne et les tests de seconde ligne. Les premiers, en général utilisés pour les troubles du langage avant cinq ans, servent à repérer une population à risque de développer un trouble des apprentissages. Les tests de première ligne visent à mettre en évidence un trouble et sa sévérité ; ils contribuent à définir les actions pédagogiques, à prescrire des examens complémentaires et à préciser les actions de soins. Enfin, les tests de seconde ligne affinent le diagnostic, évaluent le fonctionnement de l'individu, identifient les mécanismes déficitaires et préservés et contribuent donc à orienter la rééducation (Inserm, 2007).

Un utilisateur de tests compétent utilise ces outils de manière appropriée, professionnelle et éthique, en tenant compte des besoins des patients et du contexte de la passation (Société Française de Psychologie, 2003). Ces recommandations s'adressent à ceux qui possèdent des tests, qui les choisissent, les administrent, les cotent et les interprètent et qui

informent les personnes testées de leurs résultats. Il est notamment préconisé une bonne connaissance des principes de base de la psychométrie. Cette connaissance nécessite une formation des utilisateurs de tests, primordiale non seulement pour l'administration des tests, mais aussi pour leur cotation et l'interprétation des résultats (Bernaud, 2007). Les orthophonistes sont concernés à chacune de ces étapes.

Or, au Canada, les orthophonistes semblent en majorité peu conscients dans leur pratique clinique de l'importance des critères de qualité d'un test, ont peu confiance en leurs connaissances en psychométrie et sont peu au fait des mauvais usages des tests (Kerr, Guildford & Bird, 2003). Les connaissances psychométriques des utilisateurs de tests en Belgique présenteraient également des lacunes : il peut en découler une interprétation approximative, voire « carrément fausse », des résultats (Grégoire, 2014, p.461). Enfin, il apparaît que, dans leur choix d'utilisation de tests pour diagnostiquer un trouble développemental du langage, les orthophonistes américains privilégient comme critère la date de sortie du test à ses qualités psychométriques (Betz, Eickhoff & Sullivan, 2013). À la lumière de ces études concernant les pratiques orthophoniques au Canada, en Belgique ou aux Etats-Unis, nous pouvons nous interroger sur les connaissances psychométriques des orthophonistes en France et sur l'utilisation des tests qui en découle.

Les orthophonistes français disposent en effet d'un choix conséquent de tests disponibles sur le marché de l'édition spécialisée ou publiés dans un article scientifique. Dès lors, comment choisir la meilleure épreuve pour évaluer les patients ? La connaissance des qualités psychométriques d'un test peut s'avérer un argument objectif participant au choix d'un test parmi d'autres.

2. Psychométrie

La psychométrie est une discipline primordiale à connaître, quelle que soit la spécialité du praticien évaluateur. Elle permet de développer des méthodes d'évaluation des caractéristiques individuelles (ex. le niveau d'intelligence, de langage ou d'anxiété d'un sujet). La mesure de ces caractéristiques est relative et non directe : il s'agit d'évaluer comment l'individu se situe comparativement à un échantillon représentatif de personnes. Plusieurs principes définissent la psychométrie (Bernaud, 2007 ; Hogan, 2012) :

- son parti pris quantitatif,
- l'utilisation de normes et le recours aux étalonnages (principe de standardisation), qui offrent un cadre d'interprétation aux résultats d'un test,
- la nécessité d'un cadre théorique,
- la variété des instruments utilisés,
- la notion fondamentale de qualité de la mesure, qui répond à trois critères : sensibilité, validité et fidélité/stabilité de la mesure.

Les qualités psychométriques (de validité, de fiabilité et de sensibilité) des épreuves de langage spécifiques à la pratique orthophonique devraient être celles exigées des instruments d'évaluation psychologique. C'est à cette condition qu'elles peuvent constituer des instruments valables d'évaluation (Rondal, 1997).

Nous définirons ci-dessous les différents critères qui contribuent à fournir des éléments étayant ces trois composantes psychométriques : la validité ; la fidélité / fiabilité / stabilité de la mesure ; la sensibilité et la spécificité. La présence de normes et le principe de standardisation seront également développés.

2.1. Validité

La validité d'un test concerne le lien entre ce que ce test mesure en réalité et ce qu'il est supposé mesurer. Cette validité est à estimer selon les objectifs de l'évaluation, tels que définis par l'auteur du test (Bernaud, 2007 ; Rondal, 1997). Il s'agit de la notion la plus fondamentale en psychométrie ; en effet, un instrument même extrêmement précis ne serait d'aucune utilité clinique s'il ne s'attachait pas en premier lieu à évaluer le concept visé par le test. La validité d'une épreuve peut s'envisager sous différents aspects : validité interne ou de contenu, validité prédictive, validité concourante, validité théorique (Rondal, 1997) et validité de construction (Bernaud, 2007).

Néanmoins, la question de l'existence des preuves de validité d'un instrument de mesure est soumise à débat : il ne s'agit pas d'une propriété intrinsèque à un test, ni d'un label que l'on pourrait accorder à celui-ci (Bernaud, 2007 ; Grégoire, 2009). Nous pouvons plus modestement fournir des éléments qui indiquent une validité satisfaisante d'une méthode psychométrique, ce qui est l'objectif dans ce mémoire pour les tests de langage oral utilisés en orthophonie. Nous allons définir ci-dessous les différents critères de validité.

2.1.1. Validité de contenu

La validité de contenu (parfois appelée validité interne) concerne le domaine couvert par les items d'une épreuve (Beech & Harding, 1994). Les items choisis pour un test doivent non seulement être représentatifs et appropriés pour évaluer la compétence en question (Huteau & Lautrey, 1999 ; Rondal, 1997), mais aussi être suffisamment nombreux pour refléter toutes les facettes et toutes les catégories de la fonction testée (Bernaud, 2007). Une validité de contenu satisfaisante suppose donc à la fois une bonne représentativité et une exhaustivité suffisante des items. Elle s'établit dès l'élaboration d'une épreuve par ses auteurs. Nous pouvons estimer sa qualité à travers différents éléments présents dans le manuel du test : un nombre d'items par subtest (ou sous-partie d'un test) supérieur à vingt (Beech & Harding, 1994), ainsi que la mention d'une analyse approfondie des items (McCauley & Swischer, 1984), à la fois qualitative (à travers une validation du choix des items par plusieurs experts du domaine) et quantitative. Une méthode d'analyse du pourcentage de réussite des réponses à chaque item permet d'extraire un indice de difficulté et un indice de discrimination (Bernaud, 2007) : elle contribue ainsi au choix définitif des items, par exemple via la suppression après différentes versions de l'épreuve d'items non adaptés, trop faciles, trop difficiles, ou non discriminants.

2.1.2. Validité théorique et validité de construit/de construction

La validité théorique se rapporte au modèle théorique servant de base à l'élaboration d'un test : de quelle façon l'instrument est-il lié aux théories de l'organisation des capacités cognitives et aux modèles de fonctionnement cognitif ? Les résultats du test correspondent-ils aux prédictions établies à partir d'un modèle théorique (Huteau & Lautrey, 1997 ; Rondal,

1997) ? La validité de construit vise quant à elle à démontrer l'acceptabilité théorique d'un test et à prouver le lien entre ce test et un ou plusieurs modèles théoriques (Bernaud, 2007 ; Bouchard et al., 2009). Le critère de validité de construction inclut non seulement la présence d'un modèle théorique dans le manuel du test, mais aussi le lien effectif entre ce modèle et ce qui est mesuré par le test.

Le manuel du test devrait donc mentionner les éléments suivants : présence d'un ou de plusieurs modèles théoriques explicités et indication des analyses statistiques prouvant sa validité de construit. Parmi d'autres méthodes, les analyses factorielles exploratoires ou confirmatoires et l'analyse multi-traits multi-méthodes apportent une contribution à la validité de construit d'un instrument (Bernaud, 2007 ; Pichot, 1999). Au sein d'un même test (ou d'une même batterie d'épreuves), il s'agit de techniques statistiques visant non seulement à relier les variables associées entre elles, mais aussi à dégager des facteurs distincts les uns des autres. Ces analyses sont valables à la condition que l'échantillon soit suffisant (Bernaud, 2007), soit cent sujets par sous-groupe (cf. paragraphe 2.3.2). Par ailleurs, des corrélations élevées ($r > 0.80$) avec un autre instrument valide mesurant le même construit sont également un indicateur d'une bonne validité de construit (Bernaud, 2007).

2.1.3. Validité concourante

Cette forme de validité, également appelée validité empirique (Rondal, 1997), consiste à mesurer la corrélation entre les résultats au test et ceux obtenus à d'autres épreuves validées, administrées au même échantillon et évaluant les mêmes compétences (Beech & Harding, 1994 ; McCauley & Swischer, 1984 ; Rondal, 1997). Un coefficient de corrélation minimum de 0.40 est attendu, à la condition que l'échantillon d'étalonnage par sous-groupes soit supérieur à cent sujets (Beech & Harding, 1994).

2.1.4. Validité prédictive

Cette dernière forme de validité consiste à mesurer le lien entre les résultats du test et l'évaluation du fonctionnement des sujets dans des tâches de vie quotidienne apparentées à celles de l'épreuve et utilisant la fonction testée (McCauley & Swischer, 1984 ; Rondal, 1997). Pour les tests de langage oral, il peut s'agir par exemple de certains résultats scolaires ou des scores à des épreuves évaluant le langage oral quotidien de l'enfant (via des questionnaires parentaux). Une corrélation suffisante ($r > 0.30$) doit ainsi être observée entre ces résultats et les scores obtenus par l'échantillon d'étalonnage (Beech & Harding, 1994). De plus, un échantillon minimum de cent sujets par sous-groupe est également indispensable pour valider cette corrélation.

2.2. Fidélité

De façon générale, la fidélité (ou fiabilité) concerne la stabilité des résultats obtenus à une épreuve (Pichot, 1999 ; Rondal, 1997). Pour être considéré comme fiable, un instrument de mesure doit aboutir aux mêmes résultats, quels que soient le moment de passation, l'évaluateur et l'extrait de l'épreuve qui est administré (Bernaud, 2007). De ce fait, il existe plusieurs critères de fidélité : la fidélité test-retest ; la fidélité inter-juges (ou inter-évaluateurs) ; la fidélité de partage par moitié (Rondal, 1997), plus récemment appelée

cohérence interne (Bouchard et al., 2009). Nous allons définir plus précisément ces différents critères de fidélité.

2.2.1. Fidélité test-retest

Elle se définit par la stabilité dans le temps des résultats au test (Bernaud, 2007 ; Pichot, 1999). Pour ce faire, on mesure les corrélations entre les résultats obtenus par le même échantillon d'étalonnage entre deux passations du test, à plusieurs semaines, mois ou années d'intervalle. Le coefficient de corrélation entre les deux séries de résultats doit être indiqué : un coefficient minimum de 0.80 est satisfaisant (Bernaud, 2007 ; Grégoire, 2009), à la condition que la taille de l'échantillon par sous-groupe soit de cent participants minimum (Beech & Harding, 1994).

2.2.2. Fidélité inter-juges

Elle vise à vérifier le degré d'accord entre deux observateurs, deux évaluateurs, ou deux cotateurs (Bernaud, 2007 ; Pichot, 1999). La démarche pour mesurer cette forme de fidélité est la suivante : faire évaluer simultanément les mêmes sujets par deux ou plusieurs examinateurs, de façon indépendante, puis corréler les notes obtenues. Les coefficients de corrélation inter-juges doivent être mentionnés dans le manuel : le seuil minimum requis pour ces coefficients est le même que celui attendu pour la fidélité test-retest, mentionné au paragraphe 2.2.1.

2.2.3. Cohérence interne / homogénéité

La mesure de la cohérence interne garantit également la fidélité d'un test (Bernaud, 2007). Elle concerne l'homogénéité des items d'une même épreuve ou d'un même subtest : les items supposés mesurer une même compétence aboutissent-ils chacun à des résultats semblables (Bouchard et al., 2009) ? Le coefficient d'homogénéité peut s'établir en calculant le coefficient de corrélation entre le score de chaque item et la note globale de l'échelle moins l'item considéré (Pichot, 1999). D'autres méthodes sont statistiquement plus solides : la fidélité de partage par moitié (Rondal, 1997) ou la fidélité pair-impair. La méthode la plus fréquemment utilisée pour estimer la cohérence interne d'un test est le coefficient « alpha de Cronbach » (Bernaud, 2007).

Le manuel doit indiquer le niveau de cohérence interne ou d'homogénéité des items et préciser par quelle méthode cette cohérence est mesurée (corrélation inter-items, partage par moitié, coefficient pair-impair, alpha de Cronbach). Un coefficient « alpha de Cronbach » est correct à partir de 0.80 (Bernaud, 2007). Il est jugé limite à 0.70, sous réserve là aussi que l'échantillon soit supérieur à cent sujets (Beech & Harding, 1994).

2.2.4. Critères en lien avec la fidélité

D'autres critères sont étroitement liés à la fidélité d'un test. Dans le manuel, la présence d'instructions claires d'administration, de cotation et d'interprétation des résultats sont un moyen de garantir les résultats les plus objectifs possible et une meilleure fidélité inter-juges. En effet, quel que soit l'évaluateur qui administre le test, le fait de pouvoir suivre un protocole clairement établi rend la passation du test la plus proche possible d'une passation qui pourrait

être administrée par un autre collègue. Le manuel est censé fournir de nombreuses indications sur la passation, parmi lesquels : les modalités matérielles de l'épreuve, les consignes orales explicites données aux personnes évaluées, les contraintes de temps (utilisation ou non d'un chronomètre), la prise en compte d'une autocorrection, la possibilité d'un renforcement en cas de difficultés et la manière précise de l'établir. Les critères de cotation des épreuves doivent également être très précis : nombre de points accordés selon la réponse, indications des réponses attendues et exemples de réponses refusées. Des pistes d'interprétation des scores (ex. observer si les scores sont homogènes ou hétérogènes) ou la présentation d'exemples de divers cas cliniques sont utiles afin d'orienter le clinicien dans son analyse.

