

## Validation et normalisation d'une épreuve lexicale chez des enfants de 2e et 5e primaires

**Auteur :** Renzonnet, Hortense

**Promoteur(s) :** Leclercq, Anne-Lise; Comblain, Annick

**Faculté :** Faculté de Psychologie, Logopédie et Sciences de l'Éducation

**Diplôme :** Master en logopédie, à finalité spécialisée

**Année académique :** 2024-2025

**URI/URL :** <http://hdl.handle.net/2268.2/22560>

---

### Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

---

# Validation et normalisation d'une épreuve lexicale chez des enfants de 2e et 5e primaires

BAIWIR EMILIE ET RENSONNET HORTENSE

*PROMOTRICES : MADAME ANNE-LISE LECLERCQ, MADAME  
ANNICK COMBLAIN*

*SUPERVISEUSE : MADAME PAULINE MENJOT*

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION DU GRADE DE MASTER EN  
LOGOPÉDIE

## Remerciements

*Hortense :*

*Je tiens à remercier chaleureusement toutes les personnes qui m'ont soutenue lors de ce travail. Je remercie en particulier Madame Leclercq, Madame Comblain, nos promotrices, et Madame Menjot, notre superviseuse, pour leurs accompagnements, leurs disponibilités et leurs conseils lors de ces deux années.*

*Je remercie, ensuite, Madame Martinez-Perez et Madame Moyse d'avoir consacré du temps pour la lecture de ce travail.*

*Je tiens à remercier tout le corps enseignant du master en logopédie de l'Université de Liège pour les cours qu'ils nous ont donnés.*

*Mes remerciements vont également à mon binôme, Émilie Baiwir. Notre entraide et nos encouragements mutuels nous ont permis de mener à bien ce mémoire.*

*Je remercie également tous les enfants et familles qui ont accepté de participer à notre étude. Sans eux, rien n'aurait été possible.*

*Enfin, je tiens à remercier mes proches qui ont été d'un soutien précieux lors de ces années de master et lors de ce travail final.*

## Remerciements

*Emilie :*

*Avant de clore ce mémoire, je prends un moment pour remercier toutes les personnes qui nous ont soutenues tout au long de ce parcours.*

*Je vais tout d'abord remercier Madame Leclercq et Madame Comblain, nos promotrices, ainsi que Madame Menjot, notre superviseuse pour leurs conseils avisés, leur expertise et leurs retours constructifs, qui ont été d'une grande richesse pour mener à bien ce travail.*

*Je souhaite également remercier Madame Martinez Perez et Madame Moyse pour avoir accepté d'assumer le rôle de lectrices de ce mémoire et d'y avoir consacré du temps.*

*Ensuite, je remercie de tout cœur mes parents, qui m'ont accompagnée, guidée et conseillée avec bienveillance tout au long de mes études. Ils ont été là pour essuyer mes larmes, écouter mes doutes, mais aussi pour partager mes moments de joie. Les rendre fiers sera sans doute ma plus belle réussite.*

*Je remercie aussi ma sœur jumelle, Laura, qui est là depuis toujours et dont le soutien a compté plus qu'elle ne peut l'imaginer. J'ai également une pensée pour mes amies, Laura, Olivia, Tania, Justine et Kelly, qui ont toujours cru en moi et qui m'ont encouragée à persévérer. De même que Tom, mon copain, qui n'a cessé de me motiver et qui sera le premier derrière la ligne d'arrivée, fier.*

*Enfin, je tiens à adresser un mot tout particulier à mon binôme. Nous ne nous connaissons pas avant de nous lancer dans cette grande aventure qu'est le mémoire, et pourtant... je suis profondément heureuse que nos chemins se soient croisés et qu'elle laisse une empreinte si précieuse dans mon parcours universitaire. Dès le début, une compréhension mutuelle s'est installée entre nous, et notre soutien réciproque n'a jamais failli. Si c'était à refaire, je ne l'imaginerai avec personne d'autre.*

## Liste des tableaux

Tableau 1 - Éléments de normalité pour la dénomination lexicale pour les enfants de 2e primaire	- 49 -
Tableau 2 - Percentiles pour la dénomination lexicale des enfants de 2e primaire .....	- 49 -
Tableau 3 - Éléments de normalité pour la désignation lexicale pour les enfants de 2e primaire ....	- 50 -
Tableau 4 - Percentiles pour la désignation lexicale pour les enfants de 2e primaire .....	- 50 -
Tableau 5 - Éléments de normalité pour la dénomination lexicale pour les enfants de 5e primaire	- 54 -
Tableau 6 - Percentiles pour la dénomination lexicale pour les enfants de 5e primaire .....	- 54 -
Tableau 7 - Éléments de normalité pour la désignation lexicale pour les enfants de 5e primaire ....	- 55 -
Tableau 8 - Percentiles pour la désignation lexicale pour les enfants de 5e primaire .....	- 55 -
Tableau 9 - Données statistiques concernant les deux types d'erreurs réalisées .....	- 57 -
Tableau 10 - Test de Shapiro-Wilk évaluant la normalité des deux types d'erreurs .....	- 58 -

## Liste des figures

Figure 1 - Graphique de normalité de la dénomination lexicale pour les enfants de 2e primaire ....	- 48 -
Figure 2 - Graphique de normalité de la désignation lexicale pour les enfants de 2e primaire .....	- 50 -
Figure 3 - Graphique de normalité de la dénomination lexicale pour les enfants de 5e primaire ....	- 53 -
Figure 4 - Graphique de normalité de la désignation lexicale pour les enfants de 5e primaire .....	- 55 -
Figure 5 - Graphiques de comparaison des erreurs pour les enfants de 2e et 5e primaires .....	- 59 -

## Liste des abréviations

TDL : Trouble Développementale du langage

## Table des matières

Liste des tableaux .....	- 4 -
Liste des figures .....	- 4 -
Liste des abréviations .....	- 4 -
Table des matières .....	- 5 -
<b>Introduction générale.....</b>	<b>- 7 -</b>
<b>Introduction théorique .....</b>	<b>- 9 -</b>
<b>1. Le développement lexical.....</b>	<b>- 9 -</b>
1.1. Définition .....	- 9 -
1.2. Organisation du lexique mental .....	- 9 -
1.3. Facteurs d'acquisition du lexique .....	- 10 -
1.4. Étapes principales du développement lexical .....	- 11 -
<b>2. L'évaluation lexicale en logopédie .....</b>	<b>- 13 -</b>
2.1. L'intérêt de l'évaluation lexicale chez les enfants d'âge scolaire .....	- 13 -
2.2. La création d'une épreuve lexicale : les éléments clés.....	- 14 -
2.2.1. Tâche de désignation.....	- 15 -
2.2.2. Tâche de dénomination .....	- 15 -
2.3. Les caractéristiques psychométriques des tests de diagnostic.....	- 16 -
2.3.1. La standardisation .....	- 17 -
2.3.2. La normalisation .....	- 17 -
2.3.3. La validité .....	- 18 -
2.3.4. La consistance interne .....	- 19 -
2.3.5. La fidélité .....	- 20 -
2.3.6. La spécificité et la sensibilité .....	- 21 -
2.4. Les principaux outils francophones d'évaluation du lexique chez l'enfant d'âge scolaire .....	- 21 -
2.4.1. Analyse des principaux tests francophones d'évaluation du lexique .....	- 23 -
<b>Objectifs et hypothèses .....</b>	<b>- 35 -</b>
<b>Méthodologie .....</b>	<b>- 38 -</b>
<b>Comité d'éthique.....</b>	<b>- 38 -</b>
<b>Participants.....</b>	<b>- 38 -</b>
<b>Outils utilisés.....</b>	<b>- 39 -</b>
Les tests standardisés.....	- 39 -
Les épreuves expérimentales .....	- 43 -
<b>Procédure.....</b>	<b>- 45 -</b>
<b>Résultats .....</b>	<b>- 46 -</b>
<b>2<sup>e</sup> primaire .....</b>	<b>- 46 -</b>
Hypothèse 1 : Consistance interne .....	- 46 -
Hypothèse 2 : Validité concourante .....	- 47 -
Hypothèse 3 : Fidélité inter-juges .....	- 47 -
Hypothèse 4 : Normalité .....	- 48 -
Hypothèse 5 : Comparaison dénomination et désignation .....	- 51 -
<b>5<sup>e</sup> primaire .....</b>	<b>- 52 -</b>
Hypothèse 1 : Consistance interne .....	- 52 -
Hypothèse 2 : Validité concourante .....	- 52 -
Hypothèse 3 : Fidélité inter-juges .....	- 53 -
Hypothèse 4 : Normalité .....	- 53 -
Hypothèse 5 : Comparaison dénomination et désignation .....	- 56 -

<b>Comparaison 2<sup>e</sup> primaire et 5<sup>e</sup> primaire .....</b>	<b>- 57 -</b>
Hypothèse 6 : Comparaison erreurs sémantiques et phonologiques pour la tâche de désignation.....	- 57 -
<b><i>Discussion</i>.....</b>	<b>- 59 -</b>
Réponses à nos hypothèses.....	- 59 -
Forces de l'étude .....	- 61 -
Limites .....	- 62 -
<b><i>Conclusions</i> .....</b>	<b>- 62 -</b>
<b><i>BIBLIOGRAPHIE</i>.....</b>	<b>- 64 -</b>
<b><i>ANNEXES</i>.....</b>	<b>- 74 -</b>
Annexe 1 - Récapitulatif des caractéristiques psychométriques des principaux tests francophones d'évaluation du lexique.....	- 74 -
Annexe 2 - Capture d'écran d'un exemple de représentation de la tâche de dénomination lexicale expérimentale .....	- 75 -
Annexe 3 - Capture d'écran d'un exemple de représentation de la tâche de désignation lexicale expérimentale .....	- 75 -
<b><i>RÉSUMÉ</i>.....</b>	<b>- 76 -</b>

## Introduction générale

Au cours de ces dernières décennies, de nombreuses recherches s'effectuent quant au développement et l'acquisition du langage. Ce dernier a une évolution exceptionnelle dans l'histoire humaine. Il est vital pour l'être humain. Il sert à satisfaire ses besoins physiques, sociaux, émotionnels, mais aussi à se faire une place dans la société (Kern, 2019). En maîtrisant ses codes, l'Homme a la possibilité d'acquérir et de partager des savoirs et donc de développer une éducation. Nous pouvons comparer le langage humain à des milliers de pièces qui peuvent être combinées de multiples manières. Alors qu'un nombre infini de ces associations s'avèrent correctes, d'autres semblent erronées. Le défi de chacun est donc de comprendre ces structures et de les utiliser de façon appropriée (Saffran, 2003). Boerma et al. expliquaient, en 2023, que les enfants qui ont un bon développement langagier, auront un développement social ultérieur ainsi que des compétences académiques plus profitables. En effet, ils relatent que les enfants ayant un trouble développemental du langage (TDL) sont plus susceptibles de rencontrer des difficultés sociétales tout au long de leur vie (Boerma et al., 2023). L'enfant, dès sa première année de vie, est confronté au challenge de développer et d'acquérir le langage. Alors que cela semble aller de soi et se réaliser sans effort, il acquiert environ 14 000 mots, depuis sa naissance jusqu'à l'âge de 6 ans. Ce développement est donc un réel exploit intellectuel (Daviault, 2011). C'est pourquoi les logopèdes portent un intérêt particulier au développement de ce langage. Ils se soucient de la prévention, de l'évaluation et du traitement de nombreuses pathologies dont le TDL (Maillart, 2019).

Notre travail s'inscrit dans une thèse qui porte sur l'évaluation du TDL. Ce trouble touche approximativement 3 à 7% de la population (Bishop et al., 2017). Il est caractérisé par des difficultés d'acquisition et d'utilisation du langage oral dans les deux versants, à savoir la compréhension et la production. Les personnes qui présentent ce diagnostic se plaignent de répercussions dans leur vie quotidienne (les impacts fonctionnels), et ce, depuis l'enfance jusqu'à l'âge adulte. Effectivement, elles rencontrent des difficultés scolaires, des restrictions sociales et professionnelles (Bishop et al., 2017).