De même, la mention dans le manuel de la qualification de l'examinateur permet une meilleure fiabilité de l'instrument. Les qualifications de la personne qui administre, cote et interprète le test doivent être précisées afin de garantir la qualité des données recueillies (McCauley & Swisher, 1984). En effet, pour utiliser un test, l'évaluateur doit être formé à son usage. Si le test n'indique pas clairement les professionnels à qui le test est destiné, la passation et l'interprétation du test peuvent être sujets à caution.

Enfin, un dernier critère, particulièrement important en psychométrie, est une mesure directement liée à la fidélité : l'erreur standard de mesure (ou erreur type de mesure). Il faut en effet toujours considérer que les scores observés en situation de test ne sont qu'une approximation du véritable niveau de compétences des sujets. Cette estimation des réelles compétences dépend de l'erreur de mesure (Bernaud, 2007). Si les scores vrais étaient ceux mesurés, la fidélité des tests serait parfaite et l'erreur de mesure serait nulle. L'erreur type de mesure se calcule grâce au coefficient de fidélité (Grégoire, 2009). Elle correspond au niveau de fluctuation de la mesure, due à l'infidélité de la méthode, et elle permet d'estimer la marge d'erreur des résultats. On considère souvent en pratique que la marge d'erreur acceptable ne doit pas dépasser 5 % (Rondal, 1997). Si nous disposons de l'erreur de mesure du test, il est possible d'établir une zone autour du score observé, l'intervalle de confiance, où le véritable score du sujet a des chances de se situer. La valeur de l'erreur-type de mesure, ainsi que celui de l'intervalle de confiance pour chaque test ou subtest, doivent également apparaître dans le manuel du test.

2.3. Normes et standardisation

2.3.1. Description de l'échantillon d'étalonnage

Les caractéristiques de l'échantillon d'étalonnage doivent être explicitées et représentatives de la population tout-venant (Bouchard et al., 2009 ; McCauley & Swisher, 1984). Ce sont principalement l'âge, le genre / le sexe et l'origine géographique des participants, la catégorie socio-professionnelle (CSP) ou le niveau socio-culturel (NSC) ou le niveau de formation des parents, et éventuellement le caractère normal ou pathologique des sujets. Le manuel indique quels sont les critères d'inclusion et d'exclusion qui ont permis la sélection de l'échantillon.

2.3.2. Taille de l'échantillon

Elle doit être d'au moins cent participants par sous-groupe (Bernaud, 2007 ; Bouchard et al., 2009 ; Friberg, 2010 ; McCauley & Swisher, 1984).

2.3.3. Mesure de la tendance centrale

Le but d'un test est de situer le score brut du sujet au sein de l'ensemble des scores obtenus par l'échantillon (Pichot, 1999). La distribution des notes obtenues par la population de l'échantillon peut être représentée par une courbe. Quand elle est « en cloche » et symétrique, elle est dite gaussienne (ou normale) et définie par deux paramètres : une valeur centrale et un indice de dispersion. L'indice de dispersion permet de préciser si la courbe est plus ou moins « étalée » autour de la valeur centrale.

L'indice de dispersion le plus utilisé est l'écart-type. Si la distribution des notes est gaussienne (et seulement à cette condition), nous pouvons mesurer le résultat de l'individu comparativement à la population de référence en utilisant la formule suivante : score *z* (de l'individu) = (moyenne du groupe de référence – note brute de l'individu) / écart-type du groupe de référence. En revanche, si la distribution des notes obtenues n'est pas gaussienne, nous devons utiliser le procédé des percentiles : il consiste à définir la place du sujet dans la distribution en indiquant le pourcentage d'individus obtenant une note inférieure à la sienne (Pichot, 1999).

Le manuel du test doit donc préciser si la distribution des scores de l'échantillon suit une loi gaussienne, et donc si l'on peut comparer le score brut de l'individu à ceux de la population de l'échantillon en utilisant la formule précitée. Si la distribution n'est pas gaussienne, le manuel doit préciser que c'est le procédé des percentiles que l'utilisateur doit utiliser. Afin de disposer d'un seuil pathologique, l'idéal requiert la mention du percentile 5 ou des percentiles 2 - 3.

2.4. Sensibilité et spécificité

La sensibilité concerne la capacité d'un instrument de mesure à discriminer le plus finement possible les sujets et à détecter une pathologie chez des enfants effectivement atteints (Plante et Vance, 1994). Elle vise donc à limiter le nombre de faux-négatifs au test. Dans le domaine de l'orthophonie, un test sensible permet de faire émerger le maximum de différences fines entre les individus (Perdrix, 2018). La sensibilité est en lien avec le nombre d'items du test : en général, plus les items valides d'une échelle sont nombreux, plus l'épreuve devient sensible (Pichot, 1999 ; Rondal, 1997). L'échantillon d'étalonnage doit également être suffisant pour garantir une bonne sensibilité (Perdrix, 2018).

La spécificité d'un test est sa capacité à rejeter la présence d'un trouble chez les enfants qui n'en sont pas atteints (Plante et Vance, 1994) : il s'agit de son pouvoir discriminatif, c'est-à-dire celui de réduire le nombre de faux-positifs.

Le calcul d'indices de sensibilité et de spécificité est possible à partir du nombre de vrais-positifs, de vrais-négatifs, de faux-positifs et de faux-négatifs. De bons niveaux conjoints de sensibilité et de spécificité sont requis : leurs indices doivent apparaître dans le manuel et être supérieurs à 0.80 (Plante et Vance, 1994).

3. État des lieux de la recherche sur les qualités psychométriques des tests de langage oral

Aux Etats-Unis, McCauley et Swisher (1984) ont étudié trente tests de langage et ont examiné pour chacun la présence des dix critères suivants :

- la description des caractéristiques de l'échantillon d'étalonnage ;
- la taille de l'échantillon ;
- une analyse des items (les items mesurent-ils bien ce qu'ils sont censés mesurer ?) ;
- la mesure de la tendance centrale, ou mesure de comparaison à la norme ;
- la validité concourante ;
- la validité prédictive ;
- la fidélité test-retest ;
- la fidélité inter-juges ;
- la description précise des consignes d'administration, de cotation et d'interprétation du test ;
- les qualifications de l'évaluateur.

Dans cette étude, seuls 20 % des tests satisfont à la moitié de ces critères. Les informations nécessaires pour évaluer chaque critère sont souvent absentes des manuels. Les critères de validité et de fidélité sont les plus fréquemment absents.

D'autres auteurs ont également proposé des critères d'analyse des outils d'évaluation du langage. Hutchinson (1996), Bouchard et al. (2009), Friberg (2010) et Flipsen et Ogiela (2015) ont repris tout ou partie des critères de McCauley et Swischer et en ont ajouté. La synthèse de l'utilisation de ces critères (qu'ils soient satisfaits ou non) dans ces différentes recherches est présentée par ordre de fréquence dans le tableau 1 ci-après. Nous constatons que douze de ces critères sont utilisés dans au moins trois de ces études.

A propos des outils francophones d'évaluation du langage oral chez l'enfant, les résultats de recherches antérieures montrent qu'ils offrent peu de preuves de qualité psychométrique. En effet, plus de la moitié des outils utilisés au Canada ne fournissent pas de preuves de validité et de fidélité (Bouchard et al., 2009) ; ce manque serait dû à une absence d'information dans les manuels, et non obligatoirement à une mauvaise performance des tests à un ou plusieurs critères. Nous pouvons émettre l'hypothèse qu'il en est de même pour les outils utilisés en France. En effet, Leclercq et Veys (2014) ont étudié quatre batteries fréquemment utilisées par les orthophonistes (ELO, N-EEL, Exalang et L2MA-2) : moins de la moitié des critères de qualité psychométrique étaient satisfaisants pour trois batteries ; aucune d'entre elles ne satisfaisait l'ensemble des critères choisis.

Tableau 1 : Critères psychométriques étudiés dans la littérature.

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

Note. O : critère utilisé - N : critère non utilisé

4. Objectif et hypothèses

L'objectif de notre mémoire est d'analyser la qualité psychométrique des tests évaluant les troubles du langage oral chez l'enfant, utilisés par les praticiens francophones d'Europe, en nous appuyant sur différentes composantes psychométriques (validité, fidélité, sensibilité, normes et standardisation). Nous souhaitons également connaître quels critères psychométriques sont les mieux satisfaits dans l'ensemble de ces tests, et au contraire quels critères le sont le moins.

Suite à notre revue de littérature exposée ci-dessus (paragraphe 3), nous formulons l'hypothèse suivante : peu d'outils évaluant le langage oral chez l'enfant, disponibles et

utilisés par les orthophonistes, répondent aux standards psychométriques requis. Concernant l'analyse précise de chaque critère psychométrique dans l'ensemble des tests, nous nous situons dans une démarche exploratoire et ne formulons pas d'hypothèse précise pour chaque critère. Nous supposons globalement que les preuves de validité et de fidélité sont lacunaires dans les tests de langage oral.

Méthode

Nous exposerons tout d'abord comment nous avons choisi les tests examinés dans notre étude et les critères d'analyse psychométrique retenus. Ensuite, nous indiquerons comment nous avons évalué chaque critère psychométrique.

1. Choix des tests

Pour sélectionner les tests de langage oral, nous nous fondons sur deux listes répertoriant les outils utilisés par les cliniciens et les chercheurs francophones d'Europe. Nous avons utilisé en premier lieu une liste des tests et batteries de langage oral employés par les logopèdes exerçant en Belgique francophone, établie par l'Institut National d'Assurance Maladie - Invalidité (INAMI, 2017). La deuxième liste de référence a été celle publiée par l'Union Nationale pour le Développement de la Recherche et de l'Evaluation en Orthophonie (UNADREO). Elle recense les tests orthophoniques utilisés pour l'évaluation du langage oral, en phonologie, lexique, syntaxe et pragmatique (UNADREO, 2011).

Nous avons sélectionné 36 tests et batteries d'épreuves présents dans au moins une de ces listes. Nous avons au préalable vérifié que chacun de ces tests était toujours diffusé et commercialisé à ce jour ou que l'article de référence le concernant était disponible et accessible. Nous avons écarté sept batteries contenant à la fois des épreuves de langage oral et de langage écrit, qui font l'objet du mémoire de Colli-Vaast (2018). Finalement, 29 tests ou batteries d'épreuves de langage oral ont été analysés :

1. BEPL-A (Chevrie-Muller et al., 1997a),
2. BEPL-B (Chevrie-Muller et al., 1997b),
3. BILO Petits (Khomsi & Khomsi, 2009),
4. BILO-2 (Khomsi & Pasquet., 2007),
5. BILO-EC2 (Khomsi & Khomsi, 2007),
6. BILO-3 (Khomsi et al., 2007),
7. Children's Communication Checklist – C.C.C. (Bishop, 1998 ; Bishop & Baird, 2001 ; Maillart, 2003),
8. Décision lexicale (Maillart & Schelstraete, 2004),
9. DEDALE (Deltour & Hupkens, 2011),
10. Discrimination (Maillart & Schelstraete, 2004),
11. DLPF – Outil pour l'évaluation du Développement du Langage de Production en Français (Bassano et al., 2005),
12. ECSP - Evaluation de la communication sociale précoce (Guidetti & Tourrette, 2009),
13. ELDP (Macchi et al., 2012),

14. ELO (Khomsi, 2001),
15. ELOLA (de Agostini et al., 1998),
16. Étude de la motricité et des praxies oro-faciales chez l'enfant de 2 ans et demi à 12 ans et demi (Henin, 1981),
17. EVALO 2-6 (Coquet et al., 2009),
18. EVIP (Dunn et al., 1993),
19. EXALang 3-6 (Helloin & Thibault, 2006),
20. IFDC - Inventaire Français du Développement Communicatif (Kern & Gayraud, 2010),
21. ISADYLE (Piérart et al., 2010),
22. Kikou 3-8 (Boutard & Bouchet, 2009a),
23. N-EEL (Chevrie-Muller & Plaza, 2001),
24. PEES 3-8 (Boutard & Bouchet, 2009b),
25. Péléa 11-18 (Guillon et al., 2011),
26. Test de dépistage de dénomination et de désignation d'images (Kremin & Dellatolas, 1995),
27. Test des concepts de base Boehm – 3 (Boehm, 2009),
28. ThaPho – Test d'habiletés phonologiques (Ecalle, 2007),
29. T.L.O.C.C. (Maurin, 2006).

Nous nous sommes procuré les différents manuels de ces tests ou batteries soit auprès d'orthophonistes les possédant, soit à la testhotèque du département d'orthophonie de Lille. Nous avons également consulté les études de référence et les articles complémentaires traitant des tests.

2. Choix des critères psychométriques

Suite à notre revue de littérature, nous choisissons d'utiliser pour notre analyse les critères psychométriques les plus fréquemment retrouvés, présents dans au moins trois études (cf. tableau 1). Il s'agit des éléments suivants : description de l'échantillon, validité concourante, fidélité test-retest, fidélité inter-juges, taille de l'échantillon, tendance centrale, validité prédictive, qualifications de l'examinateur, procédures d'administration / cotation / interprétation du test, objectifs du test, analyse des items / validité de contenu (deux critères étroitement liés et regroupés : cf. contexte théorique, paragraphe 2.1.1), erreur standard de mesure.