C'est la raison pour laquelle il est primordial d'évaluer précocement les compétences langagières des enfants afin de pallier d'éventuelles difficultés. Actuellement, les logopèdes recourent à une série de tests pour réaliser une évaluation des compétences langagières. Parmi les compétences évaluées, le thérapeute va chercher à connaître les capacités lexicales de son patient. En effet, le lexique est capital pour les apprentissages scolaires (Dujardin et al., 2021)



et permet d'avoir une communication fonctionnelle (Ebbels et al., 2022). Plusieurs tests sont disponibles pour l'évaluer de multiples manières, c'est-à dire que nous pouvons évaluer la richesse, l'accès lexical ou encore l'étendue lexicale. Des épreuves de première ligne telles que la dénomination d'images ou la désignation d'images peuvent être proposées, tout comme des tâches de deuxième ligne en production (fluences sémantiques ou phonologiques, définitions, antonymies, synonymies, devinettes, catégorisations) ou en réception (associations sémantiques, jugement sémantique, catégorisation, désignation à partir d'indices) (Steele & Gibbons, 2023). Cependant, comme nous l'aborderons dans la suite de ce manuscrit, les tests disponibles en français respectent peu les recommandations quant aux caractéristiques psychométriques (Bragard et al., 2010; McCauley & Swisher, 1984). Nous ne possédons pas de tests lexicaux contenant des normes récentes avec des items identiques en compréhension et en production. Or, ceci nous permet de comparer ces deux versants. Nous souhaitons donc, dans le cadre de cette étude, valider et normer une épreuve de désignation (compréhension) et de dénomination (production) lexicale ayant les mêmes items. La thèse qui chapeaute notre mémoire porte sur la création et la validation d'une évaluation lexicale sensible et spécifique d'enfants ayant un TDL afin d'en améliorer le diagnostic. Dès lors, la première étape est de normer et valider ces deux épreuves lexicales auprès d'enfants scolarisés de la 3<sup>e</sup> maternelle à la 5<sup>e</sup> primaire. Plus précisément, notre travail porte sur la validation et la normalisation des tests parmi des élèves tout-venant de 2<sup>e</sup> primaire et de 5<sup>e</sup> primaire. La première partie de ce travail traite des aspects théoriques liés au développement lexical, son organisation, ses facteurs d'acquisition et ses principales étapes. Ensuite, les perspectives de l'évaluation logopédique, et plus précisément de l'évaluation lexicale, sont envisagées. Nous parcourons les différents outils disponibles dans la littérature, les éléments pertinents à la création d'épreuves lexicales ainsi que les principaux critères psychométriques. Au niveau de la partie pratique, nous présentons nos objectifs et nos hypothèses quant à la validation et la normalisation des deux épreuves lexicales. Ensuite, nous expliquons notre procédure expérimentale incluant la sélection des participants, la description des outils utilisés, et la description des tâches expérimentales. Nous exposons nos résultats et nous finissons par une discussion portant sur les résultats de notre étude. Enfin, nous clôturons ce mémoire par une conclusion reprenant les éléments principaux de notre travail.

# Introduction théorique

## 1. Le développement lexical

### 1.1. Définition

Le lexique est communément défini comme « l'ensemble des mots connus d'une langue donnée » (Daviault, 2011). Il est considéré comme l'ensemble des entités significatives comprises et produites par un individu. Celles-ci peuvent être plus ou moins complexes selon les morphèmes (forme minimum douée de sens) ou lexèmes (unité minimale de signification appartenant au lexique) (Bogliotti, 2012). En 2019, Kern ajoute qu'un adulte produit entre 20 000 et 50 000 mots, mais en comprend davantage.

Le lexique est donc une organisation indispensable à l'apprentissage d'une langue. Cette acquisition va permettre à l'enfant, dès ses premières années de vie, d'exprimer ses différents besoins, mais également de favoriser le développement langagier (phonologie, morphosyntaxe, discours, pragmatique). De plus, le développement langagier ne peut pas être séparé des autres dimensions du développement, telles que le cognitif, l'affectif, le psychomoteur, le social, etc. D'après plusieurs travaux, les compétences lexicales prédisent également la réussite scolaire (Florin, 2010).

### 1.2. Organisation du lexique mental

Le lexique mental est un ensemble organisé et interconnecté de représentations encodées en mémoire à long terme. Un mot contient des représentations sur sa signification (sémantique), son utilisation (morphosyntaxique et contraintes d'usage) et sa forme (phonologie, morphologie et orthographique). Prenons l'exemple du mot « chien ». Il véhicule une signification (c'est un mammifère à quatre pattes, qui se déplace, qui aboie, ...), une utilisation (un nom commun masculin) et une forme (une représentation phonologique : /ʃiɛ̃/ ; une représentation orthographique : chien ; et une représentation morphologique (le morphème correspond au plus petit élément significatif d'un mot) : chien (González-Fernández & Schmitt, 2020).

Par ailleurs, l'enfant va développer son stock lexical de manière verticale (plusieurs niveaux de catégories sémantiques), mais aussi horizontale (plusieurs items dans chaque catégorie

sémantique). Rakison et Yermolayeva, en 2010, expliquaient que le développement lexical se développe en trois niveaux :

- Le niveau sur-ordonné correspondant aux catégories générales (par exemple les moyens de transport)
- Le niveau de base relatif aux items distinctifs de chaque catégorie (par exemple les voitures)
- Le niveau sous-ordonné se rapportant aux items spécifiques (par exemple les citadines, les 4x4)

L'enfant va d'abord acquérir des items typiques de chaque catégorie (développement horizontal) et ensuite, développer des catégories lexicales et des items spécifiques (développement vertical) (Cronin, 2002; Rosch et al., 1976). Les unités vont alors être davantage interconnectées pour former des réseaux lexicaux (Borovsky, 2022). Cependant, toute personne comprendra davantage de mots qu'elle ne les produit, car le stock de mots compris est plus important que le stock de mots produits (Zhong, 2018).

En outre, Perraudin et Mounoud, en 2009, expliquaient que les enfants de 5 et 7 ans caractérisent les objets préférentiellement par leurs actions (par exemple « couteau » - « pain ») alors que les enfants de 9 ans et les adultes les caractérisent grâce à des termes partageant les mêmes propriétés sémantiques (par exemple « couteau » - « ciseaux »).

### 1.3. Facteurs d'acquisition du lexique

L'acquisition de cette organisation au sein du stock lexical va reposer sur différentes capacités et sur des connaissances plus larges telles que des connaissances dans d'autres domaines afin de faciliter l'apprentissage lexical (Christophe et al., 2008). Effectivement, l'enfant va pouvoir recourir à ses compétences de segmentation phonologique, c'est-à-dire ses capacités à reconnaître et à manipuler les sons présents au sein des mots ; ses compétences morphologiques, c'est-à-dire ses capacités à comprendre la structure des mots (affixes) ; ses compétences syntaxiques, c'est-à-dire la faculté de comprendre comment les mots s'agencent pour former des phrases correctes ; ses compétences pragmatiques, c'est-à-dire la manière adéquate avec laquelle il va utiliser le langage selon le contexte ; ses compétences sémantiques, la capacité à comprendre et à utiliser les mots et ses compétences cognitives comme la mémoire,

l'attention, les capacités générales d'apprentissage ainsi que la capacité de raisonnement (Bloom, 2001). Effectivement, la mémoire à court terme et plus précisément la boucle phonologique aide à la rétention d'informations et à l'apprentissage lexical (Bussy et al., 2014). Enfin, certains principes innés vont venir soutenir et guider l'enfant lors de son apprentissage tels que la généralisation lexico-sémantique (c'est lorsqu'un terme qui est utilisé à l'origine pour désigner une chose spécifique, en vient à désigner une catégorie plus large), la conventionnalité (c'est l'idée selon laquelle le sens des mots et des expressions est influencé par des conventions sociales et linguistiques), la dénomination catégorielle (il s'agit d'un processus qui consiste à utiliser un mot pour désigner un ensemble d'objets ou de concepts) ainsi que les principes d'exclusivité mutuelle (il s'agit d'un principe où l'enfant attribue le nouveau mot entendu à un objet dont il ne connaît pas le nom plutôt qu'à un objet dont il connaît déjà le nom) et de l'objet entier (principe selon lequel lorsque l'on désigne un objet, on le désigne dans son ensemble et non une partie de celui-ci) (Colunga & Sims, 2017; Krzemien et al., 2020; Perry & Samuelson, 2011). Bogliotti (2012) ajoute que la discrimination auditive est un facteur important pour la bonne représentation phonologique du stock lexical. Les enfants TDL ayant des difficultés concernant cette discrimination, entre autres, ont des représentations lexicales plus appauvries, mal définies.

#### 1.4. Étapes principales du développement lexical

Grâce à ces différents facteurs, l'enfant va, dès les premières années de vie, acquérir et suivre une trajectoire développementale quant à l'acquisition lexicale. Il va passer par différentes étapes afin de se constituer un stock lexical riche et varié (Kern, 2019).

Selon Daviault, en 2011, bien que l'enrichissement se fait tout au long de la vie, l'acquisition lexicale est principalement active lors de 4 grandes périodes :

- Le lexique précoce, entre 1 an et 1 an et demi
- L'explosion lexicale, entre 1 an et demi et 2/3 ans
- L'âge préscolaire, entre 2/3 ans et 5/6 ans
- L'âge scolaire, entre 5/6 ans et 12 ans

L'enfant va, donc, produire ses premiers mots aux alentours de son premier anniversaire. Au cours de cette première phase, il va produire environ 50 mots. Ceux-ci reflètent le quotidien et les premières expériences de vie de l'enfant. Effectivement, il produit majoritairement des noms

correspondant à des personnes proches, des vêtements, de la nourriture, des parties du corps, des animaux et des objets de la vie quotidienne. À ce stade, les productions des enfants sont davantage composées de noms alors que les verbes et les morphèmes grammaticaux (pronoms, prépositions et articles) constituent un pourcentage minime de celles-ci (MacRoy-Higgins et al., 2016). Frewin et al. (2025) expliquent que les jeunes enfants sont davantage avantagés par leurs diverses expériences motrices pour l'acquisition des verbes.

C'est pourquoi, aux alentours des 18 mois, l'enfant aura acquis environ 50 mots, apprendra plus facilement du nouveau vocabulaire et produira des combinaisons de 2 mots. Ceci marquera le début du développement de la grammaire, entre autres, grâce à l'exploration de l'environnement (aidé par la marche) (Feldman, 2019; Frewin et al., 2025) .

C'est à l'âge de 2 ans que l'acquisition lexicale est spectaculaire. Il s'agit de la période d'explosion lexicale. En 2017, Colunga et ses collaborateurs détaillent cette étape en expliquant qu'un enfant apprend entre 5 et 9 nouveaux termes par jour, ceci avec, tout de même, une variabilité interindividuelle. L'enfant combine alors des morphèmes connus c'est-à-dire les plus petites unités porteuses de sens (par exemple « elle » et « mange » deviendra « Elle mange »).

Lorsque l'enfant entre à l'école maternelle, cet apprentissage semble ralentir pour laisser place au développement phonologique, morphologique et syntaxique. Daviault, en 2011, relate qu'à cette période, l'enfant acquiert environ 2 nouveaux termes par jour. La complexité rencontrée dans les autres composantes semble donc être la cause de ce ralentissement (Tabouret-Keller, 2021).

Enfin, lors de la période scolaire, l'apprentissage lexical est fortement lié au développement du langage écrit. Il s'agit d'un lien bidirectionnel entre ces deux compétences (Cain & Oakhill, 2011). En effet, le stock lexical actif s'élève à 14 000 mots à l'âge de 6 ans et évolue jusque 12 ans pour atteindre environ 55 000 termes (Daviault, 2011). Nous observons alors des réorganisations sémantiques, ainsi que des approfondissements de connaissances lexicales en plus de l'acquisition de nouveaux lexèmes (l'unité de base porteuse de sens ; exemple « manger »). Nous parlons alors d'étendue (la quantité de lexèmes connus) et de profondeur (la qualité des représentations lexicales) du vocabulaire (Steele & Gibbons, 2023).

## 2. L'évaluation lexicale en logopédie

L'évaluation de cette acquisition, plus généralement appelée « évaluation lexicale » remonte au temps des premiers tests d'intelligence de Binet et Thurstone (Pearson et al., 2007). Les élèves devaient définir ou expliquer des termes qui étaient susceptibles de se retrouver dans des textes scolaires. Lors de la Première Guerre mondiale, l'évaluation formelle a commencé son essor. Afin de tester les nouvelles recrues, il était nécessaire de passer à des tests plus standards et plus facilement administrables. C'est au cours des années 70 que nous voyons apparaître l'évaluation lexicale contextualisée où les enfants doivent comprendre un texte pour pouvoir répondre aux questions concernant le vocabulaire. Cependant, à l'heure actuelle, il n'est pas rare d'utiliser des tests d'évaluation lexicale isolée. Il s'agit alors principalement de tâches de désignation et de dénomination (Bragard et al., 2010).

### 2.1. L'intérêt de l'évaluation lexicale chez les enfants d'âge scolaire

Duff et al. (2015) exposent le fait que le vocabulaire des enfants d'âge scolaire doit être évalué, car il constitue un prérequis crucial pour divers aspects de leurs apprentissages. Au cours des trois dernières décennies, la recherche sur le vocabulaire a connu une expansion significative. Le vocabulaire représente un élément capital intermédiaire entre la langue orale et écrite, influençant diverses compétences linguistiques.

Des études (Brinchmann et al., 2015 ; Cain & Oakhill, 2014 ; Ouellette & Beers, 2010 ; Quinn et al., 2015 ; Tunmer & Chapman, 2012) ont mis en évidence son impact sur des aspects tels que les capacités linguistiques générales, à savoir lire, parler, écrire et écouter ainsi que sur le décodage et la compréhension.

Bishop et al. (2017) relatent que les enfants qui présentent un retard au niveau du vocabulaire entre 18 et 24 mois pourront entraîner des répercussions à long terme. Ils peuvent présenter des faiblesses quant au premier stade du développement de la syntaxe, la combinaison syntaxique. Jullien (2008) ajoute que la composante syntaxique peu développée va jouer un rôle négatif quant au développement discursif.

De plus, le lexique et la phonologie s'influencent réciproquement (le lexique aide à affiner la perception phonologique et cette dernière encourage la bonne compréhension lexicale de termes

proches) ; le développement lexical exerce un rôle sur la composante phonologique (Arias-Trejo et al., 2025). Enfin, le vocabulaire joue un rôle dans le développement du langage écrit, capacité indispensable pour une scolarité ordinaire (lecture, orthographe et compréhension écrite) (Bianco, 2021; Duff et al., 2015; Dujardin et al., 2021; Ecalle et al., 2021)

## 2.2. La création d'une épreuve lexicale : les éléments clés

Un des buts principaux d'un test diagnostique est d'identifier si le sujet présente un déficit dans la compétence observée. Bragard et al., en 2010, exposent plusieurs paramètres pour la mise au point d'une épreuve d'évaluation lexicale.

Tout d'abord, pour une meilleure comparaison des performances, il est souhaitable de proposer les mêmes items dans une épreuve de désignation et de dénomination. Effectivement, cela laisse penser à un éventuel manque du mot si le sujet a de bonnes performances lors de la tâche de désignation, mais une faiblesse voire un déficit dans celle de dénomination. Différentes catégories sémantiques doivent être comprises dans l'épreuve afin d'évaluer le vocabulaire le plus écologiquement possible, c'est-à-dire selon le développement lexical (une série de mots appartenant à divers domaines).

Ensuite, au sein des épreuves basées sur la dénomination et la désignation, ces items doivent être imageables et ce, de façon non ambiguë. Effectivement, il faut que les concepts reposent sur un consensus entre le nom et sa représentation. Les mots qui ont une grande imageabilité sont reconnus et dénommés plus vite que ceux qui ont une faible imageabilité (Dymarska et al., 2023). Par ailleurs, les illustrations colorées permettent une meilleure représentation de la réalité (photographie, dessin). Ces dernières représentent davantage la réalité plutôt que des images en noir et blanc. Pour éviter d'apporter une aide contextuelle, un fond uni commun doit être utilisé pour tous les items (Bragard et al., 2010).

Les éléments sélectionnés peuvent également être contrôlés selon d'autres critères psycholinguistiques comme l'âge d'acquisition, la longueur des mots et la complexité phonologique. Ces derniers seraient responsables du temps de récupération d'un mot (Bragard et al., 2010). Manoiloff et al. (2013) ajoutent que les termes les plus fréquents sont plus vite dénommés. Cette dernière variable est considérée comme l'une des plus importantes à contrôler (Hashimoto, 2021). Plusieurs auteurs expliquent qu'elle impacte la récupération lexicale du sujet. La fréquence correspond à l'occurrence d'un mot dans un corpus donné. Effectivement,

un mot plus fréquent nécessite moins de traitement et donc moins de temps pour être récupéré dans le stock lexical. En revanche, un item moins fréquent aura un seuil d'activation plus lent (Dent et al., 2008; Morrison & Ellis, 1995). Il s'agit donc d'un effet de fréquence (Ferrand, 1997). La fréquence d'un mot est fortement corrélée avec l'âge d'acquisition. C'est l'âge approximatif auquel un individu acquiert un mot dans son enfance (Bonin et al., 2021). Il a été reconnu comme la variable prédictive principale pour les tâches de dénomination des mots et le traitement lexical à tous les âges (Raman et al., 2014). Enfin, la longueur lexicale concerne le nombre de syllabes, phonèmes, graphèmes inclus dans l'item alors que la complexité phonologique relève de la relation entre les séquences de phonèmes du mot (Gierut, 2007).

### *2.2.1. Tâche de désignation*

Une épreuve de désignation lexicale a pour objectif d'évaluer la compréhension des mots chez un individu. Elle consiste à lui demander d'associer un mot donné à l'item imagé correspondant. Plus particulièrement, pour une épreuve de désignation lexicale, le choix des distracteurs est important. Nous en relevons principalement 3 types : un distracteur phonologique (ayant une relation au niveau des sons avec l'item cible), un distracteur sémantique (ayant une relation au niveau du sens avec l'item cible) et un distracteur neutre (sans relation avec l'item cible). Si nous prenons l'item « chat » comme item cible, nous pouvons imaginer que « bas » soit le distracteur phonologique, « chien » soit le distracteur sémantique et « table » soit le distracteur neutre (Bragard & Schelstraete, 2006). Bragard et al. (2010) ainsi que Sheng et McGregor (2010) expliquent que les enfants produisent davantage d'erreurs de type sémantique en grandissant. Dès lors, une comparaison des deux types d'erreurs peut être réalisée.

### *2.2.2. Tâche de dénomination*

Pour la construction d'une tâche de dénomination, il est intéressant d'obtenir le temps de dénomination. Malheureusement, peu de batteries le permettent, par exemple, la NEEL, l'ELO ou encore l'EVIP ne le permettent pas. De plus, il est également intéressant d'avoir les mêmes items en production et en compréhension afin de pouvoir comparer les compétences du patient entre le versant production et le versant compréhension. De plus, avoir les mêmes items permet de souligner un possible manque du mot. Il est important de savoir faire la différence entre un manque du mot et un manque de vocabulaire. Un manque du mot sera caractérisé par une



lenteur d'accès lexical et une imprécision lors des réponses (Bragard et al., 2010). Qui plus est, certaines variables affectent ce type d'épreuves au niveau du temps et/ou de la précision de dénomination comme la complexité visuelle, la fréquence des mots, leur âge d'acquisition, la complexité phonologique, la familiarité des concepts et la longueur des mots. Il est donc crucial de considérer ces variables lors de la création d'une tâche de dénomination (Bragard et al., 2010).

Dans le cadre d'une tâche de dénomination, il est également pertinent de proposer une ébauche phonologique. Les résultats d'une étude (Wing, 1990) montrent que les activités de segmentation phonologique aident les enfants d'âge scolaire à récupérer un mot en mémoire lorsqu'ils présentent des difficultés. Plus précisément, le fait d'obtenir une représentation partielle du mot comme le son initial ou le nombre de syllabes peut permettre à l'enfant de récupérer plus facilement un mot. Un autre auteur a une conclusion similaire en relatant que les enfants bénéficient de l'aide phonologique pour la récupération des mots lorsque ceux-ci sont présents dans le stock lexical (Best, 2005). De même que Kazanina et al., en 2018, écrivent que pour accéder aux représentations lexicales, les phonèmes jouent un rôle crucial.

Enfin, Bragard et al. (2010) expliquent que les enfants commettent davantage d'erreurs phonologiques au début du développement pour ensuite, au cours de celui-ci, réaliser plus de réponses erronées de type sémantique. En revanche, les enfants ayant un trouble développemental du langage (TDL) produisent moins d'erreurs sémantiques que les enfants sans TDL. Effectivement, un enfant TDL ayant une profondeur lexicale et un réseau lexical amoindris, il produiront davantage de réponses de type « Je ne sais pas » en dénomination (Sheng & McGregor, 2010). Ces enfants ont des représentations phonologiques et sémantiques plus faibles et moins structurées qu'un enfant ayant un développement lexical typique (Bogliotti, 2012).

### 2.3. Les caractéristiques psychométriques des tests de diagnostic

En 1984, McCauley et Swisher relatent qu'un test, qui bénéficie de caractéristiques psychométriques appropriées, offre au clinicien une certaine sensibilité par rapport au potentiel du test à identifier un trouble.

Les tests standardisés constituent des éléments essentiels de l'évaluation du langage, mais ces derniers présentent des limites comme des échantillons de normalisation inappropriés, peu d'informations sur la sensibilité et la spécificité et un manque de validité pour pouvoir tester les compétences communicatives dans la vie quotidienne (Caesar & Kohler, 2009; Spaulding et al., 2006). De plus, certains tests de vocabulaire standardisés sont limités et n'offrent pas assez d'informations. Et enfin, les tests standardisés ne permettent pas de suivre les progrès étant donné qu'ils ne peuvent pas être administrés plusieurs fois sur une courte période. Pourtant, les logopèdes utilisent les tests pour adapter au mieux les prises en charge et pour suivre les progrès (Steele & Gibbons, 2023). Youngstrom et al. (2017) différencient trois types de tests : les prédictifs, ceux qui sont administrés en premier et qui permettent de prendre la décision de creuser davantage l'évaluation ; les prescriptifs, ceux qui creusent les hypothèses du clinicien et permettent d'orienter la prise en soin ; les progressifs, ceux qui évaluent les progrès du patient au fil de la prise en soin.

À présent, nous allons aborder les caractéristiques psychométriques essentielles pour les tests diagnostiques.

### *2.3.1. La standardisation*

Selon McCauley et Swicher (1984), la standardisation, qui correspond à la capacité de placer les individus dans les mêmes conditions d'évaluation, procure à l'utilisateur une certaine confiance que le score obtenu par le participant est une véritable indication de son niveau de compétence. Celui-ci ne doit pas être influencé par des facteurs externes tels que la manière dont le test a été administré, les instructions ou les questions posées, le contexte de passation du test ainsi que les règles de cotation.

### *2.3.2. La normalisation*

Les normes d'un test récapitulent statistiquement les scores obtenus par un échantillon normatif. Elles sont cruciales pour évaluer un enfant testé en le comparant à des pairs de même âge. Aussi, nous savons que les échantillons normatifs devraient idéalement présenter d'autres caractéristiques démographiques communes avec l'enfant évalué comme le sexe, le statut socio-économique, l'origine ethnique et la région géographique (Spaulding et al., 2012).

Les normes provenant de divers groupes normatifs peuvent être nécessaires pour fournir divers types d'informations sur la présence d'un éventuel problème. De manière optimale, les créateurs de tests devraient établir des normes couvrant une diversité de groupes assez étendue afin de permettre à l'utilisateur du test d'évaluer de manière pertinente tous les candidats potentiels.

Néanmoins, nous savons qu'il est rare de trouver un échantillon normatif représentatif, même pour la question d'évaluation la plus courante, à savoir « est-ce qu'il y a un déficit ? ».

Afin que l'utilisateur du test puisse comparer le score du participant par rapport aux normes, le manuel du test doit contenir assez d'informations sur l'échantillon normatif (McCauley & Swisher, 1984).