Par ailleurs, nous conservons trois autres critères : validité théorique / de construit, cohérence interne et sensibilité / spécificité. Au vu de nos lectures (cf. contexte théorique, paragraphes 2.1.2, 2.2.3 et 2.4) et malgré leur moindre fréquence dans les articles de recherche sur les qualités psychométriques des tests, ces trois critères nous ont paru importants. En effet, ils garantissent respectivement : un appui théorique solide en lien avec une analyse de la construction du test ; l'homogénéité des items de l'épreuve ; la possibilité de détecter spécifiquement un trouble. Notre analyse psychométrique s'appuie donc sur quinze critères, présentés dans le tableau 2 ci-après.

Tableau 2 : Critères psychométriques étudiés dans notre mémoire.

Validité	1. Validité de contenu/analyse des items 2. Validité théorique/de construit 3. Validité concourante 4.Validité prédictive
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
Normes et standardisation	12. Description de l'échantillon d'étalonnage 13. Taille de l'échantillon 14. Tendance centrale
15. Sensibilité/spécificité	

3. Méthode de cotation pour chaque critère

Une méthode utilisant deux ou trois niveaux de cotation pour chaque critère est fréquemment retrouvée dans les études traitant de ce sujet. Cependant, une méthode de cotation binaire (critère satisfaisant / non satisfaisant) nous a paru insuffisante dans notre analyse, de même qu'une cotation à trois niveaux (critère absent du manuel, critère présent mais non satisfaisant, critère présent et satisfaisant). Nous avons jugé pertinent d'ajouter une catégorie intermédiaire quand le critère n'était pas satisfaisant. Quatre niveaux de cotation ont donc été retenus. Le tableau 3 ci-dessous présente les éléments attendus dans le manuel ou dans l'article de référence du test pour coter les quinze critères de 0 (critère absent) à 3 (critère parfaitement satisfaisant).

Tableau 3 : Echelle de cotation des critères psychométriques.

Critères	0	1	2	3
Validité de contenu / analyse des items	Non indiquées	Analyse uniquement qualitative[1] des items, quel que soit le nombre d'items	Analyse qualitative et partiellement quantitative[2] des items Nombre d'items par épreuve < 20	Analyse qualitative et quantitative des items Nombre d'items par épreuve ≥ 20
Validité théorique / 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 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 analyse statistique complète[3] et plus de 100 sujets par sous-groupe
Validité concourante (seuil : $r = 0.40$)	Non indiquée	Indiquée mais indications incomplètes (pas de preuve chiffrée)	Indiquée mais inférieure au seuil ; ou moins de 100 sujets par sous-groupe	Supérieure au seuil et plus de 100 sujets par sous-groupe
Validité prédictive (seuil : $r = 0.30$)				
Fidélité test-retest (seuil : $r = 0.80$)				
Fidélité inter-juges (seuil : $r = 0.80$)				
Cohérence interne (seuil : $r = 0.80$)				
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 99 % 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
Instructions d'administration / cotation / interprétation	Non indiquées	Clair et précis sur seulement un type d'instructions[4]	Clair et précis sur 2 types d'instructions	Clair et précis sur les 3 types d'instructions

1 Nous appelons analyse qualitative une mention des critères de choix des items (ex. complexité, structure, fréquence), sans preuve chiffrée, ou l'indication des pourcentages de réussite des items, sans analyse statistique approfondie contribuant au choix des items.

2 Nous appelons analyse quantitative partielle l'analyse statistique effectuée. Ex. analyse classique des items avec un seul indice renseigné : indice de difficulté ou de discrimination.

3 Ex. Analyse factorielle, analyse multi-traits multi-méthodes.

4 Nous considérons trois types d'instructions : l'administration, la cotation et l'interprétation.

Critères	0	1	2	3
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[5]	Indication précise en lien avec l'objectif du test : orthophoniste, ou liste précise de professionnels (dont l'orthophoniste)
Erreur standard de mesure	Non indiquée	Pas de tableau des erreurs type de mesure (ETM), ni de tableau des intervalles de confiance (IC), mais présence de données[6] permettant à l'utilisateur de calculer les ETM et donc les IC	Présence d'un tableau des ETM, mais pas de tableau des IC	Présence de tableaux des IC
Description de l'échantillon d'étalonnage	Non indiquée	1 ou 2 critères principaux présents[7], 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 preuve de représentativité de l'échantillon
Taille de l'échantillon	Non indiquée	Moins de 50 participants par sous-groupe	Entre 50 et 99 participants par sous-groupe	100 participants ou plus par sous-groupe
Tendance centrale	Aucune mesure de tendance centrale	Mesure de tendance centrale indiquée, mais non adaptée à la distribution de la population contrôle[8]	Mesure de tendance centrale adaptée à la distribution de la population contrôle, mais indication confuse ou imprécise ou mal placée dans le manuel ou le logiciel	Mesure de tendance centrale adaptée à la distribution de la population contrôle et indication claire, précise[9] et bien placée dans le manuel ou le logiciel
Sensibilité / Spécificité (seuil : $r = 0.80$)	Non évoquées	Evocation de l'une ou l'autre, sans indication du seuil	Indication du taux de sensibilité et/ou de spécificité, mais inférieur au seuil	Indication du taux de sensibilité et/ou de spécificité, égal ou supérieur au seuil

4. Fiabilité de la méthode de cotation

La fiabilité intra-examinatrice de la cotation des quinze critères psychométriques a été mesurée sur 5 tests et batteries sélectionnés aléatoirement parmi les 29 étudiés dans notre mémoire. Ceci représente 17 % des données totales. Cette fiabilité a été mesurée de deux manières. Premièrement, 93 % de nos mesures sont identiques entre la première et la seconde

5 Ex. Le test peut être utilisé dans une école, mais les enseignants ne sont pas mentionnés comme potentiels utilisateurs.

6 La connaissance du coefficient de fidélité test-retest et des écarts types permettent de calculer les ETM et donc les IC. Les IC peuvent être calculés à partir des ETM selon la formule : IC = [SB – ETM seuil ; SB + ETM seuil] (SB : score brut).

7 Critères principaux : âge, sexe, origine géographique, CSP ou NSC des parents.

8 Ex. Les moyennes et ET sont indiqués, mais sans analyse de la distribution des données contrôle, ou avec des éléments montrant une distribution non gaussienne pour au moins une note, et l'utilisation des percentiles n'est pas recommandée en cas de distribution non gaussienne.

9 Ex. pour les percentiles : indication des P2-3 ou du P5.

cotation de ces cinq tests, à trois semaines d'intervalle. Deuxièmement, une corrélation positive significative a été trouvée entre ces deux cotations (r_s = .98, p < .0001).

Par ailleurs, la fiabilité inter-examinatrices de la cotation des critères psychométriques a été mesurée sur 3 tests et batteries sélectionnés aléatoirement parmi les 29 étudiés. Ceci représente 10 % des données totales. Les deux examinatrices (Colli-Vaast et nous-même) fournissaient 87 % de mesures identiques. Une corrélation positive significative a été trouvée entre les deux cotations (r_s = .93, p < .001).

L'ensemble de ces résultats est satisfaisant. La liste des tests ayant fait l'objet de ces mesures de fiabilités intra et inter-examinatrices se trouve en annexes (annexe n°1).

Résultats

Les résultats de l'analyse des 29 tests sont indiqués dans le tableau 4 en annexes (annexe n°2). Pour chaque critère, nous avons utilisé la cotation en quatre niveaux développée précédemment. Chaque test est donc coté globalement sur 45 pour l'ensemble des 15 critères. A titre d'exemple, les éléments de preuve étayant ou non chaque critère psychométrique sont présentés en annexes (annexes n° 3 et 4) pour deux de ces tests (l'ELO et l'EVIP). En premier lieu, nous analyserons globalement l'ensemble des cotations, puis nous développerons plus précisément nos résultats, en suivant deux axes : analyse des tests et analyse des critères psychométriques.

1. Analyse globale des résultats de cotation

L'objectif de cette analyse est de connaître globalement le degré de précision avec lequel les critères psychométriques sont communiqués, décrits et / ou analysés par les auteurs des 29 tests de langage oral objets de notre étude. Pour cela, nous avons réalisé le calcul suivant : le nombre total de cotes 0 divisé par le nombre total de cotations de critères, tous tests et critères confondus, c'est-à-dire 435 cotes (29 tests x 15 critères). Nous avons réalisé de même pour chacune des cotes 1, 2 et 3. Chacun de ces quatre ratios a été rapporté en pourcentages pour plus de lisibilité des résultats (figure 1).

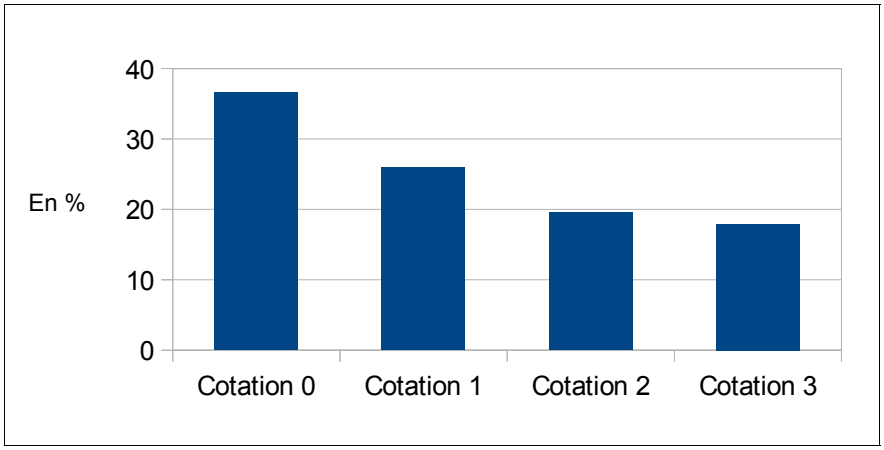

Figure 1: Pourcentages de cotes 0, 1, 2 et 3 pour l'ensemble des critères et des tests.

Nous constatons les éléments suivants : moins de 20 % des cotes sont à 3 (critère totalement satisfaisant), et moins de 20 % à 2 (critère comportant des éléments de preuve encore insuffisants). Plus de la moitié des cotes indiquent soit une simple mention du critère sans aucun élément de preuve (cote 1), soit une absence d'information dans le manuel (cote 0). La cote 0 est la cote que l'on retrouve le plus fréquemment (37 % de l'ensemble des cotes).

2. Analyse des tests

Nous présenterons d'abord une analyse globale de nos résultats, puis une analyse de chacun des 29 tests.

2.1. Analyse globale

L'objectif de cette analyse est d'abord de connaître le score total moyen de qualité psychométrique par test (score maximum : 45), ainsi que le nombre moyen de critères psychométriques par test entièrement satisfaits, donc cotés 3 (score maximum : 15). Ensuite, nous souhaitons présenter le nombre moyen de critères par test apportant des éléments de preuve : soit insuffisamment prouvés (cote 2), soit entièrement satisfaits (cote 3). Enfin, nous y ajoutons le nombre moyen de critères cités et présents par test : soit simplement mentionnés sans preuve (cote 1), soit apportant des éléments de preuve (cotes 2 ou 3). L'objectif de de cette présentation est de comparer nos résultats lorsque nous augmentons notre niveau d'exigence des éléments de preuve attendus pour satisfaire un critère.

Pour répondre à ces objectifs nous avons réalisé une analyse statistique descriptive illustrant la distribution de ces données. L'ensemble de ces informations est présenté dans les figures ci-après (figures 2 et 3).

La figure ci-dessous (figure 2) représente l'étendue et la distribution des scores totaux (sur 45) de notre cotation pour l'ensemble des tests.

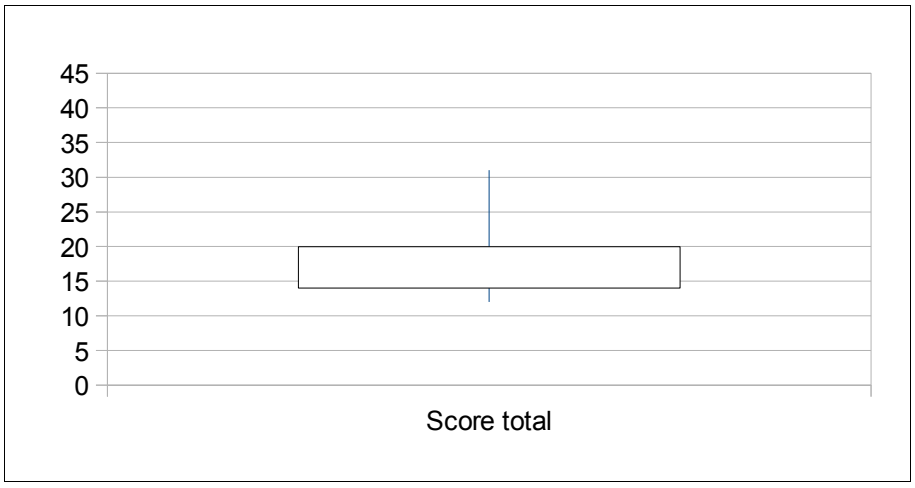

Figure 2 : Distribution des scores totaux par test.

Nous constatons les résultats suivants. La moyenne du score total par test est de 17,8 sur 45 (*ET* = 4,8). Les scores s'étendent de 12 à 31. La médiane se situe à 17 : la moitié des 29 tests obtient un score total sur 45 inférieur ou égal à 17. Le quartile 1 se situe à 14, le quartile 3 à 20. La majorité de scores est faible (entre 12 et 20) et il existe quelques scores qui étirent la distribution vers le haut.

La figure 3 ci-dessous nous indique le nombre moyen de critères par test, lorsque nous augmentons notre niveau d'exigence.