En outre, idéalement les manuels de tests devraient avoir des normes construites sur un échantillon d'enfants ayant un trouble du langage dans le but d'une évaluation clinique (Spaulding et al., 2012) mais aussi dans le but d'évaluer la sensibilité et la spécificité du test. Ce tableau permet aux cliniciens de comparer les performances d'un enfant ayant des troubles du langage avec celles d'enfants ayant divers degrés de sévérité de troubles du langage, ce qui aide à évaluer la sévérité du déficit langagier de l'enfant. A l'heure actuelle, les enfants présents dans les normes sont des individus tout-venant respectant des critères d'inclusion tels que l'âge, le niveau socio-culturel ainsi que le genre. Nous pouvons évaluer la normalisation d'un test grâce à des tests statistiques, tels que le test de normalité. Le résultat obtenu est comparé au seuil de significativité, à savoir 0,05.

### *2.3.3. La validité*

La validité psychométrique d'un instrument de mesure est déterminée par sa capacité à évaluer avec précision ce qu'il est censé mesurer. Par exemple, « un test de vocabulaire réceptif est un instrument de mesure valide du vocabulaire réceptif dans la mesure où il mesure avec précision le vocabulaire réceptif » (McCauley & Swisher, 1984).

Néanmoins, certaines caractéristiques peuvent nuire à la validité d'un test. Par exemple, si le test est trop long, la fatigue peut être un facteur qui peut nuire aux performances. Une autre caractéristique est aussi fondamentale, à savoir la fiabilité. Si le test n'est pas fiable, il ne mesurera pas de manière cohérente ce qu'il est censé mesurer et donc il s'agira d'un instrument de mesure moins valide (McCauley & Swisher, 1984).

Il existe plusieurs types de validité, nous reviendrons uniquement sur la validité concourante dans la partie résultat. À savoir :

- La validité théorique est le fait de vérifier que le test mesure réellement le concept qu'il est censé mesurer. Elle est examinée par une comparaison minutieuse de la définition du concept à tester par l'auteur du test avec le contenu réel du test. Pour l'évaluer, il faut qu'un clinicien compétent dans le domaine du comportement évalué par le test examine à la fois les éléments spécifiques et l'ensemble du contenu du test (McCauley et Swisher, 1984).
- « La validité concourante d'un test est déterminée en évaluant dans quelle mesure le score d'un individu au test est lié à son score sur une variable critère qui est mesurée à peu près au même moment où le score du test est obtenu ». Il s'agirait, par exemple, de comparer les résultats du test avec les jugements des experts sur l'importance du problème de langage d'un enfant (McCauley & Swisher, 1984). De plus, selon Friberg, en 2010, la validité concourante est importante car elle démontre que les résultats d'un outil d'évaluation sont plus fiables lorsque ses résultats sont cohérents avec ceux d'un autre outil, lui-même validé, mesurant un concept similaire. Il relate qu'il s'agit d'une corrélation entre des résultats obtenus à un test avec les résultats obtenus à un autre test évaluant la même chose.
- La validité prédictive d'un test est évaluée en analysant dans quelle mesure le résultat obtenu par un individu à un test peut être utilisé pour prédire ses performances futures. Entre autres, les données provenant d'un test validé pourraient servir de base essentielle pour distinguer différents sous-groupes en fonction de leur performance (McCauley & Swisher, 1984).

#### *2.3.4. La consistance interne*

La consistance interne fait référence à la cohérence des items qui sont normalement censés évaluer les mêmes concepts dans un test. Une bonne consistance interne est un indicateur d'une bonne fiabilité (Bourque et al., 2020).

Selon Zhang (2022), le coefficient alpha de Cronbach est utilisé pour examiner la fiabilité en termes de consistance interne. Elle est considérée comme excellente lorsque les valeurs sont supérieures à 0.9, bonne pour les valeurs entre 0.8 et 0.9 et acceptable entre 0.7 et 0.8. D'autres tests, comme le KR-20, le permettent. Ce dernier admet un seuil d'acceptabilité de 0.7 (Ntumi et al., 2023).

#### *2.3.5. La fidélité*

La fidélité est l'une des caractéristiques les plus importantes, tout comme la validité d'un test. En effet, « la fidélité fait référence à la cohérence avec laquelle un test mesure un attribut ou un comportement donné. Un instrument de mesure parfaitement fiable donne systématiquement la même valeur lorsque la même variable est mesurée » (McCauley & Swisher, 1984).

Plusieurs sortes de fidélités existent :

- La fidélité inter-juges qui permet à la personne qui va utiliser le test de savoir dans quelle mesure les différents administrateurs peuvent influencer les résultats du test. Les scores obtenus par une même personne testée par plusieurs administrateurs sont comparés pour évaluer sa valeur. Nous n'aborderons que ce type de fidélité dans notre travail. En effet, nous sommes plusieurs étudiantes à administrer les tâches et donc nous suivons les mêmes consignes d'administration et de correction. Les instructions ont été élaborées pour minimiser les différences entre les évaluateurs. Youngstrom et al (2017) estimaient qu'une corrélation de 0.8 était bonne pour cette fidélité.
- La fidélité test-retest fait référence à la stabilité des résultats des tests dans le temps (McCauley & Swisher, 1984). Une corrélation de 0.9 est recommandée (Friebert, 2010). Cette caractéristique psychométrique ne sera pas analysée dans notre mémoire car nous n'allons, à aucun moment, administrer à deux reprises une même tâche à un participant car le vocabulaire est une composante assez stable qui ne varie que très peu entre deux passations.

### *2.3.6. La spécificité et la sensibilité*

« La sensibilité est considérée comme la probabilité qu'un enfant ayant déjà reçu un diagnostic de trouble du langage soit identifié comme tel lors de l'utilisation d'un outil d'évaluation du langage. En revanche, la spécificité est définie comme la possibilité qu'un enfant dont le développement est typique soit identifié comme tel lorsqu'il est testé à l'aide d'un outil d'évaluation donné. » (Friberg, 2010).

Elles sont mesurées en pourcentage et ce pourcentage permet d'identifier la précision globale d'un outil d'évaluation pour établir un diagnostic précis. Les valeurs de sensibilité et de spécificité se situent entre 0% et 100%, lorsque les valeurs se rapprochent de 100%, elles reflètent les diagnostics les plus sensibles et spécifiques (Friberg, 2010).

Ces caractéristiques psychométriques importantes ne seront pas analysées car notre échantillon se compose uniquement d'enfants sans trouble développemental du langage. Le but des nouvelles épreuves étant de diagnostiquer un trouble lexical, nous devons prendre en considération plusieurs caractéristiques psychométriques : la consistance interne, la fidélité inter-juges, la présence de normes, la validité de contenu, la validité prédictive, la validité convergente et discriminante ainsi que la sensibilité et la spécificité (Cattini & Lafay, 2024; Youngstrom et al., 2017). Cependant, nous allons réaliser des normes auprès d'enfants n'ayant pas de troubles du langage. Dès lors, nous prendrons en compte la consistance interne, la fidélité inter-juges, la validité concourante et la présence de normes.

## *2.4. Les principaux outils francophones d'évaluation du lexique chez l'enfant d'âge scolaire*

Nous disposons de différents tests pour l'évaluation du vocabulaire en langue française. Ils évaluent principalement le niveau lexical en compréhension et en production grâce à des tâches respectives de désignation et de dénomination (Bragard et al., 2010).

De façon générale, nous retrouvons les batteries d'évaluation du langage. Hadley et Dickinson, en 2020, définissent les batteries de testing comme des tests qui utilisent plusieurs subtests pour mesurer les compétences lexicales au sein d'une appréciation plus globale du langage. Dès lors, ces tests n'évaluent pas l'ensemble des aspects des connaissances lexicales comme la morphologie dérivationnelle ou les capacités d'apprentissage de l'enfant (évaluation dynamique).

Au sein de ces batteries, nous pouvons retrouver des subtests d'évaluation de première ligne du vocabulaire, faciles et rapides à administrer. Ils évaluent un aspect du vocabulaire comme la compréhension de mots isolés. Il est demandé à l'enfant de pointer des images selon le mot énoncé pour évaluer sa compréhension par exemple. En ce qui concerne la production, l'enfant doit souvent dénommer des images, mais il est également possible de lui demander des définitions, des explications relatives à la fonction, la catégorie, les relations sémantiques, des synonymes et antonymes des mots administrés.

Nous pouvons également recourir à d'autres types d'évaluation tels que l'analyse d'échantillons de langage et les tâches créées par les cliniciens, qui sont davantage qualitatives et qui les aident à identifier les déficits de vocabulaire (Steele & Gibbons, 2023). Les logopèdes peuvent, par exemple, créer leurs complétions de phrases, leurs tâches de catégorisation, leurs questionnaires lexicaux selon les apprentissages scolaires. Celles-ci servent davantage d'analyse qualitative des déficits et de monitoring de prise en soin comme des routines d'évaluation. De façon générale, nous pouvons évaluer la communication, dont le lexique, grâce à des grilles ou des questionnaires proposés à l'entourage de l'enfant. Ils montrent une validité et une fiabilité plus élevées que l'évaluation directe pour le diagnostic du trouble du langage chez le jeune enfant lorsqu'il est timide ou lorsque la présence de l'examineur pourrait fausser les résultats (Bishop & McDonald, 2009). Cependant, l'utilisation de questionnaires semble moins pertinente lors de l'évaluation des enfants d'âge scolaire. En effet, l'évaluation directe est plus opportune au vu des capacités d'attention et de concentration par exemple (Bishop & McDonald, 2009).

Enfin, certains auteurs mettent en évidence que la plupart des tests, jusqu'en 2021, sont en format « papier-crayon » (Dujardin et al., 2021). En revanche, les tests informatisés permettent une meilleure fidélité et validité grâce au contrôle des temps de réponse, le nombre de réponses correctes ou encore les modalités de passation (temps d'affichage des items par exemple) (Csapó et al., 2014). Ecalle et al (2021) ajoutent que les participants sont davantage engagés dans la tâche.

À l'heure actuelle, plusieurs batteries contiennent des épreuves évaluant le lexique en français comme l'ELO (Khomsî, 2001), L2MA-2 (Chevrié-Muller et al., 2011), NEE-L (Chevrié-Müller, C. & Plaza, M., 2001), EXALANG 5-8 (Helloin et al., 2010), EVALÉO 6-15 (Launay, L. et al., 2018) et EVALO 2-6 (Coquet F. et al., 2009).

### *2.4.1. Analyse des principaux tests francophones d'évaluation du lexique*

#### ELO (Khomsi, 2001)

Dans cette batterie de tests, l'échantillon se compose de 970 sujets ayant entre 3 ;3 ans et 10 ;3 ans, c'est-à-dire de la 1<sup>ère</sup> maternelle à la 5<sup>e</sup> primaire. Nous y retrouvons une épreuve de désignation lexicale et, ensuite, une épreuve de dénomination lexicale. Cette batterie se repose sur des bases de données relatives au langage écrit (lecture, orthographe). De façon générale, la batterie ne nous donne pas d'informations quant à la distribution des données, la structure factorielle, la validité (prédictive et discriminante) et la fidélité (consistance interne, inter-juges et test-retest). Nous relevons, en revanche, des informations concernant la standardisation (les consignes d'administration et de cotation), les scores (les tendances centrales de la batterie) et les objectifs de chaque épreuve. Les consignes d'administration et de cotation sont précises et claires, les résultats de chaque épreuve peuvent être analysés selon des mesures statistiques et toutes les épreuves ont des objectifs spécifiques.