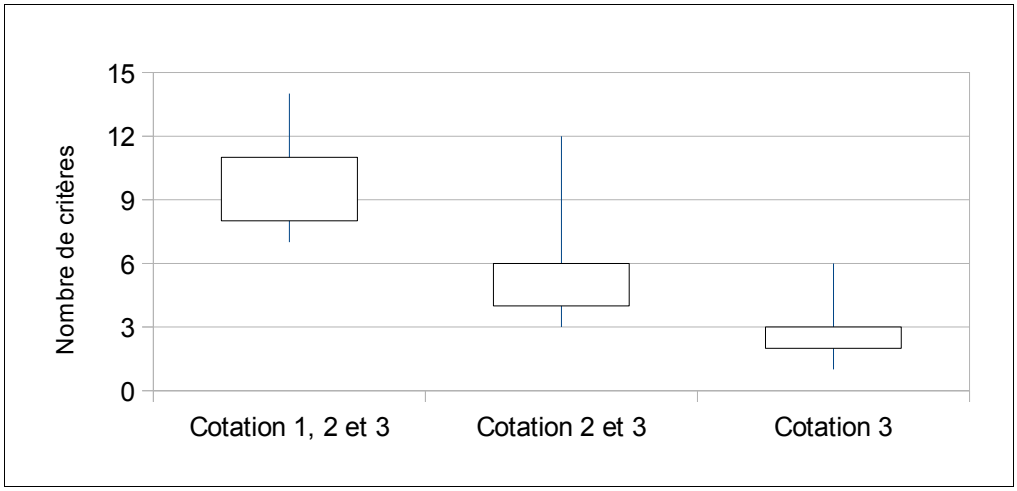

Figure 3 : Comparaison de la répartition du nombre de critères par test selon leur niveau de cotation.

La figure de gauche nous indique l'étendue du nombre de critères présents par test : ce nombre s'étend de 7 à 14 avec une médiane à 9. Celle du milieu montre une étendue entre 3 et 12 du nombre de critères par test cotés 2 ou 3, avec une médiane à 5 : la moitié des tests fournissent donc des éléments de preuve étayant seulement cinq critères ou moins, sans être toujours suffisants.

Enfin, la figure de droite représente le nombre de critères entièrement satisfaits par test : ce nombre s'étend de 1 à 6, avec une médiane à 2. Quinze tests présentent donc des niveaux de preuve totalement suffisants pour seulement deux critères ou moins. Ainsi, plus nous augmentons le niveau d'exigence en termes de preuves attendues pour satisfaire un critère, plus le nombre de critères par test baisse. Nous notons que, pour chacune de ces distributions, la moyenne est très légèrement supérieure à la médiane (car quelques tests étirent la distribution vers le haut).

2.2. Analyse par test

Afin d'approfondir ces résultats, le graphique ci-après (figure 4) représente le nombre de critères entièrement satisfaits (cotés 3) pour chacun des 29 tests.

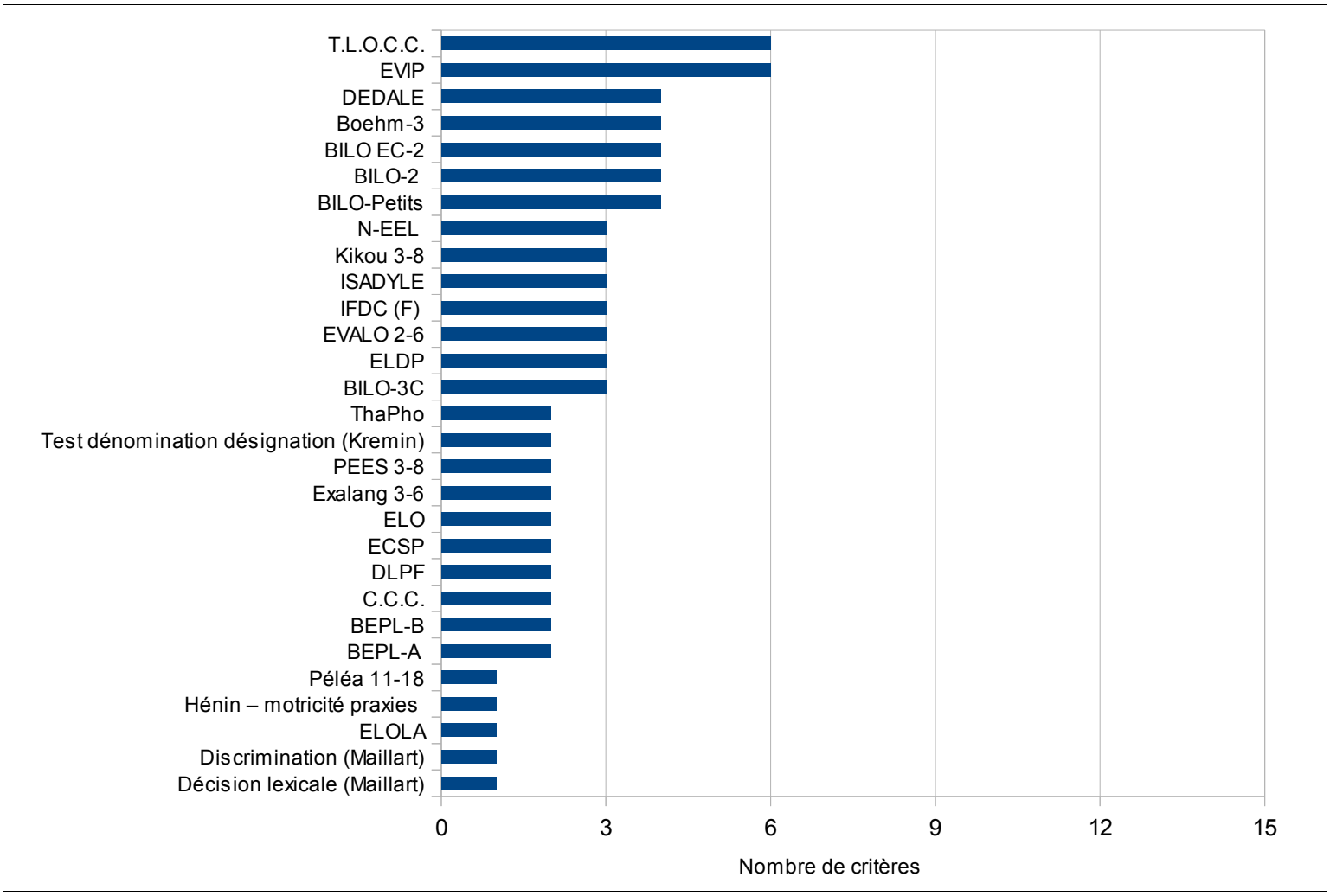

Figure 4 : Nombre de critères cotés 3 par test.

Aucun test ne remplit de façon totalement satisfaisante la moitié des critères (soit huit critères). Seuls un quart des tests (7 tests sur 29) satisfont à un quart ou un peu plus d'un quart des critères (soit quatre ou cinq critères sur quinze) : il s'agit du TLOCC, de l'EVIP, de DEDALE, du Boehm-3 et de trois des épreuves BILO. Cinq tests ne remplissent parfaitement qu'un seul critère : les épreuves de Maillart et ses collègues, l'ELOLA, l'épreuve de Hénin et le Péléa 11-18.

Afin d'affiner notre analyse, il nous paraît intéressant de tenir compte de la cotation intermédiaire de niveau 2, qui indique des éléments en faveur de la présence d'un critère, mais qui ne sont pas complètement satisfaisants ou suffisants pour en fournir la preuve. Le graphique ci-dessous (figure 5) est construit sur le même principe que le précédent, mais indique également les critères cotés au niveau 2, en sus de ceux cotés au niveau 3 (déjà complètement satisfaisants). L'objectif est de visualiser quels tests présentent des éléments intéressants pour étayer un critère, que celui-ci soit excellemment rempli ou encore lacunaire.

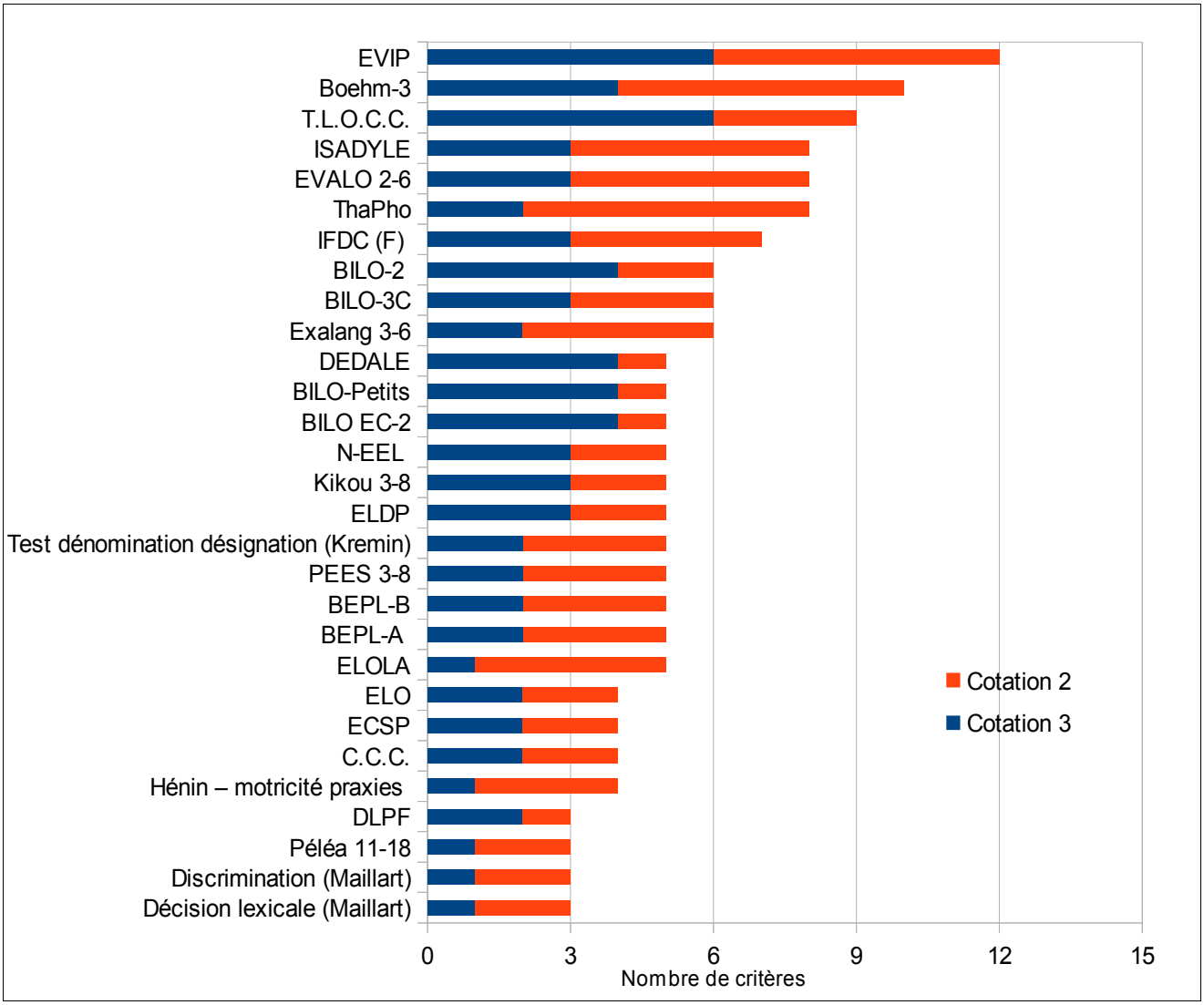

Figure 5 : Nombre de critères cotés 2 ou 3 par test.

En tenant compte conjointement des niveaux 2 et 3 de cotation, six tests fournissent des éléments intéressants, malgré une insuffisance de preuves parfois, concernant la moitié des critères : il s'agit de l'EVIP, du Boehm-3, du TLOCC, d'Isadyle, d'Evalo 2-6 et du ThaPho. L'EVIP est même relativement correct, voire très bon, pour environ 75 % des critères (12 critères cotés à 2 ou 3). Dix-neuf tests, soit la majorité des tests, sont évalués à 2 ou 3 pour une proportion de 25 à 50 % des critères, ce qui reste faible : pour ces dix-neuf tests, au moins la moitié des critères est donc cotée à 0 ou 1. De plus, quatre tests respectent moins de quatre critères (soit 25 %) cotés 2 ou 3, ce qui signifie que les trois quarts des critères sont absents ou très largement insuffisants pour ces quatre tests (il s'agit du DLPF, du Péléa 11-18, et des épreuves de Discrimination et de Décision lexicale de Maillart et al.).

3. Analyse des critères

L'analyse des critères fera l'objet de cette partie. Quels critères sont les mieux remplis, ou au contraire les moins satisfaisants, parmi les 29 tests de notre étude ?

La figure ci-dessous représente pour chaque critère la somme de nos cotations des 29 tests (le score maximal pour chaque critère est donc de 87).

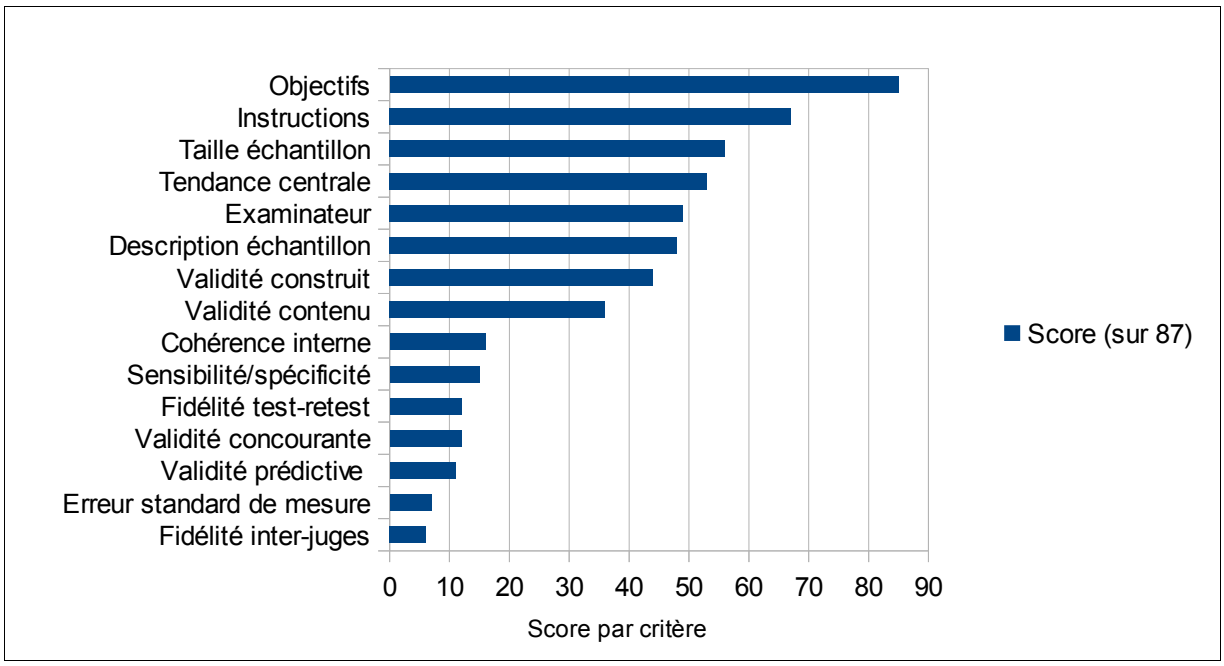

Figure 6 : Score total de chaque critère

Nous constatons que certains critères sont plus fréquemment retrouvés dans les manuels : les objectifs du test, les instructions de passation, la taille et la description de l'échantillon, la mesure de tendance centrale, les qualifications de l'examinateur et la validité de construit. Mais cela ne signifie pas que ces critères soient satisfaisants, puisque ce score total ne tient pas compte du niveau de cotation pour chacun des tests considéré séparément (nous ne savons pas si la somme des cotes pour chaque critère est composée majoritairement de cotes 1, 2 ou 3).

Afin d'affiner notre analyse, la figure ci-après représente le nombre de tests entièrement satisfaisants (cotés 3) pour chaque critère.

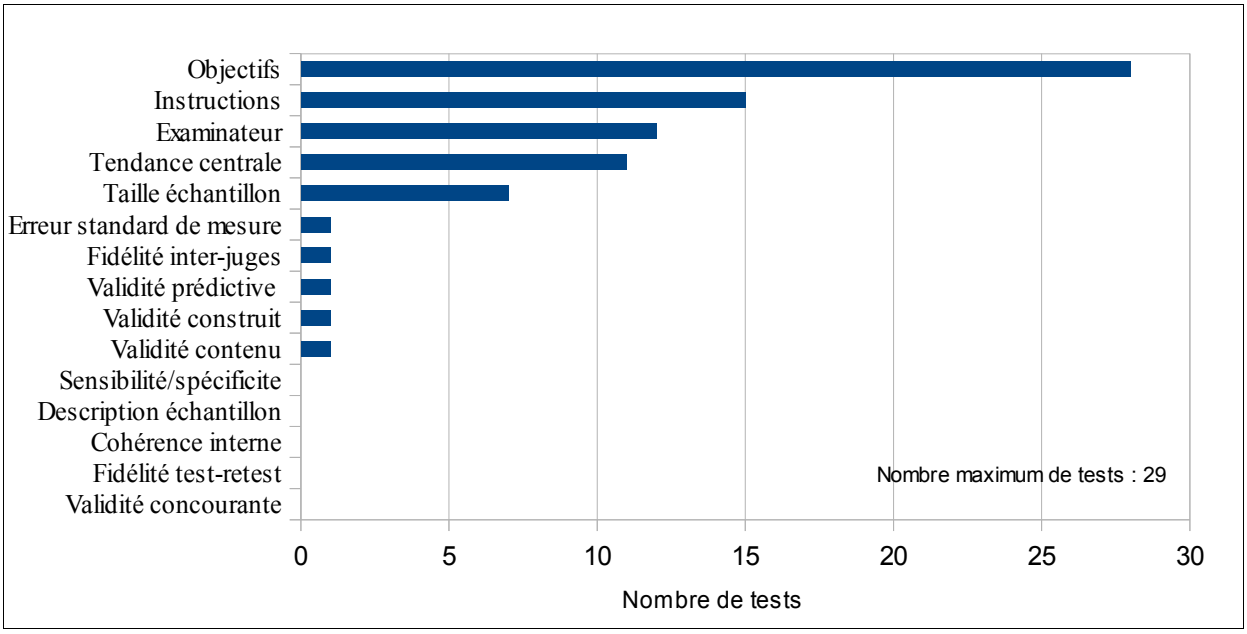

Figure 7 : Nombre de tests cotés 3 pour chaque critère.

L'indication précise des objectifs du test est le seul critère que l'on retrouve dans la quasi-totalité des manuels. Les consignes de passation, de cotation et d'interprétation sont complètes dans la moitié des tests. Trois critères sont moins fréquemment retrouvés : les qualifications de l'examinateur, la mesure de tendance centrale et la taille de l'échantillon sont totalement satisfaisantes pour 24 à 31 % des tests. Concernant cinq autres critères, un bon niveau de preuves est atteint pour un seul des tests parmi les vingt-neuf : il s'agit des validités prédictive, de contenu, de construit, de la fidélité inter-juges et de l'erreur standard de mesure. Les autres critères (validité concourante, fidélité test-retest, cohérence interne, description de l'échantillon et sensibilité/spécificité) ne sont jamais retrouvés de façon totalement complète ou satisfaisante dans l'ensemble des tests.

Il nous paraît également intéressant d'observer quels critères sont cotés au niveau 2 : certes, cela signifie que le critère est jugé insuffisamment rempli, mais cette cotation indique que le test offre davantage d'éléments qu'une simple mention du critère sans preuve (ce qui correspond à la cotation 1). En tenant compte conjointement de la cotation 2 et de la cotation 3, nous retrouvons les résultats suivants, résumés dans la figure 8 ci-après. Pour l'ensemble des 29 tests, celle-ci nous montre à la fois quels critères sont excellemment remplis (cotés 3), et ceux qui offrent des éléments intéressants sans être suffisants (cotés 2).

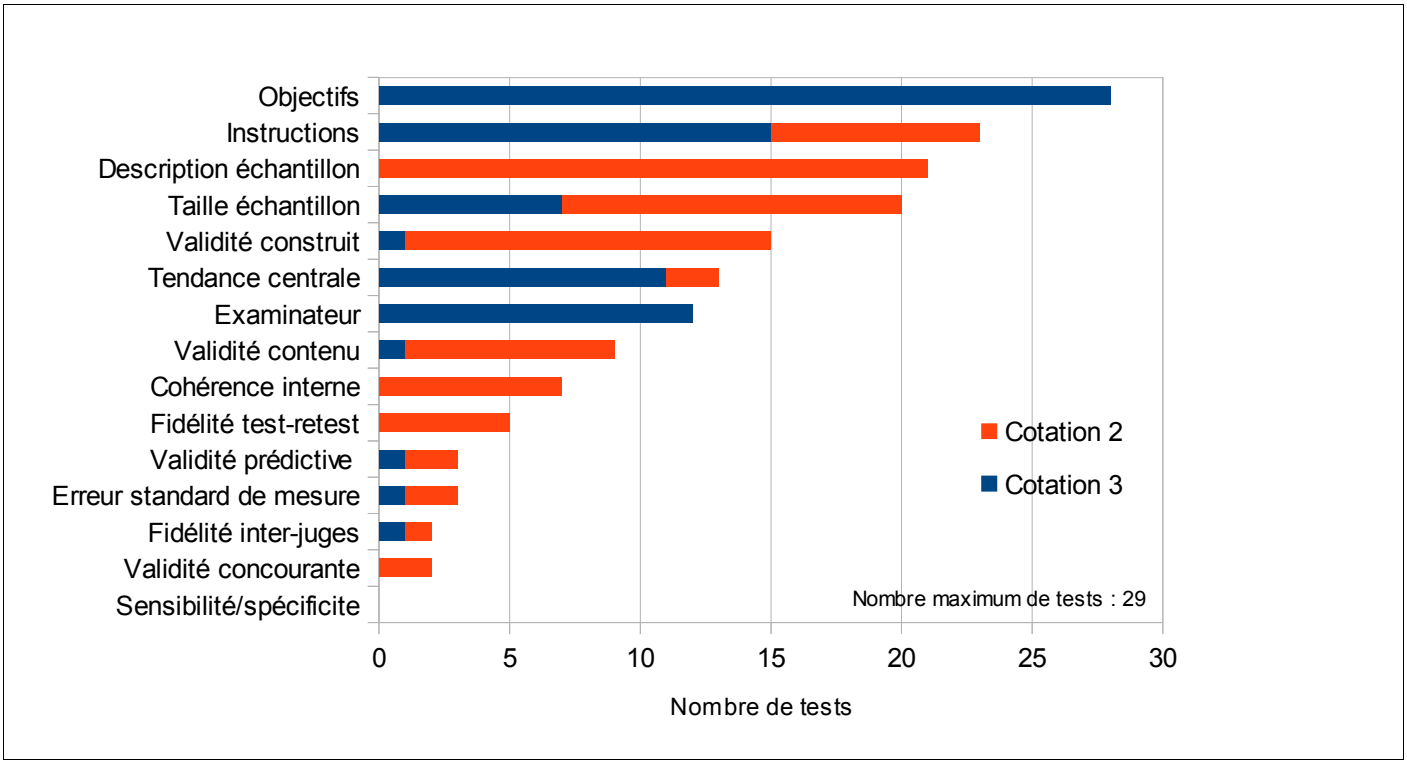

Figure 8 : Nombre de tests cotés 2 ou 3 pour chaque critère.

Cinq critères sont retrouvés et cotés à 2 ou 3 dans plus de la moitié des tests : l'indication des objectifs, les instructions de passation, de cotation et d'interprétation, la description et la taille de l'échantillon et la validité de construit. Les trois premiers d'entre eux le sont même dans 21 des tests sur 29 (soit environ 75 % des tests). Même si le niveau de preuve reste insuffisant, des éléments indiquant la présence de quatre autres critères sont

relevés pour 25 à 50 % des tests : il s'agit de la mesure de tendance centrale, des qualifications de l'examinateur, de la validité de contenu et de la cohérence interne. En revanche, cinq critères restent très rarement l'objet d'analyses statistiques dans les manuels, même en tenant compte d'un niveau chiffré de preuves insuffisant coté à 2 : la fidélité test-retest, l'erreur standard de mesure, la validité prédictive, la fidélité inter-juges et la validité concourante ont été analysées ou calculées dans moins de 25 % des tests étudiés. Enfin, la sensibilité et la spécificité ne font l'objet d'aucune analyse statistique pour les 29 tests, même si elles sont parfois évoquées sans preuve chiffrée.

Nous constatons donc un manque très net de preuves statistiques démontrant la validité, la fidélité et la sensibilité des outils d'évaluation du langage oral chez l'enfant. Soit les indications sont absentes des manuels, soit le niveau de preuves requis est insuffisant. Les critères de standardisation sont également perfectibles. Seuls les objectifs du test ainsi que les instructions de passation et de cotation sont présents et relativement corrects dans la plupart des manuels.

Discussion

Nous commenterons dans un premier temps nos résultats, puis en indiquerons les implications pratiques. Nous aborderons ensuite les limites de notre étude, ainsi que les perspectives de potentielles futures recherches.

1. Discussion des résultats

Nous discuterons de nos résultats concernant les tests de langage oral étudiés dans ce mémoire, puis commenterons ceux traitant des critères psychométriques. Nous comparerons nos résultats avec ceux d'autres recherches.

1.1. Discussion sur les tests

Nos résultats indiquent que le nombre médian de critères psychométriques mentionnés (cotés 1 ou plus) par test est de neuf sur quinze, ce qui est moyen. Le nombre médian de critères totalement satisfaits (cotés 3) par test est de deux sur quinze, ce qui est extrêmement faible. Notre analyse laisse supposer des lacunes concernant la qualité psychométrique des tests orthophoniques évaluant le langage oral de l'enfant : rappelons qu'aucun test ne satisfait complètement la moitié des critères. Nous trouvons fréquemment dans les manuels des affirmations de validité, de fidélité ou de sensibilité, mais malheureusement sans analyse statistique complète ni preuve chiffrée suffisante. Le plus souvent, ces informations sont absentes des manuels : 37 % de l'ensemble de nos cotes sont au niveau 0, ce qui est un pourcentage très important. Ces résultats sont en accord avec notre hypothèse initiale, qui supposait que peu d'outils d'évaluation francophones du langage oral de l'enfant répondaient aux standards psychométriques attendus d'un test.