Plus précisément, en ce qui concerne la dénomination lexicale, nous avons des informations quantitatives relatives aux échantillons de chaque année scolaire évaluée. Pour chacune de celles-ci, des sujets masculins et féminins ont été recrutés. Nous retrouvons 68 participants en 1<sup>ère</sup> maternelle, 191 en 2<sup>e</sup> maternelle, 158 en 3<sup>e</sup> maternelle, 129 en 1<sup>ère</sup> primaire, 103 en 2<sup>e</sup> primaire, 140 en 3<sup>e</sup> primaire, 86 en 4<sup>e</sup> primaire et 95 en 5<sup>e</sup> primaire. Aucun étalonnage n'est normalement distribué mais les auteurs ne donnent que des moyennes et des écart-types dans le manuel. Une analyse sera donc moins précise qu'avec des percentiles (utilisés en cas d'absence de normalité). L'enfant doit dénommer une image (noms ou verbes) présentée en noir et blanc (42 items pour les enfants de 1<sup>ère</sup> maternelle à la 3<sup>e</sup> primaire et 50 pour ceux de 4<sup>e</sup> et 5<sup>e</sup> primaires). L'expérimentateur donne la consigne au début du test (« Qu'est-ce que c'est ? » ou « Qu'est-ce qu'il fait ? » pour les enfants de 1<sup>e</sup> maternelle à ceux de 3<sup>e</sup> primaire) et ne peut pas donner de moyens d'aide au cours de l'épreuve. Les items présents dans le subtest ont une fréquence qui varie entre 33,16 et 69,79, ce qui indique que les items présents sont des items de faible fréquence (<45), de fréquence moyenne (45-65) et quelques items sont de haute fréquence (>65) (Peereman et al., 2007). Ceci est donc un atout pour observer un éventuel effet de fréquence (Ferrand, 1997). Le subtest ne présente pas de critère d'arrêt selon les difficultés du participant mais bien selon son année scolaire. Enfin, la cotation est binaire (0 ou 1) et



l'étalonnage de certaines années présente un effet plafond. En effet, pour l'étalonnage de la 3<sup>e</sup> primaire (maximum 32 points), nous observons une moyenne de 26,4 et un écart-type de 3. De plus, pour les 4<sup>e</sup> et 5<sup>e</sup> primaires, le subtest ne propose pas de verbes (« qu'est-ce qu'il fait ? »). Effectivement, nous n'avons qu'un étalonnage pour les noms.

En ce qui concerne la désignation lexicale, nous avons également des échantillons ayant une population mixte et une distribution en percentile. Nous retrouvons le même nombre de participants dans chaque étalonnage que pour l'épreuve de dénomination. À nouveau, nous présentons des images en noir et blanc (20 items) et l'enfant doit désigner la cible parmi 4 images (distracteurs sémantiques). Les cibles de l'épreuve ont une fréquence lexicale qui varie entre 38,74 et 64,71 bien que nous n'ayons pas l'information dans le manuel, il s'agit d'items de faible fréquence et de fréquence moyenne. Des items de haute fréquence auraient pu être bénéfiques (Ferrand, 1997). Les distracteurs possèdent des fréquences similaires à leurs cibles. L'activation des distracteurs et des cibles sont donc relativement identiques (Dent et al., 2008; Morrison & Ellis, 1995). C'est équivalent à l'épreuve de dénomination. L'épreuve ne présente pas de critère d'arrêt et l'examineur ne peut pas venir en aide à l'enfant lors de l'épreuve. La cotation est à nouveau binaire (0 ou 1) et nous observons un effet plafond pour les étalonnages à partir de la 2<sup>e</sup> primaire.

Bien que la batterie propose une tâche pour chaque versant lexical, nous remarquons que les normes ne sont plus récentes (2001) et non normales. En 2009, des auteurs expliquaient qu'on observait un effet Flynn dans la population. Les performances évoluent dans le temps à cause de facteurs sociaux, éducationnels ou encore biologiques (Nicollet et al., 2009). C'est pourquoi les normes doivent aussi évoluer. De plus, les tâches n'utilisent pas les mêmes items comme la littérature le recommande ou encore une bonne qualité concernant la présentation des items (items noirs et blancs) (Bragard et al., 2010). Les items ont une fréquence assez similaire (ainsi que les distracteurs avec les cibles). Elles sont entre 33,16 et 69,79. Dès lors, le contrôle de la fréquence a été réalisé. En revanche, dans l'épreuve de désignation lexicale, nous ne remarquons que des distracteurs sémantiques. Or, la littérature explique que les enfants plus jeunes font davantage d'erreurs phonologiques et les enfants plus âgés commettent plus d'erreurs sémantiques (Sheng & McGregor, 2010). C'est pourquoi des distracteurs phonologiques auraient été opportuns pour réaliser une comparaison.

Au vu des caractéristiques importantes à considérer pour un diagnostic (Cattini & Lafay, 2024; Youngstrom et al., 2017), cette batterie ne contient pas d'informations quant à la validité

prédictive, la consistance interne, la sensibilité et la spécificité, les fidélités test-retest et inter-juges, la structure interne, la validité concourante et la validité théorique. Enfin, le calcul de l'erreur standard n'est pas calculable.

#### L2MA-2 (Chevrié-Muller et al., 2011)

Dans cette batterie, l'échantillon se compose de 500 sujets de la 2<sup>e</sup> primaire à la 6<sup>e</sup> primaire. Elle comporte 31 tâches. De façon générale, le manuel nous informe de la validité discriminante de la batterie (sensibilité et spécificité). En revanche, nous n'avons aucune information quant à la distribution des données, la fidélité (inter-juges, test-retest et consistance interne) et la validité prédictive. La batterie est normée pour des enfants de 7 à 11 ans. Quant à la fréquence des items pour les enfants de 2<sup>e</sup> primaire, elle se situe entre 42,17 et 60,13 et entre 41,05 et 60,13 pour les enfants de 5<sup>e</sup> primaire. Pour les deux années scolaires, les items sont majoritairement de fréquence moyenne et quelques items de faible fréquence. La comparaison des items lors de la passation est donc plus compliquée en termes d'effet de fréquence (Ferrand, 1997).

Une épreuve de dénomination d'images est présente dans cette batterie. Les items sont sélectionnés selon le niveau scolaire et la fréquence. Trente items sont affichés à l'écran pendant 4 secondes. Alors que nous n'avons pas différentes catégories grammaticales, nous retrouvons plusieurs champs sémantiques dans cette épreuve. Les critères de cotation sont 0 ou 1. Nous lisons l'ensemble des consignes de correction (0 ou 1) et les consignes d'administration dites à l'enfant (« Comment ça s'appelle là ? » et « Maintenant, tu vas voir des images. Pour chaque image, tu me diras ce que c'est. Tu regardes bien chaque image et tu me dis ce que c'est ») dans le manuel. De plus, l'objectif des épreuves (apprécier le stock lexical), les tendances centrales, les informations sur l'échantillon sélectionné sont relatés. L'épreuve ne propose pas de critère d'arrêt mais une aide phonémique (voyelle ou consonne qui débute le mot). Un gain sera calculé avec ces ébauches phonémiques éventuelles.

Les normes datent de 2014 et présentent un échantillon mixte (féminin/masculin).

Une épreuve de désignation est également présente et est réservée aux enfants de 2<sup>e</sup> et 3<sup>e</sup> primaires. Les auteurs ayant utilisé des items facilement imageables expliquent que les étalonnages au-delà du CE2 présenteraient un risque d'effet plafond. C'est pourquoi ils n'ont pas réalisé de normes après le CE2. Les items sont les mêmes que pour la dénomination. Ils

sont sélectionnés selon l'année scolaire et la fréquence. Précisément, l'enfant doit désigner l'item cible parmi 9 images. Des distracteurs phonologiques ou sémantiques sont utilisés dans ce subtest. Effectivement, le test propose un distracteur phonologique et sémantique par item cible. Enfin, le score est calculé informatiquement, l'épreuve est chronométrée et ne présente pas de critère d'arrêt. De plus, les auteurs contrôlent la validité théorique et expliquent l'objectif de l'épreuve (observer le stock lexical passif).

Les consignes d'administration sont les suivantes : « Tu vas voir des images et entendre des mots. Tu dois cliquer sur l'image qui correspond au mot que tu as entendu ». La correction se fait automatiquement par le logiciel. Nous obtenons le résultat relatif à la justesse des réponses mais aussi le temps de désignation.

Cette batterie contient des normes plus récentes mais elle ne présente que très peu de critères psychométriques dans son manuel pour permettre une évaluation de qualité. Or, celles-ci assurent un meilleur diagnostic (McCauley & Swisher, 1984). Nous avons la présence d'une validité théorique. Cependant, la batterie ne présente pas d'informations quant à la validité concourante, la validité prédictive, la fidélité test-retest et inter-juges, la consistance interne, la structure interne et la distribution des données. De plus, l'erreur standard de mesure n'est pas calculable. Or, ces dernières sont importantes pour un bon diagnostic (Cattini & Lafay, 2024; Youngstrom et al., 2017).

#### NEE-L (forme G) (Chevrié-Müller, C. & Plaza, M., 2001)

Au vu de la population de notre travail, nous avons décidé de discuter de la forme « G » (grand) de ce test et non de la version « P » (petit).

Dans cette partie (forme G) de la batterie, l'échantillon se compose de 541 sujets de 5 ;7 ans à 8 ;7 ans. Elle présente les caractéristiques psychométriques suivantes : fidélité test-retest et le calcul de l'erreur standard. Ce dernier est une mesure statistique qui informe de la variation et de la part d'incertitude des résultats obtenus à partir d'un test. En revanche, elle ne présente pas d'informations par rapport à la validité (discriminante, prédictive) et la fidélité inter-juges et la consistance interne. La distribution des données n'est pas relatée, tout comme la sensibilité au changement. Les normes datent de 2001. Cependant, des moyennes et écart-types nous sont donnés pour quantifier les capacités du patient. Ceci est moins précis que des percentiles au vu de l'absence de normalité.

Deux épreuves évaluant le lexique se retrouvent dans la NEE-L. Une épreuve de dénomination lexicale est envisagée avant une épreuve de désignation. Lors de la dénomination, l'enfant doit dénommer 57 images (mots concrets et abstraits, des couleurs et des formes géométriques), sans critère d'arrêt. Une ébauche phonémique est proposée afin d'avoir un indice quant à l'accès au stock lexical. Nous proposons le phonème initial au patient et nous pouvons obtenir un score concernant le gain obtenu grâce à ce dernier. Un commentaire peut également aider l'enfant pour certains items moins évidents visuellement. Enfin, la cotation est 0, 1 ou 2 selon les consignes de correction. La note de 2 est attribuée lorsque le mot attendu est produit correctement, la note de 1 est attribuée lorsque le mot qui est produit contient une légère altération phonologique (maintien du nombre de syllabes dans les items de plus de deux syllabes et qui sont reconnaissables ou encore des difficultés phonologiques telles que les groupes consonantiques comme « brouette ») et la note de 0 est attribuée lorsque le mot n'est pas produit correctement. Les deux épreuves utilisent les mêmes items, ce qui permet une comparaison entre les deux versants (Bragard et al., 2010). Les fréquences des items varient entre 42,19 et 65,73, ce qui montre que le test comprend des items de faible fréquence, de fréquence moyenne et de haute fréquence. La fréquence des items est contrôlée dans ces subtests. Ceci est un point positif de la batterie car nous pouvons comparer les performances de l'enfant selon ce critère (Ferrand, 1997).