Toutefois, ces résultats généraux cachent des variations importantes entre les tests. En effet, certains tests (ex. les épreuves de Maillart et al., l'épreuve de Henin) montrent des scores très faibles (avec un seul critère coté à 3) alors que d'autres (ex. EVIP, T.L.O.C.C.) disposent de côtes psychométriques plus élevées (avec six critères cotés à 3). Nous pouvons proposer les hypothèses suivantes pour expliquer ces différences. Certains tests sont très anciens (ex. épreuve de Henin) : les connaissances psychométriques ont en effet beaucoup progressé depuis. D'autres tests sont le résultat de recherches préliminaires, sans volonté d'établir des épreuves parfaitement normées, et sont en accès libre (ex. épreuves de Maillart et al.) : nous pouvons supposer le manque de moyens financiers affectés à ces travaux de recherche. D'autres tests sont en revanche plus récents ou ont fait l'objet de versions précédentes : la dernière version qui en résulte est probablement plus solide sur le plan psychométrique. Par exemple, l'EVIP est une adaptation d'un test anglophone et le fruit de longues années de recherche.

1.2. Discussion sur les critères psychométriques

Certains critères sont excellemment remplis par une majorité de tests : l'indication des objectifs et les instructions d'administration, de cotation et d'interprétation. Il s'agit des critères répondant à des impératifs cliniques, aidant le professionnel dans sa pratique quotidienne auprès des patients. Notons cependant que les consignes d'interprétation sont le plus souvent lacunaires. Parfois, les consignes d'administration laissent explicitement une liberté au testeur dans les consignes à formuler à l'enfant, selon son âge (ex. Kikou 3-8, PEES 3-8). Ce fait est heureusement assez rare et les consignes sont le plus souvent précises ; de plus, il arrive que les auteurs soulignent qu'il est impératif de les respecter pour garantir la validité des résultats (ex. EVIP).

Quant aux critères de standardisation, la comparaison entre les résultats résumés dans les figures 7 et 8 ci-dessus est très parlante concernant la description de l'échantillon. En effet, nous trouvons souvent dans les manuels les informations requises sur la population d'étalonnage, mais aucun élément probant de représentativité n'est apporté (ce qui explique une cotation à 2 pour vingt-et-un tests). L'échantillon est souvent dit « représentatif », sans aucune preuve. Seul un test, la N-EEL, apporte une preuve de représentativité de la distribution des CSP, en comparant la répartition de l'échantillon d'étalonnage à celle de la population française (les données de l'Insee figurent dans le manuel) ; mais l'échantillon de ce test n'est pas représentatif pour l'origine géographique (72 % des enfants habitent l'Ile de France). Nous comprenons qu'il n'est pas aisé sur un plan logistique d'administrer des épreuves à un échantillon d'enfants originaires de toutes les régions de France, mais un échantillon provenant de trois ou quatre régions ne peut garantir une bonne représentativité.

La mesure de tendance centrale serait quant à elle largement perfectible. En effet, quelques manuels n'apportent pas de preuve d'une distribution gaussienne (certains indiquent qu'elle ne l'est pas), mais les auteurs laissent au clinicien la libre utilisation des moyennes et des écarts-types. Heureusement, d'autres manuels précisent clairement d'utiliser les percentiles en raison d'une distribution non gaussienne, mais la mention du percentile 5 ou du percentile 3 reste très minoritaire (ex. les BILO, Kikou 3-8) ; seul le percentile 10 est indiqué assez fréquemment (ex. ELO).

Les critères apportant le moins de preuves suffisantes sont la sensibilité/spécificité, les critères de validité et de fidélité et la mention de l'erreur standard de mesure.

Concernant les critères de validité, nous pouvons émettre diverses hypothèses pour expliquer ces résultats. En premier lieu, la validité concourante (type de validité le moins étudié dans les manuels) ne peut s'établir que si l'on dispose de tests antérieurement validés qui évaluent les mêmes compétences. Or, les tests apportant des preuves de validité sont très rares en orthophonie. Il paraît donc logique que ce critère soit difficile à satisfaire.

Au sujet de la validité théorique et de construit, la mention d'une analyse factorielle est extrêmement rare : elle est retrouvée dans l'ECSP et l'ELOLA, mais en partie fragilisée par un nombre de sujets par sous-groupes insuffisant. Nous trouvons en revanche très souvent un facteur âge qui est dégagé à travers une analyse statistique montrant un effet significatif de l'âge sur les scores aux épreuves (ex. BILO 3C, EVIP) ; c'est une indication indispensable contribuant à la validité d'une épreuve où l'on compare les résultats d'un individu à ceux de sa tranche d'âge. Cependant, cela reste insuffisant pour juger du construit de l'épreuve, par exemple en extrayant d'autres facteurs expliquant des résultats proches entre deux subtests d'une batterie, ou en dégageant des variables associées entre elles. En revanche, des éléments d'appui théorique sont très souvent indiqués dans les manuels, et des modèles théoriques du fonctionnement cognitif sont parfois même longuement développés (ex. IFDC, Péléa). Cela ne suffit pas à dégager statistiquement des facteurs du construit de l'épreuve ou de la batterie, mais ces éléments sont tout à fait intéressants et restent indispensables.

La validité de contenu, quant à elle, s'appuie très souvent sur une simple mention qualitative du choix des items : les auteurs expliquent par exemple qu'ils ont choisi tels items en fonction de leur complexité syllabique ou de leur fréquence dans la langue (ex. Test de dépistage de dénomination et de désignation d'images, Kremin & Dellatolas, 1995). D'autre part, concernant une analyse quantitative, nous trouvons régulièrement en annexes du manuel une simple mention du pourcentage de réussite item par item. Certains auteurs ont utilisé ces pourcentages pour dégager un indice de difficulté et sélectionner ainsi les items les plus pertinents (l'indice de discrimination n'est en revanche pas utilisé dans les tests étudiés ici, sauf dans l'EVIP, qui est le seul test coté 3 dans notre étude pour la validité de contenu). Mais tous les auteurs n'utilisent malheureusement pas ces pourcentages de réussite pour sélectionner leurs items, alors que le travail d'analyse des items a été entamé. Dans ce cas, les pourcentages de réussite sont simplement utiles à l'utilisateur du test pour interpréter la réponse à un item. Enfin, notons que ce critère de validité est souvent insatisfaisant car le nombre d'items de certains subtests est inférieur à vingt (seuil en dessous duquel la validité de contenu d'une épreuve ne peut être garantie d'après la littérature).

Enfin, la validité prédictive reste un élément difficile à mesurer. En effet, comment peut-on établir la mesure des compétences en langage oral dans la vie quotidienne, prérequis indispensable pour établir la corrélation avec les résultats au test ? Les résultats scolaires sont une possibilité, mais ils sont également en lien avec les compétences en langage écrit. L'option de questionnaires aux adultes côtoyant l'enfant régulièrement (parents, professeurs) est envisageable, afin d'estimer son niveau de langage oral. Mais elle serait coûteuse en temps ; elle suppose également la validation de ces questionnaires et une formation de ces adultes à leur cotation pour que la mesure soit fiable, ce qui demande également du temps et des moyens.

Il en est de même pour le calcul des fidélités inter-juges et test-retest. Elles requièrent du temps et des moyens humains et financiers. En effet, l'une suppose de dédoubler la cotation pour chaque enfant de l'échantillon, avec deux examinateurs différents (à l'aide d'enregistrements vidéo par exemple, ce qui représente un certain budget). L'autre consisterait à faire passer deux fois les mêmes épreuves à chaque participant. Cet aspect peut paraître fastidieux à un enfant et rendre plus délicate la recherche de volontaires pour établir l'échantillon d'étalonnage.

Le calcul de l'erreur standard de mesure et de la cohérence interne est également peu fréquent dans les manuels. Nous pouvons suggérer l'hypothèse d'un manque de formation des créateurs de tests dans le domaine statistique pour expliquer ces faiblesses psychométriques.

Enfin, nous notons que les calculs des coefficients de validité ou de fidélité, avec certains résultats à première vue excellents (supérieurs au seuil attendu), sont parfois partiellement invalidés en raison d'un échantillon par sous-groupe inférieur à cent sujets. C'est le cas du Thapho, de l'ELOLA et de l'EVIP.

En dernier lieu, la sensibilité et la spécificité sont les critères les moins satisfaits. Beaucoup de manuels de tests mentionnent une bonne sensibilité de l'épreuve (ex. Exalang 3-8, N-EEL), mais cela n'est pas prouvé par le calcul d'un coefficient de sensibilité. La spécificité est quant à elle totalement absente des manuels. Les auteurs, tels ceux du Boehm-3 ou du PEES 3-8, s'appuient parfois sur les scores obtenus par un échantillon restreint (moins de cent sujets) de cas d'enfants diagnostiqués dysphasiques ou porteurs de trouble développemental du langage oral (le nombre de cas pathologiques est parfois inconnu). Ils comparent ces résultats avec la population normale de leur échantillon, en mentionnant simplement des résultats inférieurs. Sur cette base, les auteurs en déduisent souvent que l'épreuve est sensible. Or, comme nous l'avons vu, le calcul d'un taux de sensibilité s'avère plus complexe et un taux chiffré significatif de sensibilité doit apparaître dans le manuel.

1.3. Mises en perspective avec les résultats d'autres recherches

Notre étude corrobore les résultats retrouvés dans la littérature au sujet des qualités psychométriques des tests francophones évaluant le langage oral (Bouchard et al., 2009 ; Leclercq et Veys, 2014). Elle est également en accord avec les résultats des recherches menées sur les tests anglophones (McCauley & Swischer, 1984). Celles-ci montraient une absence de preuves suffisantes de validité, de sensibilité et de fidélité pour la majorité des tests étudiés. La majorité des informations relatives aux qualités psychométriques étaient là aussi le plus souvent absentes des manuels.

2. Implications pratiques

Cette recherche nous amène à nous questionner sur l'utilisation des tests en orthophonie et les diagnostics qui en découlent. Nous nous appuyons en effet sur les scores aux épreuves pour identifier un trouble, poser un diagnostic, construire notre projet thérapeutique et évaluer les progrès de l'enfant lors du bilan de renouvellement. Les chercheurs se fondent également sur ces résultats aux tests de langage pour sélectionner les participants d'une étude. Or, notre

analyse suggère que la validité et la fiabilité de ces scores peuvent être remises en question. Ce constat confirme l'évidence que les résultats chiffrés à une épreuve ne suffisent pas pour diagnostiquer un trouble chez un patient. L'analyse clinique, l'interprétation des résultats, l'expérience acquise, l'observation des aptitudes et des difficultés de l'enfant hors des situations de tests, les informations apportées par les parents lors de l'anamnèse participent au diagnostic et sont d'autant plus indispensables. La qualité psychométrique médiocre des tests peut aussi expliquer un écart important que l'on peut parfois constater chez le même enfant entre deux résultats à des épreuves aux objectifs similaires. Si un score nous paraît étrange en regard de ce que l'on observe par ailleurs chez l'enfant, il est suggéré de conserver un esprit critique vis-à-vis de ce résultat.

En dépit de cet apparent manque de solidité psychométrique, la plupart des outils orthophoniques possèdent d'autres qualités, parmi lesquelles, pour certains tests, la facilité d'administration, la rapidité de passation et de cotation ou leur aspect attrayant. Ainsi, un nombre d'items très élevé, pourtant gage d'une bonne validité de contenu, peut rendre la passation longue et fastidieuse pour un enfant, surtout si celui-ci est fatigable ou présente des difficultés attentionnelles. Lui proposer un test trop long pourrait donc nuire à une bonne évaluation de ses compétences : l'enfant risquerait par exemple de répondre au hasard pour aller plus vite, ou de ne pas vouloir terminer l'épreuve. Certains tests bons d'un point de vue psychométrique comprennent des subtests complexes à administrer pour le clinicien (ex. subtests de morphosyntaxe du T.L.O.C.C.) : l'épreuve peut se trouver biaisée si le clinicien est hésitant dans sa manière de conduire la passation. En tant que clinicien, nous devons opérer des choix au cas par cas, entre une épreuve courte et facile d'utilisation clinique, mais dont les résultats chiffrés sont à interpréter avec prudence, et une épreuve plus longue, donc a priori plus solide statistiquement, mais moins aisée à administrer. Gardons à l'esprit qu'un test statistiquement peu valide peut nous fournir des indications qualitatives malgré tout très intéressantes, si l'on prend le temps nécessaire à son analyse et son interprétation. De plus, certains tests sont rares pour évaluer une compétence précise : ainsi, le Péléa est l'un des seuls tests visant l'évaluation du langage élaboré de l'adolescent. D'autres tests ont un aspect ludique et attrayant (ex. les épreuves informatisées des Exalang) : leur utilisation s'avère intéressante auprès d'enfants présentant des difficultés attentionnelles ou des traits autistiques. Malgré les lacunes psychométriques de certains de ces tests, ils ont tout de même le mérite d'exister et d'être utiles aux cliniciens. Nos choix d'outils d'évaluation sont donc le fruit d'une réflexion tenant compte de plusieurs critères, dont leurs qualités psychométriques.

3. Limites de l'étude

La première concerne le choix des tests étudiés. Le manque de listes récentes établies en France (la liste de l'UNADREO datant de 2011) nous a conduit à nous appuyer en premier lieu sur celle établie par l'INAMI, réactualisée quant à elle chaque année. Il s'agit cependant d'une liste correspondant aux pratiques belges et non spécifiquement françaises : certes, les pratiques belges et françaises sont proches, mais elles ne sont pas tout à fait identiques. Notre procédure de sélection des tests aurait pu provenir des pratiques des orthophonistes français, par exemple en s'appuyant au préalable sur des questionnaires adressés aux praticiens pour connaître les tests les plus usités. Ainsi, certains tests de notre liste sont peu utilisés en France

(ex. les BILO). La liste de tests de l'INAMI correspond dans la pratique aux tests recommandés par l'assurance santé belge, dont l'usage est remboursé. Cette pratique n'existe pas en France, ce qui peut expliquer l'absence de listes de tests orthophoniques récentes à utiliser.