L'échantillon (mixte : féminin/masculin) est de 324 (108 par tranche d'âge). Les moyennes d'âge sont respectivement de 6,06 ans, 7,04 ans, et 8,08 ans. Le score maximal est de 72 pour la première partie et 42 pour la seconde. La distribution ne suit pas une normalité et la fidélité test-retest n'est pas satisfaisante selon la littérature, ( $r=0.9$ ) (Friberg, 2010). Effectivement, elle varie entre 0,72 et 0,8 ( $r=0.74$ ). Enfin, un effet plafond est remarqué pour les étalonnages de cette deuxième partie :

- 5,7-6,6 ans : 37,15 (3,46)
- 6,7-7,6 ans : 37,98 (3,12)
- 7,7-8,7 ans : 39,74 (2,43)

L'épreuve de désignation reprend les mêmes items. L'enfant doit choisir la cible parmi 8 images. Nous retrouvons des distracteurs phonologiques et sémantiques. Six cibles sont à désigner sur une même planche (ceci diminue la sensibilité des performances car l'enfant peut travailler par élimination) et deux distracteurs sont insérés dans cette dernière. Un distracteur

sera phonologiquement proche d'une cible (ex : 'baignoire' comme distracteur de 'lavabo') et un autre sera phonologiquement proche (ex : 'radiateur' comme distracteur de 'aspirateur'). Tous les items n'ont donc pas de distracteurs pour cette épreuve, ceci réduit la possibilité d'erreurs si l'enfant n'a pas de représentations fines de la cible. La cotation est de 0 ou 1 et les images sont toutes en couleur. L'épreuve se découpe en 2 parties (36 items concrets et 21 items plus abstraits, des couleurs ou encore des formes géométriques). Cependant, aucune variabilité grammaticale n'est observée dans l'épreuve, comme pour la dénomination. Nous ne pouvons donner aucune aide ou arrêter le test avant la fin. Le manuel nous explique que cette épreuve permet d'observer le développement du stock lexical passif. Enfin, nous avons des consignes précises de passation (« Montre-moi ... ») ainsi que des informations sur l'échantillon (identique à l'épreuve de dénomination ci-dessus) et les tendances centrales. Pour la première partie, le score maximal est de 36. Nous remarquons donc un effet plafond pour les étalonnages proposés :

- 5,7-6,6 ans : 34,68 (1,86)
- 6,7-7,6 ans : 35,2 (1,18)
- 7,7-8,7 ans : 35,59 (0,96)

La fidélité test-retest est meilleure ( $r=0.89$ ) et atteint presque les recommandations littéraires ( $r=0.9$ ).

Pour la seconde partie, nous obtenons une fidélité test-retest insatisfaisante ( $r=0.53$ ), un score maximal de 21 et donc un effet plafond pour tous les étalonnages :

- 5,7-6,6 ans : 19,78(1,28)
- 6,7-7,6 ans : 20,08 (1,15)
- 7,7-8,7 ans : 20,6 (0,81)

C'est donc pour cette raison mais aussi parce que les normes ne sont pas récentes, qu'il est opportun de sélectionner un autre test (d'en valider et normer un nouveau dans le cadre de ce travail) pour observer le vocabulaire des enfants. De plus, la littérature recommande une corrélation de 0,9 pour la fidélité test-retest (Friberg, 2010). Ceci est donc insuffisant selon la littérature pour l'ensemble des parties. L'erreur standard de mesure est calculable et nous avons une présence de validité théorique. En revanche, la batterie ne présente pas d'informations par rapport à la validité concourante, la validité prédictive, la sensibilité et la spécificité, la consistance interne, la fidélité inter-juges, la structure interne et la distribution des données. Cependant, ces données sont importantes si nous voulons diagnostiquer un trouble du langage (Cattini & Lafay, 2024; Youngstrom et al., 2017).

### EXALANG 5-8 (Helloin et al., 2010)

L'Exalang 5-8 ans a un échantillon de 300 sujets allant de la fin de la 3<sup>e</sup> maternelle à la fin de la 2<sup>e</sup> primaire. Au niveau des tests, il contient une dénomination lexicale : 44 noms sont administrés à l'enfant. À nouveau, la diversité grammaticale n'est pas prise en compte. Concernant la fréquence des items, elle se situe entre 50,38 et 69,79, ce qui démontre que les items sont de fréquence moyenne et de haute fréquence. Ce test ne contient pas de mots de faible fréquence, ce qui aurait été un atout pour observer un éventuel effet (Ferrand, 1997). La cotation est de 0 ou 1. Il s'agit de l'examineur qui décide si l'item a correctement été dénommé ou non et ensuite le logiciel additionne les points encodés par l'examineur.

Le manuel de l'EXALANG 5-8 expose les critères de fidélité suivants : une fidélité test-retest insatisfaisante ( $r=0.8$ ) car une corrélation de 0,9 est recommandée et une erreur standard de mesure calculable ( $\geq .80$ ). De plus, la validité théorique est présente. Même si la distribution des données n'est pas spécifiée, les auteurs fournissent des éléments d'interprétation. Cependant, il n'y a aucune donnée disponible concernant la validité prédictive, la validité concourante, la structure interne, la sensibilité et la spécificité, la fidélité inter-juges et la consistance interne.

Pour les enfants de 3<sup>e</sup> maternelle, l'échantillon est composé de garçons et de filles, ils sont au nombre de 95 à avoir été testés. L'âge moyen est de 6.1 ans et la distribution suit une normalité. Pour les enfants de 1<sup>ère</sup> primaire, l'échantillon est composé de garçons et de filles, ils sont au nombre de 81 à avoir été testés. L'âge moyen est de 7.2 ans et la distribution ne suit pas une normalité.

Et pour les enfants de 2<sup>e</sup> primaire, l'échantillon est composé de garçons et de filles, ils sont au nombre de 121 à avoir été testés. L'âge moyen est de 8.2 ans et la distribution suit une normalité.

Les normes datent de 2010, ce qui la rend relativement ancienne. De plus, dans l'idéal, il aurait été intéressant qu'il y ait une tâche de désignation qui reprenne les mêmes items que dans la tâche de dénomination.

### EVALEO 6-15 (Launay, L. et al., 2018)

Cette batterie contient un échantillon allant de la 1<sup>ère</sup> primaire à la 3<sup>e</sup> secondaire. Elle est informatisée et propose divers tests évaluant le lexique :

- Une dénomination répartie selon des tranches d'âge : le nombre d'items à dénommer est différent selon les tranches d'âge. Ces items sont issus de plusieurs champs sémantiques et de différentes catégories grammaticales. Une ébauche est envisagée lorsque les mots sont mal dénommés, il faut donner le 1<sup>er</sup> phonème afin de voir si l'ébauche aide l'enfant. Une répétition est aussi envisagée lorsque les items sont mal produits phonologiquement. Au niveau de la cotation, le patient obtiendra 2 points en cas de réponse correcte et 1 point s'il a eu recours à un indice ou s'il a dû répéter.
- Une désignation : l'enfant doit désigner la cible parmi 4 images. Les distracteurs sont sémantiquement proches. Les auteurs ne fournissent pas davantage de précisions sur les distracteurs sémantiques, bien que le choix des distracteurs soit un élément crucial. Nous distinguons trois types de distracteurs : phonologiques, sémantiques et neutres, chacun jouant un rôle essentiel dans l'analyse des performances. Tout comme pour la tâche de dénomination, le nombre d'items à dénommer est différent selon les tranches d'âge et les items sont également issus de plusieurs champs sémantiques et de différentes catégories grammaticales.

Concernant la fréquence des items pour les niveaux CP, CE1 et CE2, elle se situe entre 47,94 et 67,06. Pour ceux de la CM1 à la 6<sup>e</sup>, elle se situe entre 36,60 et 62,46. Plus particulièrement, pour les niveaux CP, CE1 et CE2, il s'agit d'items de fréquence moyenne et de haute fréquence. Alors que pour les niveaux de la CM1 à la 6<sup>e</sup>, il s'agit d'items de faible fréquence et de fréquence moyenne. Nous n'observons donc pas les 3 types de fréquence et donc cela pourrait entraver l'observation de l'effet de fréquence (Ferrand, 1997).

Cette batterie contient également des épreuves de fluence sémantique et phonologique, un subtest d'antonymie ainsi que des tests évaluant la morphologie dérivationnelle mais nous n'y reviendrons pas non plus car ces épreuves ne s'accordent pas avec notre objectif qui est de pouvoir diagnostiquer et évaluer le lexique des enfants.

Le manuel de l'Evaléo 6-15 met en avant les aspects suivants concernant la fidélité : une consistance interne inadéquate (Alpha de Cronbach :  $\alpha=0.78$ ), étant donné que nous l'avons dit adéquate si elle était égale ou supérieure à 0.8 et une erreur standard de mesure quantifiable. De plus, la structure interne et la validité théorique sont exposées. La distribution des données

n'est pas normale. Toutefois, aucune indication n'est donnée par rapport à la validité prédictive, la validité concourante, la sensibilité et la spécificité, ainsi que la fidélité test-retest et la fidélité inter-juges. Le groupe sujet comprend 1321 participants, avec une représentation tant masculine que féminine.

Bien que cette batterie propose une tâche de dénomination et de désignation, elle ne reprend pas la totalité des mêmes items dans chacune des épreuves. De plus, elle ne présente pas le même nombre d'items entre l'épreuve de dénomination et de désignation alors que cela pourrait permettre une meilleure comparaison des performances dans les deux versants.

#### EVALO 2-6 (Coquet F. et al., 2009)

Cette batterie contient un échantillon de 880 sujets se situant entre 2 ;3 ans et 6 ;3 ans. Bien que nous ne connaissions pas le nombre exact d'enfants par tranches d'âge, nous pouvons supposer que chaque tranche d'âge comprend environ 93 enfants. Concernant la fréquence des items, elle se trouve entre 42,53 et 71,64. Cette batterie contient donc des items de faible fréquence, de fréquence moyenne et de haute fréquence. Il s'agit d'un point positif de la batterie pour objectiver les capacités de l'enfant selon la fréquence (Ferrand, 1997). Elle est divisée selon l'âge et propose plusieurs tâches dont :

- Une tâche de désignation lexicale à partir d'un mot : l'enfant doit désigner un item parmi 6 images. Les 57 items sont issus de plusieurs catégories sémantiques. Enfin, les distracteurs sont au nombre de 5 et sont uniquement des distracteurs sémantiques. La cotation est de 0 ou 1 point.
- Une tâche de dénomination phonologie/lexique : l'enfant voit des images issus de différentes catégories grammaticales et doit dénommer l'item, l'action ou la couleur. Il est coté pour le niveau lexical et pour le niveau phonologique.

Cette batterie contient également une tâche de mise en réseaux ainsi que des tests de compréhension de termes topologiques ou de qualificatifs mais nous ne nous attarderons pas dessus car ces tâches ne concordent pas avec notre objectif qui est de diagnostiquer un déficit de stock lexical (épreuves de 1<sup>ère</sup> ligne) dans un TDL.

L'Evalo 2-6 présente une fidélité inter-juges correcte et permet de calculer l'erreur standard de mesure. De plus, le manuel met en avant la validité théorique. Aucune information n'est apportée quant à la relation avec d'autres variables, la validité prédictive, la validité



concourante, la structure interne, la sensibilité et spécificité, la fidélité test-retest, la consistance interne et la distribution des données.

Au niveau de l'épreuve de désignation et de dénomination :

- 2,3 – 2,9 ans : l'échantillon est composé de 47 hommes et femmes contrôles dont l'âge moyen est de 2,6 ans. La distribution ne suit pas une normalité et il n'y a pas de consensus entre les juges.
- 2,9 – 3,3 ans : l'échantillon est composé de 68 hommes et femmes contrôles dont l'âge moyen est de 3,0 ans. La distribution ne suit pas une normalité et il n'y a pas de consensus entre les juges.
- 3,3 – 3,9 ans : l'échantillon est composé de 73 hommes et femmes contrôles dont l'âge moyen est de 3,6 ans. La distribution ne suit pas une normalité et il n'y a pas de consensus entre les juges.
- 3,9 – 4,3 ans : l'échantillon est composé de 90 hommes et femmes contrôles dont l'âge moyen est de 4,0 ans. La distribution ne suit pas une normalité et il n'y a pas de consensus entre les juges.
- 4,3 – 4,9 ans : l'échantillon est composé de 96 hommes et femmes contrôles dont l'âge moyen est de 4,6 ans. La distribution ne suit pas une normalité et il n'y a pas de consensus entre les juges.
- 4,9 – 5,3 ans : l'échantillon est composé de 100 hommes et femmes contrôles dont l'âge moyen est de 5,0 ans. La distribution ne suit pas une normalité et il n'y a pas de consensus entre les juges.
- 5,3 – 5,9 ans : l'échantillon est composé de 125 hommes et femmes contrôles dont l'âge moyen est de 5,6 ans. La distribution ne suit pas une normalité et il n'y a pas de consensus entre les juges.
- 5,9 – 6,3 ans : l'échantillon est composé de 147 hommes et femmes contrôles dont l'âge moyen est de 6,0 ans. La distribution ne suit pas une normalité et il n'y a pas de consensus entre les juges.

Pour cette batterie, les normes datent de 2009 et ne sont donc pas récentes. De plus, tout comme pour l'Evaleo 6-15, elle ne reprend pas la totalité des mêmes items dans chacune des épreuves et elle ne présente pas le même nombre d'items entre l'épreuve de dénomination et de désignation.

### EVT-3 (Williams, 2023)

Ce test standardisé contient 188 items (dont 63 items en commun avec le PPVT-5) pour évaluer la production lexicale des enfants de 2 ;6 ans à 18 ;11 ans. Il s'organise en ordre croissant de difficultés. Au niveau de la fréquence des items, elle se situe entre 19,61 et 71,64 (pour les enfants de 2<sup>e</sup> et 5<sup>e</sup> primaires). Ce test contient un large éventail d'items étant donné qu'il y a des items de faible et moyenne fréquence ainsi que quelques items de haute fréquence. Ceci est un atout pour observer un éventuel effet de fréquence (Ferrand, 1997). Cependant, nous notons le peu d'items ayant une fréquence élevée (>65) (Peereman et al., 2007). L'échantillon de la batterie est de 803. Plus précisément, la répartition de l'échantillon est le suivant :

- 554 sujets contrôles pour la tranche d'âge de 2 à 10 ans ;
- 28 sujets contrôles pour la tranche d'âge de 9 à 11 ans ;
- 249 sujets contrôles pour la tranche d'âge de 11 à 18 ans.

Elle contient plusieurs catégories grammaticales dont des verbes, des substantifs et des adjectifs. L'enfant observe une représentation imagée et l'examineur lui pose une question (soit « Qu'est-ce que c'est ? », soit « Qu'est-ce qu'il fait ? » ou « Dis-moi un autre mot pour ... »). L'épreuve évalue donc la production d'items précis à dénommer ou la production de synonymes. L'administration de ce test est d'environ 25/30 minutes. On l'utilise dans le cadre du dépistage du trouble d'accès lexical actif, pour évaluer l'évolution du patient au cours de la prise en charge, dans la recherche ou encore, conjointement avec le PPVT-5, l'obtention d'une vue globale des capacités lexicales de l'enfant/adolescent.

Cette batterie ne suit pas une distribution normale. Nous avons des percentiles à notre disposition pour évaluer les performances du patient. De plus, la fidélité test-retest est de 0,88, ce qui ne suit pas la recommandation littéraire (0.9) (Friberg, 2010). En revanche, l'erreur standard est calculable.

Au niveau des éléments importants pour un diagnostic, l'EVT-3 contient des informations quant à la validité concourante, la validité théorique, la distribution des normes (non-normale) et l'erreur standard (calculable). En revanche, nous n'avons pas d'informations par rapport à la sensibilité et la spécificité, la consistance interne, la validité prédictive, la structure interne et la fidélité inter-juges.

Enfin, alors que les normes sont récentes et que le test contient un nombre élevé d'items, nous remarquons que le test PPVT-5 ne contient que 63 items en commun. Dès lors, la comparaison des capacités du patient selon les versants langagiers est plus compliquée. De plus, la passation est chronophage au vu du nombre d'items.

### PPVT-5 (Dunn, 2023)

Cette batterie repose sur un échantillon de 803 enfants et adolescents. Il s'agit d'un test standardisé, administré individuellement, visant à évaluer le vocabulaire réceptif chez les enfants et adolescents âgés de 2 ans et 6 mois à 18 ans et 11 mois. Plus précisément, il mesure l'étendue du lexique passif en fonction de différentes catégories grammaticales, notamment les substantifs, les verbes et les adjectifs attributs.

Le PPVT-5 comprend un total de 239 items, dont 63 sont communs avec l'EVT-3, organisés selon un ordre de difficulté croissante. Chaque item se compose d'un mot-stimulus accompagné de quatre images, parmi lesquelles une correspondance correcte et trois distracteurs (phonologique, sémantique et visuel). Cette conception permet d'analyser les types d'erreurs lexicales en réception. Les items (pour les enfants de 2<sup>e</sup> et 5<sup>e</sup> primaires) ont une fréquence entre 19,4 et 71,64, de même que pour l'EVT-3, et les distracteurs sont similaires au niveau de celui-ci. L'administration du test dure environ 20 minutes.

Un atout majeur du PPVT-5 est la possibilité de l'associer à l'EVT-3 afin d'obtenir une évaluation complète des connaissances lexicales des participants. Il peut être utilisé dans divers contextes, notamment pour mesurer le vocabulaire passif dans le cadre de la recherche, effectuer un dépistage, en tant que batterie diagnostique ou encore réaliser un re-test afin d'évaluer la pertinence d'une prise en charge thérapeutique.

Le PPVT-5 présente une fidélité test-retest insatisfaisante selon la littérature ( $r=0.72$ ) alors que c'est recommandé à partir de 0.9, et son erreur standard de mesure est calculable. La distribution des données est non normale. De plus, le manuel met en avant les informations concernant la validité concourante et théorique. Toutefois, aucune information n'est disponible concernant la validité prédictive, la validité discriminante (sensibilité et spécificité), la structure interne, la fidélité inter-juges ainsi que la consistance interne.

En ce qui concerne l'échantillonnage, on observe :

- 554 sujets contrôles pour la tranche d'âge de 2 à 10 ans ;
- 28 sujets contrôles pour la tranche d'âge de 9 à 11 ans ;
- 249 sujets contrôles pour la tranche d'âge de 11 à 18 ans.

Bien que cette batterie soit récente et partage certains items avec l'EVT-3, elle comporte un nombre important d'items, ce qui allonge considérablement la durée de passation. De plus, les items du PPVT-5 et de l'EVT-3 ne sont pas strictement identiques.

L'annexe 1 montre un résumé des batteries analysées selon notre objectif de diagnostic du trouble lexical.

## Objectifs et hypothèses

Notre mémoire poursuit un objectif principal qui est de valider et normaliser une tâche de désignation lexicale et une tâche de dénomination lexicale dans le cadre du diagnostic du TDL. Par rapport aux tâches existantes, ces tâches nouvellement créées présentent les avantages suivants :

- normes récentes reprenant les caractéristiques culturelles, socio-économiques et géographiques de l'échantillon pour permettre une meilleure comparaison du patient avec celui-ci (Nicollet et al., 2009).
- items similaires en production et en compréhension pour permettre une meilleure comparaison des performances dans les deux versants (Bragard et al., 2010)
- test informatisé : cela permet une meilleure fidélité et validité (Csapó et al., 2014)

En validant les deux nouvelles épreuves, nous tenterons d'objectiver leurs capacités à évaluer le lexique des élèves de 2<sup>e</sup> et 5<sup>e</sup> primaires. Nous vérifierons certaines caractéristiques psychométriques pour leur construction : la consistance interne, la validité concourante et la fidélité inter-juges.

Ensuite, nous les normerons pour, ultérieurement, pouvoir comparer les performances d'enfants ayant un TDL à des enfants contrôles. Ceci fera l'objet de la thèse de Pauline Menjot.

Notre hypothèse principale est que les épreuves expérimentales respecteront les caractéristiques psychométriques contrôlées pour valider les épreuves expérimentales à savoir la validité

(validité concourante) et la fidélité inter-juges et la consistance interne. En effet, nous nous attendons à ce que ces deux épreuves expérimentales fournissent à l'utilisateur la confiance que le score obtenu par le participant reflète fidèlement son niveau de compétence, sans être influencé par des facteurs externes ; qu'elles permettent de comparer les participants à d'autres enfants du même âge, afin d'identifier un éventuel déficit.

Plus précisément, nos hypothèses principales, pour l'échantillon de 2<sup>e</sup> primaire et celui de 5<sup>e</sup> primaire respectivement, sont les suivantes :

- En ce qui concerne la consistance interne : si chaque épreuve mesure bien un seul et même concept (le vocabulaire en compréhension d'une part, en production d'autre part), nous nous attendons à observer une bonne homogénéité dans les scores obtenus à chacun des items. Ceci sera vérifié avec le score KR-20 vu les données dichotomiques. Si celui-ci est égal ou supérieur à 0.7, nous concluons que chacune des tâches a une consistance interne acceptable et si celui-ci est égal ou supérieur à 0.8, nous concluons que chacune des tâches a une bonne consistance interne (Ntumi et al., 2023).
- Pour ce qui est de la validité concourante : si nos épreuves expérimentales mesurent bien ce qu'elles sont censées mesurer, les résultats obtenus à ces épreuves expérimentales seront comparables aux résultats obtenus à des épreuves standardisées évaluant les mêmes composantes et ayant déjà été validées. Pour vérifier cette hypothèse, nous calculerons la corrélation entre les scores obtenus au PPVT-5 et à la tâche expérimentale de désignation d'une part, ainsi qu'entre les scores obtenus à l'EVT-3 et à la tâche de dénomination d'autre part. Plus celle-ci s'approchera d'un, plus il y aura une bonne corrélation et plus les tâches expérimentales auront une bonne validité concourante. Une corrélation égale ou supérieure à 0.7 est souhaitable (Souza et al., 2017). Les tests de l'EVT-3 et PPVT-5 étant non normalement distribués, nous utiliserons le test de corrélation de Spearman.
- Concernant la fidélité inter-juges : si nos épreuves permettent de donner des scores fiables, quel que soit l'évaluateur qui les administre, nous nous attendons à ce qu'elles présentent de bons niveaux de fidélité inter-juges. Pour cela, deux juges réévalueront une deuxième fois 20% des audios des passations des enfants de 2<sup>e</sup> primaire et de 5<sup>e</sup> primaire et nous objectiverons ce type de fidélité. Le taux de 20% a été choisi car c'est ce qui est fréquemment employé dans la littérature comme étant représentatif de l'échantillon total (Graham, 2012). Ces audios seront analysés par un évaluateur (Madame Pauline Menjot). Ainsi, on comparera l'évaluation d'un autre expérimentateur

sur une même passation. Le score obtenu par celui-ci sera comparé au nôtre. Nous allons comparer les scores obtenus item par item via le coefficient de Kappa, que nous jugerons acceptable si celui-ci est supérieur à 0.6 (Landis & Koch, 1977). Nous n'évaluerons cette fidélité que pour l'épreuve de dénomination qui est sujette à davantage de subjectivité de l'expérimentateur par rapport à la désignation. En effet, lors d'une tâche de désignation, la réponse attendue est unique et identifiable sans ambiguïté. De plus, il s'agit d'une cotation binaire qui ne laisse place à aucune subjectivité (Landis & Koch, 1977).

Ensuite, concernant la normalisation, nous vérifierons que nos données récoltées ne présentent pas d'effet plancher ou plafond. Nous recueillerons des moyennes et écarts-types, des percentiles et des étendues. De plus, nous contrôlerons la distribution normale des données. Concernant cette dernière, nous réaliserons un test de Shapiro-wilk pour chaque épreuve et chaque année scolaire de notre étude.

Secondairement, nous supposons que les enfants, en 2<sup>e</sup> primaire comme en 5<sup>e</sup>, auront plus de facilités lors de la tâche de compréhension lexicale que lors de la tâche de production lexicale. Étant donné que le lexique passif (compris) est plus élevé que le lexique produit chez des enfants tout-venant, la tâche de désignation lexicale doit être mieux réussie que celle de dénomination (Daviault, 2011).

De plus, nous estimons que les enfants vont davantage commettre des erreurs de type sémantique en 5<sup>e</sup> primaire et réaliser plus d'erreurs de type phonologique en 2<sup>e</sup> primaire (Bragard et al., 2010). Des précédentes études ont montré que les enfants commettent de plus en plus d'erreurs sémantiquement proches de la cible avec l'âge. Effectivement, des enfants de 5 ans commettent davantage d'erreurs phonologiques alors que des enfants de 8 ans réalisent plus d'erreurs sémantiques (Sheng & McGregor, 2010).

En revanche, nous ne pourrons pas comparer davantage les données obtenues par les participants des deux niveaux scolaires, car nous n'utilisons pas les mêmes items pour nos tâches expérimentales en 2<sup>e</sup> primaire et en 5<sup>e</sup> primaire.

## Méthodologie

### Comité d'éthique

Nous avons introduit notre demande au comité d'éthique en début d'année scolaire 2023-2024. Cependant, ce mémoire s'inclut dans un plus large projet de thèse de doctorat. Une demande antérieure avait donc déjà été introduite et acceptée. Dès lors, notre requête a également reçu un retour positif.