Le choix des critères psychométriques, sur lesquels nous avons bâti notre analyse, s'est appuyé sur une revue de littérature traitant du sujet. Nous avons fait le choix de regrouper certains critères dénommés différemment par les auteurs, en raison de leur recoupement ou de leurs points communs dans les définitions que les auteurs en donnent (ex. la validité théorique et de construit). Notre liste de critères est le fruit d'une méthodologie se voulant rigoureuse (critères présents dans au moins trois études de référence), mais le choix de conserver certains critères moins retrouvés dans la littérature reste personnel. Leur utilisation met cependant en lumière le manque criant de données dans les manuels de tests sur ces aspects, notamment pour la sensibilité. Tous ces choix ont fait l'objet de longues réflexions, qui nous ont conduits à notre liste finale de critères.

Dans certains cas, il a parfois été délicat de coter une batterie entière composée d'épreuves inégales en qualité. Considérons par exemple le nombre d'items par subtest. Nous avons adopté un choix restrictif de ne pas faire de moyennes du nombre d'items par subtest, mais de considérer que le nombre d'items n'était pas suffisant si un des subtests comportait moins de vingt items. Ce choix reste discutable.

Par ailleurs, la procédure d'addition des critères dans nos résultats globaux peut paraître simpliste. Cela permet de rendre l'étude plus lisible et de synthétiser les résultats, mais nos recherches théoriques ont montré que tous les critères ne sont pas équivalents : certains critères sont plus importants, en premier lieu les critères de validité.

Enfin, la méthodologie de cotation a été souhaitée la plus rigoureuse possible, afin d'évaluer chaque test de la même manière. Cependant, coter tous les tests de notre liste avec la même rigueur est discutable. En effet, certains tests sont issus d'articles de recherche et sont en accès libre, comme l'ELDP. Nous pouvons supposer qu'ils ne disposent pas des mêmes moyens financiers que les maisons d'édition proposant des tests payants. Dans certains de ces articles, il est clairement précisé que l'étude des données n'est pas complète. Parfois, la publication de la totalité des données a été annoncée, mais aucun autre article n'est paru par la suite : c'est par exemple le cas du DLPF, dont l'article de référence précise clairement qu'il ne s'agit que d'une version préliminaire du protocole. Les auteurs, tels Maillart et ses collègues dans l'élaboration de l'épreuve « Décision lexicale », précisent souvent dans ce cas que le test découlant de leur recherche reste qualitatif : il convient d'interpréter les résultats obtenus avec prudence et ceux-ci ne sont qu'un complément à d'autres évaluations normées. Notre cotation peut donc paraître sévère quand le but n'était pas d'établir un test parfaitement normé et validé.

4. Pistes de recherches et perspectives

L'élaboration et la construction de tests orthophoniques solides sur un plan psychométrique reste à mener. Concernant les tests dont la taille de l'échantillon est déjà suffisante (cent sujets par sous-groupe), comme DEDALE, les épreuves BILO ou le TLOCC,

des analyses de validité et de fidélité restent possibles a posteriori. Les tests où la taille de l'échantillon est insuffisante paraissent plus complexes à améliorer d'un point de vue psychométrique : les possibilités d'établir des preuves de validité et de fidélité sont conditionnées par le ré-étalonnage des épreuves auprès d'un échantillon suffisant.

La formation en psychométrie des orthophonistes (et des futurs concepteurs de tests) semble indispensable afin de comprendre un manuel, de maîtriser l'utilisation et l'interprétation des résultats. Les statistiques font partie de notre programme dans nos études d'orthophonie. Nous pouvons espérer que la mise en place du master favorisera l'amélioration des compétences de la profession en la matière. Nous pouvons également émettre l'hypothèse que la hausse du niveau des cliniciens dans ce domaine incitera les éditeurs de tests à publier des outils plus solides sur le plan psychométrique.

Il serait par ailleurs intéressant de mener une étude sur le lien entre la fréquence d'utilisation des tests par les orthophonistes et les qualités psychométriques des épreuves. Globalement, quels sont les facteurs influençant les orthophonistes dans leur choix de tests ? Les qualités psychométriques en font-elles partie, et si oui, lesquelles ? Quels sont les autres critères retenus par les cliniciens dans leur décision d'utilisation des outils (ex. aspect ludique, facilité d'administration, rapidité de cotation) ? Cette problématique pourrait faire l'objet de futures recherches.

Conclusion

L'objectif de notre étude était d'analyser les qualités psychométriques des tests et batteries d'épreuves francophones utilisés dans les bilans de langage oral de l'enfant. Vingt-neuf tests et batteries ont été sélectionnés, en nous fondant sur deux listes de référence, l'une belge, l'autre française. Quinze critères ont été utilisés pour mener cette analyse : il s'agissait de critères de validité, de fidélité, de standardisation et de sensibilité. La méthode d'évaluation de chaque critère s'est fondée sur une cotation en quatre niveaux : du moins satisfaisant au plus satisfaisant.

Les résultats de notre recherche montrent qu'une large majorité de tests orthophoniques de langage oral ne satisfont pas les critères psychométriques attendus d'un outil d'évaluation. Aucun des 29 tests ne remplit de façon complètement satisfaisante la moitié des critères ; la moitié de ces tests indiquent des niveaux de preuve parfaitement suffisants pour seulement deux critères psychométriques ou moins sur quinze. L'indication claire des objectifs du test est le seul critère présent dans la quasi-totalité des manuels ; les consignes de passation, de cotation et d'interprétation sont complètes pour la moitié des épreuves. Tous les autres critères sont entièrement satisfaits pour moins de la moitié des tests. Les manuels de la plupart des tests de langage oral présentent ainsi des lacunes au sujet des preuves de validité, de fidélité, de standardisation et de sensibilité des épreuves.

Ces résultats nous questionnent sur l'utilisation clinique de nos outils d'évaluation, sur lesquels nous nous appuyons en partie pour diagnostiquer la présence ou l'absence d'un trouble de langage. Notre formation initiale nous encourage vivement à ne pas nous fonder uniquement sur les scores quantitatifs d'une série d'épreuves lors de nos bilans. L'évaluation qualitative du patient, à travers une anamnèse approfondie, une comparaison avec le langage spontané et l'attitude de l'enfant et une analyse qualitative des résultats aux épreuves, est

indispensable. Notre étude confirme cette évidence. Les tests que nous utilisons au quotidien possèdent certes de nombreuses qualités cliniques, mais nous devons les utiliser de manière éclairée, en toute conscience de leurs lacunes psychométriques, ce qui rend d'autant plus nécessaire la qualité de notre expertise clinique.

La conception de futures épreuves orthophoniques plus solides d'un point de vue psychométrique, ainsi que l'amélioration potentielle d'outils déjà existants, constituent un champ de recherches à explorer. Des études complémentaires de validité, de fidélité et de sensibilité pourraient être conduites sur certains tests dont les qualités psychométriques sont déjà relativement satisfaisantes. L'orthophonie est une profession jeune, qui a beaucoup évolué depuis sa naissance. Le souci de la qualité psychométrique de nos tests devient depuis quelques années un réel sujet de recherche appliquée, ce qui laisse présager une amélioration de nos outils. Nous pouvons également espérer que la réforme des études permettra une meilleure connaissance du domaine de la psychométrie chez les futurs orthophonistes et chez les futurs concepteurs de tests, et que les éditeurs répondront à cette exigence croissante de qualité de la part de la profession.

Il serait également intéressant de savoir si la connaissance des qualités psychométriques des tests de langage influence les orthophonistes actuellement en exercice en France : sur quels critères choisissent-ils d'utiliser tel ou tel test ? Les preuves de qualité psychométrique d'une épreuve font-ils partie de leurs critères de choix ? Ces questionnements pourraient initier de futurs travaux de recherche.

Bibliographie

Bassano, D., Labrell, F., Champaud, C., Lemétayer, F & Bonnet, P. (2005). Le DLPF : Un nouvel outil pour l'évaluation du développement du langage de production en français. *Enfance, n°2/2005*, 171-208.

Beech, J.R. & Harding, L. (1994). *Tests, mode d'emploi... Guide de psychométrie.* Traduction et adaptation française sous la direction de J.-P. Rolland & J.-L. Mogenet. Paris, France : Editions du Centre de Psychologie Appliquée.

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

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, Vol. 44*, 133-146.

Bishop, D.V.M. (1998). Development of the Children's Communication Checklist (CCC) : a method for assessing qualitative aspects of communicative impairment in children. *Journal of Child Psychology and Psychiatry, 39*, 6, 879-891.

Bishop, D.V.M. & Baird, G. (2001). Parent and teacher report of pragmatic aspects of communication : use of the Children's Communication Checklist in a clinical setting. *Developmental Medicine and Child Neurology, 43*, 809-818.

Boehm, A.-E. (2009). *Boehm-3 : Test des concepts de base – 3e édition.* Paris, France : ECPA.

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, Vol. 33*, 129-139.

Boutard, C. & Bouchet, M. (2009). *PEES 3-8 : protocole d'évaluation de l'expression syntaxique*. Isbergues, France : Ortho Edition.

Boutard, C. &, Bouchet M. (2009). *KIKOU 3-8 : protocole d'évaluation de la compréhension syntaxique et narrative*. Isbergues, France : Ortho Edition.

Chevrie-Muller, C., Simon, A.-M., Le Normand, M.-T. & Fournier, S. (1997). *Batterie d'évaluation psycholinguistique (BEPL-A)*. Paris, France : ECPA.

Chevrie-Muller, C., Simon, A.-M., Le Normand, M.-T. & Fournier, S. (1997). *Batterie d'évaluation psycholinguistique (BEPL-B)*. Paris, France : ECPA.

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

Colli-Vaast, L. (2018). *Caractéristiques psychométriques des tests de langage écrit chez l'enfant* (mémoire de master). Université de Lille, Lille.

Coquet, F., Ferrand, P. & Roustit, J. (2009). *EVALO 2-6: évaluation du développement du langage oral chez l'enfant*. Isbergues, France : Ortho Edition.

De Agostini, M., Metz-Lutz, M.-N., Van Hout, A., Chavance, M., Deloche, G., Pavao-Martins, I. & Dellatolas, G. (1998). Batterie d'évaluation du langage oral de l'enfant aphasique (ELOLA) : standardisation française (4-12 ans). *Revue de Neuropsychologie, Vol. 8, n°3,* 319-367.

Décret n° 2002-721 du 2 mai 2002 relatif aux actes professionnels et à l'exercice de la profession d'orthophoniste (2002). *JORF n°104 du 4 mai 2002 page 8339 texte n° 56.*

Deltour, J.-J. & Hupkens, D. (2011). *DEDALE : dépistage expérimental des difficultés d'apprentissage de la lecture et de l'écriture.* Paris, France : Eurotests Editions.

Dunn, L., Theriault-Whalen, C. & Dunn, L. (1993). *EVIP : échelle de vocabulaire en images Peabody*. Toronto, Canada : Psycan.

Ecalle, J. (2007). *ThaPho : test d'habiletés phonologiques pour enfants de 5 à 8 ans*. Paris, France : Mot à Mot Editions.

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, Vol. 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.

Grégoire, J. (2009). *L'examen clinique de l'intelligence de l'enfant. Fondements et pratiques du WISC IV.* Bruxelles, Belgique : Mardaga.

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

Guidetti, M. & Tourrette, C. (2009). *ECSP : échelle d'évaluation de la communication sociale précoce.* Paris, France : Eurotests Editions.

Guillon, A., Boutard, C. & Charlois, A.-L. (2010). *PELEA : protocole d'évaluation du langage élaboré de l'adolescent*. Isbergues, France : Ortho Edition.

Helloin, M.-C. & Thibault, M.-P. (2006). *EXALang 3-6 : batterie d'examen des fonctions langagières chez l'enfant de 3 à 6 ans.* Grenade, France : Orthomotus.

Hénin, N. (1981). Étude de la motricité et des praxies oro-faciales chez l'enfant de 2 ans et demi à 12 ans et demi. *Bulletin d'Audiophonologie*, *12*, 17–42.

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

Huteau, M. & Lautrey, J. (1997). *Les tests d'intelligence.* Paris, France : La Découverte.

Huteau, M. & Lautrey, J. (1999). *Evaluer l'intelligence. Psychométrie cognitive.* Paris, France : Presses Universitaires de France.

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

Institut National d'Assurance Maladie-Invalidité (2017). *Liste limitative des tests pour l'évaluation du langage oral et de la dysphasie.* Repéré le 4 janvier 2018 sur http://www.riziv.fgov.be/fr/professionnels/sante/logopedes/Pages/logopedes-liste-limitative-tests.aspx#.WPuRvWekKM9

Institut National de la Santé Et de la Recherche Médicale. (2007). *Dyslexie Dysorthographie Dyscalculie : Bilan des données scientifiques.* Paris, France : éditions Inserm.

Kern, S. & Gayraud, F. (2010). *IFDC : inventaire français du développement communicatif.* Grenoble, France : Editions de la Cigale.

Kerr, M.A., Guildford, S. & Bird E.K-R. (2003). Standardized language test use: A Canadian survey. *Journal of Speech-Language Pathology and Audiology, Vol. 27(1)*, 10-28.

Khomsi, A. (2001). *ELO : évaluation du langage oral*. Paris, France : ECPA.

Khomsi, A., Khomsi, J. & Pasquet, F. (2007). *BILO 2 : bilan informatisé du langage oral pour le cycle 2.* Paris, France : ECPA.

Khomsi, A., Khomsi, J., Pasquet, F. & Parbeau-Guéno, A. (2007). *BILO 3C : bilan informatisé du langage oral pour le cycle 3 et le collège*. Paris, France : ECPA.

Khomsi, A. & Khomsi, J. (2007). *BILO EC 2 : évaluation des contraintes pour le cycle 2.* Paris, France : ECPA.

Khomsi, A. & Khomsi, J. (2009). *BILO Petits : bilan informatisé du langage oral pour le cycle 1.* Paris, France : ECPA.

Kremin, H. & Dellatolas, G. (1995). L'accès au lexique : une étude de standardisation chez l'enfant d'âge pré-scolaire. *Revue de Neuropsychologie, Vol. 5, n°3,* 309-338.

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.

Macchi, L., Descours, C., Girard, E., Guitton, E., Morel, C., Timmermans, N. & Boidein, F. (2012). *ELDP : épreuve lilloise de discrimination phonologique destinée aux enfants de 5 ans à 11 ; 6 ans*. Université de Lille, Lille. Repéré le 20 janvier 2018 sur http://orthophonie.univ-lille2.fr/stocks/stock-contents/epreuve-lilloise-de-discrimination-phonologique.html

Maillart, C. (2003). Les troubles pragmatiques chez les enfants présentant des difficultés langagières. Présentation d'une grille d'évaluation : la *Children's Communication Checklist* (Bishop, 1998). *Les Cahiers de la SBLU, 13*, 13-32.

Maillart, C. & Schelstraete, M.-A. (2004). *L'évaluation des troubles phonologiques : illustration de la démarche diagnostique.* Université catholique de Louvain, Louvain-la-Neuve.

Maurin, N. (2006). *T.L.O.C.C. : test de langage oral complexe pour collégiens*. Isbergues, France : Ortho Edition.

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

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, n°273*, 47-70.

Pichot, P. (1999). *Les tests mentaux*. Paris, France : Presses Universitaires de France.

Piérart, B., Comblain, A., Grégoire, J. & Mousty, P. (2010). *Batterie ISADYLE*. Marseille, France : Solal Editeur.

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

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

Société Française de Psychologie (2003), Recommandations internationales sur l'utilisation des tests : Version 2000. *Pratiques Psychologiques, Numéro spécial hors-série.*

Union Nationale pour le Développement de la Recherche et de l'Evaluation en Orthophonie (2011). *Tests et modules*. *Annexes.* Repéré le 4 janvier 2018 sur http://www.unadreo.org/articles/getArticle/59/353

Annexes

Annexe n°1 : Liste des tests ayant fait l'objet de mesures de fiabilités intra et inter-examinatrices.

Mesure de la fiabilité intra-examinatrice :

- BILO Petits (Khomsi & Khomsi, 2009),
- Children's Communication Checklist – C.C.C. (Bishop, 1998 ; Bishop & Baird, 2001; Maillart, 2003),
- EVIP (Dunn et al., 1993),
- EXALang 3-6 (Helloin & Thibault, 2006),
- N-EEL (Chevrie-Muller & Plaza, 2001).

Mesure de la fiabilité inter-examinatrices :

- ELO (Khomsi, 2001),
- Kikou 3-8 (Boutard & Bouchet, 2009a),
- ThaPho – Test d'habiletés phonologiques (Ecalle, 2007).

Annexe n°2 : Tableau 4 : Cotation des critères psychométriques par test.

Tests	Validité contenu	Validité construit	Validité concourante	Validité prédictive	Fidélité test-retest	Fidélité inter-juges	Cohérence interne	Objectifs	Instructions	Examinateur	Erreur standard de mesure	Description échantillon	Taille échantillon	Tendance centrale	Sensibilité/spécificité	Score (sur 45)
BEPL-A	0	1	0	0	0	0	0	3	2	3	0	2	2	1	0	14
BEPL-B	0	1	0	0	0	0	0	3	2	3	0	2	2	1	0	14
BILO-Petits	1	1	0	0	0	0	0	3	3	1	0	2	3	3	0	17
BILO-2	1	2	0	0	0	0	0	3	3	1	0	2	3	3	0	18
BILO EC-2	1	1	0	0	0	0	0	3	3	1	0	2	3	3	0	17
BILO-3C	1	2	0	0	0	0	0	3	3	1	0	2	2	3	0	17
Boehm-3	1	1	1	2	2	0	2	3	3	3	2	2	2	3	1	28
C.C.C.	0	1	1	0	0	2	2	3	3	1	0	1	1	1	1	17
Décision lexicale (Maillart)	1	2	0	0	0	0	0	3	1	1	0	2	1	1	1	13
DEDALE	2	1	0	0	0	0	0	3	3	3	0	1	3	1	1	18
Discrimination (Maillart)	1	2	0	0	0	0	0	3	1	1	0	2	1	1	1	13
DLPF	2	1	0	0	0	0	0	3	1	3	0	0	1	1	0	12
ECSP	1	2	2	0	0	0	0	3	3	1	0	1	1	1	1	16
ELDP	1	1	0	0	0	0	0	3	2	3	0	2	1	3	0	16
ELO	1	1	0	0	0	0	0	3	3	0	0	1	2	2	1	14
ELOLA	2	2	0	0	0	0	0	3	2	0	0	1	1	2	0	13
EVALO 2-6	2	2	0	1	0	1	2	3	3	3	0	2	2	1	1	23
EVIP	3	2	2	1	2	0	2	3	3	3	3	2	2	3	0	31
Exalang 3-6	1	1	1	1	2	0	0	3	2	3	0	2	2	1	1	20
Hénin – motricité praxies	2	0	0	0	1	0	1	3	2	0	0	0	2	1	0	12
IFDC (F)	2	2	1	1	1	0	1	3	2	3	0	2	1	3	0	22
ISADYLE	2	2	0	0	0	0	2	3	3	3	2	1	2	1	0	21
Kikou 3-8	1	1	1	0	0	0	0	3	1	3	0	2	2	3	1	18
N-EEL	1	1	1	0	2	0	0	3	3	1	0	2	3	1	1	19
PEES 3-8	1	2	1	0	0	0	0	3	1	1	0	2	2	3	1	17
Péléa 11-18	1	2	0	0	0	0	0	3	1	1	0	2	1	1	1	13
Test dénomination désignation (Kremin)	1	2	0	0	0	3	0	1	2	0	0	2	3	1	0	15
ThaPho	1	2	1	2	2	0	2	3	3	1	0	2	2	1	1	23
T.L.O.C.C.	2	3	0	3	0	0	2	3	3	1	0	2	3	3	1	26

Annexe n°3 : Etude de l'ELO (Khomsi, 2001).

Critères	Pages	Éléments présents dans le manuel
Validité de contenu / analyse des items	pp.5 à 14 pp.45 à 71	Nombre d'items : LexR : 20 – LexP : 50 noms et 10 verbes – RépM : 32 – C1 : 20 – C2 : 21 ou 32 - ProdE : 16 ou 25 – Répétition Enoncés : 15. Analyse qualitative « rapide » de certains items (entre 2 et 8 items selon les épreuves) selon la fréquence en % des réponses en fonction du groupe d'âge + analyse des erreurs et stratégies de l'enfant. Résultats item par item en annexes : distribution des réponses en % → pas de suppression / modification d'items selon ces pourcentages.
Validité théorique / de construit	p.2 pp.17 à 21	Théories dominantes de type « défectologique »(décrire les dysfonctionnements selon ce que les enfants ne savent pas faire) et « catégoriel » (dysfonctionnements purs vs retards simples). Théories citées mais non détaillées. Analyse ANOVA pour déterminer si la différence des résultats est significative entre tous les groupes d'âge → oui pour toutes les épreuves jusqu'au CE2, mais non à partir du CM1.
Validité concourante		Pas d'information.
Validité prédictive		Pas d'information.
Fidélité test-retest		Pas d'information.
Fidélité inter-juges		Pas d'information.
Cohérence interne		Pas d'information.
Objectifs du test	p. 1, p.39 p.5 à 17	Décrire divers aspects du fonctionnement langagier oral à partir de 3 ans en réception et production, du mot à l'énoncé. Objectifs des six épreuves détaillés : LexR+LexP+RépM → analyse des modalités de traitement du mot oral ; C1/C2 : compréhension syntaxique ; RépE/ProdE : production linguistique d'énoncés.
Instructions d'administration / cotation / interprétation	pp.39 à 44 pp.3- 6-7 + feuille de passation pp.5 à 17 pp.25 à 36	Ordre des épreuves à respecter. Consignes orales explicitées, détaillées et claires. Indication des niveaux d'arrêt sur le protocole. Cotation : indications sur les réponses acceptées/refusées (parti pris restrictif) Etalonnage au 1er trimestre en 2000 → comparaison avec classe inférieure si nécessaire. Coter selon niveau d'arrêt. Interprétation : « Indispensable » d'analyser les erreurs. Proposition intéressante d'analyse des stratégies de l'enfant pour certains items. Typologie des erreurs pour C2 et ProdE. Ex. d'analyse de 5 cas cliniques.
Qualifications de l'examinateur		Pas d'information.
Erreur standard de mesure		Pas d'information.
Description de l'échantillon d'étalonnage	p.3	Etalonnage dans des écoles publiques du Loiret dotées d'un Réseau d'aide (proportion d'enfants à risque ou en difficulté « devrait être relativement plus importante qu'ailleurs »). Informations sur l'âge. Rien sur sexe et CSP. Pas l'objectif d'un échantillon représentatif.
Taille de l'échantillon	p.3	PSM : 68 ; MSM : 191 ; GSM : 158 ; CP : 129 ; CE1 : 103 ; CE2 :140 ; CM1 : 86 ; CM2 : 95 → 3 groupes sur 8 < 100 sujets
Tendance centrale	p.25 (note bas de page) pp.73 à 75	Indication du caractère non normal de la distribution → choix des centiles préconisé. Centilages en annexes (jusqu'au p10)
Sensibilité / Spécificité	p.9	Supposition de sensibilité pour RépM, sans preuve apportée

Annexe n°4 : Etude de l'EVIP (Dunn et al., 1993).

Critères	Pages	Éléments présents dans le manuel
Validité de contenu / analyse des items	p.52 p.54, p.60-61	Sélection des items du PPVT-R → Représentativité des items : preuve de validité de contenu. 340 mots dans l'EVIP : échantillon représentatif du vocabulaire français. Items nombreux : deux formes de 170 items. Analyse d'items traditionnelle + analyse Rasch Wright → % de réussite sur chaque item à chaque niveau d'âge. La très grande majorité des des items discriminent entre .40 et .60 ; suppression de 5 items.
Validité théorique / de construit	p.73 p.51	Validité « interne » calculée avec indice de Claparède pour savoir si effet de l'âge est significatif sur le vocabulaire → échelle valide jusque 13-14 ans. Appui théorique : historique du PPVT-R, avec études s'y rapportant.
Validité concourante	p.74	Corrélation élevée du PPVT avec d'autres tests de vocabulaire (subtest Stanford-Binet, échelles de Weschler, …) : médiane à .71. Corrélation avec tests de même format à .86
Validité prédictive	p.70	Test de dépistage pour le rendement scolaire. Vocabulaire = meilleur genre de test pour prédire le rendement scolaire (Dale et Reichert, 1957) → « Validité hypothéco-déductive » supposée sans élément de preuve.
Fidélité test-retest	p.41	Corrélation calculée pour chaque groupe d'âge grâce à l'application des 2 formes A et B pour 90% des sujets (délai = 1 semaine) → médiane du coefficient de corrélation à .72 (étendue de .55 à .78).
Fidélité inter-juges		Pas d'information.
Cohérence interne	p.39	Calcul de fidélité pair-impair pour chaque groupe d'âge, avec méthode Spearman-Brown. Forme A : médiane à .81 (étendue de .68 à .88). Forme B : médiane à .80 (étendue de .66 à .85)
Objectifs du test	p. XII	Deux objectifs : 1) mesure du lexique en réception, de l'étendue du vocabulaire français acquis ; 2) test de dépistage d'aptitude scolaire. Cible : 2ans 6 mois à 18 ans.
Instructions d'administration / cotation / interprétation	pp.8 à 15, 16 à 24, 26 à 32, 44, 48.	Consignes très détaillées : matériel, disposition du sujet et de l'examinateur (avec dessin). Consignes orales explicitées (notamment possibilité de répéter et prise en compte de l'auto-correction) et attitude de l'examinateur. Auteurs qui insistent sur la nécessité de respecter les consignes pour obtenir des scores valides. Explications sur la cotation + exemples. Calcul de l'âge du sujet, choix du point de départ (item base) et critères d'arrêt (item plafond), et calcul du score brut. 3 exercices pour apprendre à coter. Interprétation : ex. de 6 cas-types
Qualifications de l'examinateur	p.XII, p.6	Cliniciens, psychologues scolaires, psychopédagogues, orthophonistes, travailleurs sociaux, enseignants, psychiatres, linguistes, conseillers d'orientation.
Erreur standard de mesure	p.42, 44, 48	Seuil de confiance à 68%. ETM de 8 unités en score standardisé → calcul de l'intervalle de fiabilité des scores dérivés. Calcul de l'IC pour chaque score directement sur la feuille de réponse.
Description de l'échantillon d'étalonnage	p.55, p.33	4 régions du Canada. Langue maternelle française. Même nombre de filles et de garçons. Echantillon « représentatif » de 6 à 14 ans (sans preuve). Rien sur CSP.
Taille de l'échantillon	p.55, 63, 64	Forme A : 2038 sujets ; étendue de 45 à 126 sujets par groupe d'âge. Forme B : 1993 sujets ; étendue de 47 à 128 sujets par groupe d'âge. 20 groupes d'âge en tout (pour chaque forme) : 12 groupes > 100 sujets ; 7 groupes entre 50 et 100 sujets ; 1 groupe < 50 sujets
Tendance centrale	pp.36, 67,79 à 128	Scores distribués normalement selon courbe de Gauss → utilisation possible des ET ou des percentiles. Tableau des scores normalisés par groupe d'âge en annexes
Sensibilité/Spécificité		Pas d'information.