### Participants

Nous avons recruté un total de 52 élèves de 2<sup>e</sup> primaire, comprenant 26 filles et 26 garçons. L'âge moyen des participants est de 7 ; 6 ans, avec une amplitude allant de 6 ; 10 ans à 8 ; 6 ans. De plus, dans cet échantillon, environ 11,5% des enfants sont bilingues. Les enfants devaient avoir le français comme langue majoritaire et ne devaient pas être dans une école immersive ou dans un enseignement non-francophone. En ce qui concerne les enfants de 5<sup>e</sup> primaire, ils sont au nombre de 51 (26 filles et 25 garçons). Ils ont en moyenne 10 ; 5 ans, avec une amplitude de 9 ; 11 à 11 ; 4 ans. Au sein de l'échantillon des enfants de 5<sup>e</sup> primaire, environ 1,96% des enfants sont bilingues. Ils ont été recrutés via des écoles, les réseaux sociaux et le bouche à oreilles. Une fois le formulaire d'informations et de consentements signés par les tuteurs légaux de l'enfant et les enfants de plus de 8 ans, une anamnèse a permis de vérifier les critères d'inclusion de l'étude, à savoir :

- Être en 2<sup>e</sup> ou 5<sup>e</sup> primaire
- Suivre sa scolarité dans un enseignement francophone
- Parler au moins le français à domicile (monolinguisme ou multilinguisme)
- Ne pas présenter un trouble neurodéveloppemental de type dyslexie/dysorthographe, trouble de l'attention avec hyperactivité (TDAH), trouble développemental du langage (TDL)
- Ne pas présenter un trouble sensoriel de type surdité, déficience visuelle non traitée
- Ne pas avoir doublé ou sauté une année scolaire

Afin de vérifier que les enfants testés ne présentent pas de trouble langagier non détecté, nous leur avons administré un ensemble de tests langagiers normés. La phonologie a été évaluée à l'aide de la répétition de non-mots (en production) et l'ELDP (en réception) ; la morphosyntaxe

a été évaluée à l'aide de la répétition de phrases ; le lexique a été évalué à l'aide de l'EVT-3 et du PPVT-5. En ce qui concerne les scores seuils d'exclusion utilisés lors de ces évaluations, plusieurs auteurs (Hamann & Abed Ibrahim, 2017; Schwob, 2020) ont suggéré les critères suivants :

- Utiliser le critère de -1,25 ET pour les enfants monolingues TDL
- Utiliser celui de -1,5 ET pour les enfants bilingues TDL dans leur langue dominante et -2,25 ET dans leur langue minoritaire
- Utiliser -1,75 ET pour les enfants bilingues TDL en contexte de bilinguisme équilibré

Thordardottir (2011) ajoute que le seuil de -2 ET semble trop peu sensible et donc trop spécifique. Le seuil de -1,28 ET correspond davantage à la combinaison d'une bonne sensibilité et d'une bonne spécificité.

Au préalable, au vu de notre recrutement (enfants monolingues et bilingues), nous avons décidé de prendre les scores seuils suivants (en fonction des normes dont nous disposons dans les tests standardisés sélectionnés) : -1,5 écart-type (ET) ou le percentile 6, la note standard de 5 ou encore la classe 1. Nous avons décidé de prendre -1,5 ET au lieu de -1,28 ET, car lors des testings, nous nous sommes rendues compte que nous allions tester des enfants monolingues, mais également bilingues et qu'il était plus intéressant d'avoir une plus grande marge d'erreur. Nous avons donc décidé d'exclure de notre échantillon les enfants qui présentaient des scores inférieurs à ces scores seuils afin d'écarter la présence d'un trouble non détecté.

Les tâches concernées par ces scores seuils d'exclusion sont la répétition de non-mots, l'ELDP, la répétition de phrases, l'EVT-3 et le PPVT-5.

## Outils utilisés

### Les tests standardisés

Afin de vérifier que les enfants sélectionnés se situent dans la norme au niveau du développement du langage, nous allons leur administrer des tâches standardisées :

- L'épreuve de répétition de pseudomots de l'Évaléo 6-15 (Launay et al., 2018) : cette tâche évalue les traitements phonologiques, la mobilisation des représentations phonologiques et la mémoire à court terme phonologique, avec un matériel dépourvu



de sens. Cette répétition requiert également des représentations et des traitements phonologiques. La personne écoute 20 items un à un et doit les répéter directement. Les non-mots respectent les règles phonotactiques du français. Cette épreuve ne présente ni de critère d'arrêt ni d'item d'essai. Enfin, la note maximale est de 20. La note de 1 est accordée si le pseudomot est correctement répété. Cette tâche a notamment été choisie car plusieurs auteurs rappellent que la répétition de pseudomots est une épreuve très discriminante pour diagnostiquer un TDL (Boerma et al., 2015; Fu et al., 2024; Maillart, 2022).

- L'épreuve de répétition de phrases de la CELF-5 (Wiig et al., 2019) : cette épreuve vise à objectiver les capacités du sujet à produire des structures morphosyntaxiques, en répétant des énoncés. Il s'agit d'un outil sensible et spécifique pour l'élaboration du diagnostic de TDL (Rujas et al., 2021). Dans l'épreuve de la CELF 5 (Wiig et al., 2019), le participant doit écouter une phrase énoncée par l'examineur et doit la répéter directement. Une note de 3 sera accordée si le sujet ne commet aucune erreur de répétition, 2 s'il réalise une erreur, 1 s'il commet 2 ou 3 erreurs et 0 s'il commet plus de trois erreurs. Nous arrêtons l'épreuve après 4 notes consécutives égales à 0 (plus de 3 erreurs réalisées lors de la répétition).
- L'Epreuve Lilloise de Discrimination Phonologique (ELDP, Macchi et al., 2012) : Cette épreuve de discrimination auditive permet d'évaluer la capacité de percevoir les phonèmes de la langue. Bogliotti (2012) rappelle que les enfants porteurs de TDL ont une capacité limitée concernant la discrimination auditive. C'est pourquoi nous utiliserons l'ELDP (Macchi et al., 2012) pour objectiver cette compétence. Le participant doit comparer des paires de non-mots produits par l'ordinateur et doit dire s'ils sont « pareils » ou « pas pareils ». 1 point est attribué par paire correctement identifiée (identique ou différente). Le subtest comprend 36 non-mots qui varient selon le nombre de syllabes, la position de l'opposition et le type d'opposition.

Ensuite, toujours pour vérifier l'inclusion de nos participants mais aussi pour vérifier la validité concourante des épreuves expérimentales, nous allons proposer :

- Le Peabody Picture Vocabulary Test 5<sup>e</sup> édition (PPVT-5, Dunn, 2023) : Cette épreuve de désignation d'images permet d'apprécier la validité concourante de la tâche

expérimentale de désignation d'images (Friberg, 2010) et d'évaluer l'étendue du stock lexical passif. Lors de cette épreuve, le participant regarde une planche d'items face à lui et désigne l'image qui correspond à l'item énoncé par l'examineur. Pour chaque item, une nouvelle planche est montrée au participant. Un point lui est accordé par bonne image désignée parmi 4. Les distracteurs sont : un distracteur sémantique, un distracteur phonologique et un distracteur visuel. Le score maximum possible est de 239. De plus, il bénéficie de deux items d'exemple. Enfin, s'il échoue à 7 items consécutifs, nous arrêtons l'épreuve. Dès lors, un score seuil est calculé en comptabilisant le nombre de bonnes désignations.

- Le Test de Vocabulaire actif en images 3<sup>e</sup> édition (EVT-3, Williams, 2023) est une épreuve de dénomination d'images qui permettra de mettre en lumière des déficits éventuels liés à un manque du mot ou à un stock lexical actif trop faible/ trop peu étendu. Cela nous aidera à objectiver la validité concourante du subtest expérimental de dénomination d'images. Le participant doit répondre à des questions relatives à des images qui lui sont proposées, du type « Qu'est-ce que c'est ? » ou « Dis-moi un mot qui veut dire la même chose que « noter », comme dans la phrase « Il est en train de noter » ». Il s'agit donc soit de dénommer, soit de trouver un synonyme. On comptabilise un point si la réponse donnée est présente dans les protocoles, le score maximal possible étant de 188. Nous arrêtons l'épreuve après 7 erreurs consécutives.

Enfin, pour avoir davantage d'informations quant au fonctionnement cognitif des enfants sélectionnés, nous administrerons trois épreuves. Celles-ci ont pour but d'apprécier leur raisonnement non-verbal et leur mémoire à court terme verbale pour évaluer les fonctions cognitives indispensables à la compréhension et à la gestion de l'information, indépendamment du langage. Effectivement, Bussy et al (2014) relatent l'importance d'une mémoire à court terme verbale efficiente pour le développement et l'apprentissage lexical. En revanche, un score trop faible dans ces différentes tâches n'exclut pas le participant. En effet, une mémoire verbale à court terme ou un raisonnement non-verbal trop faible n'induit pas forcément que le participant n'est pas considéré comme « tout-venant ». Or, nous recherchons bien ce type d'individus.

- L'épreuve de répétition de chiffres (à l'endroit) de l'Evaléo 6-15 (Launay et al., 2018) : cette tâche permet d'évaluer les capacités de l'enfant concernant la mémoire verbale à

court terme. En effet, la boucle phonologique sert à retenir des informations phonologiques pendant un court instant. Ceci entre en interaction avec d'autres fonctions cognitives comme le langage (Bussy et al., 2014). Dès lors, cette épreuve nous permettra de penser que les enfants sélectionnés présenteront de bonnes capacités concernant ce type de mémoire.

- L'épreuve de répétition de chiffres (à l'envers) de l'Evaléo 6-15 (Launay et al., 2018) : cette tâche consiste à évaluer les capacités de l'enfant à retenir une information auditivo-verbale temporairement ainsi que de réaliser sa manipulation. Nous aurons une appréciation des capacités de l'enfant concernant la rétention et la manipulation d'informations.

Pour ces deux tâches précitées, le sujet écoute une suite de chiffres et doit la restituer dans le même sens (répétition de chiffres à l'endroit) ou en commençant par le dernier chiffre entendu (répétition de chiffres à l'envers). Le participant commence par réaliser deux exemples avant de passer au test. Lors de celui-ci, les séries s'allongent progressivement (allant de 3 chiffres à 8). Il peut avoir 18 points et un empan de 8 au maximum. Effectivement, un point est accordé pour chaque suite correctement répétée. Lorsque le participant commet une erreur aux 3 items de chaque série, nous arrêtons l'épreuve.

- Le subtest des matrices de l'échelle WNV (Wechsler & Naglieri, 2009) est utilisé afin d'apprécier le raisonnement non-verbal des sujets. Les capacités cognitives non verbales sont essentielles pour le développement typique du lexique (Leonard et al., 2007). Ceci nous montrera un aperçu des capacités de raisonnement de l'enfant. La tâche du WNV (Wechsler & Naglieri, 2009) demande au participant de compléter une matrice (complétion de figures) avec le bon élément parmi 5. Nous proposons 4 exemples avant le test. La note maximale est de 41 et nous arrêtons l'épreuve après 4 erreurs au sein de 5 matrices consécutives.

## Les épreuves expérimentales

Les deux épreuves expérimentales (désignation et dénomination d'images) seront administrées lors de ce testing. L'épreuve de dénomination sera proposée lors de la première séance alors que celle de désignation sera administrée lors de la deuxième (cf. annexe 2 et 3).

Concernant la tâche de dénomination, des images sont présentées une par une sur un écran d'ordinateur. Pour commencer, 3 items d'exemples sont dénommés par le participant. Suite aux exemples, 33 images vont suivre. Le participant doit d'abord dénommer spontanément l'image, s'il répond correctement, il obtient une note de 2 et l'item suivant est présenté. S'il répond erronément, il obtient une note de 0 et nous lui demandons « Connais-tu un autre mot ? ». À nouveau, s'il répond correctement, il obtient une note de 2 dans la catégorie « autre mot » et nous passons à l'image suivante, mais s'il ne répond pas correctement, il obtient une note de 0 dans cette même catégorie et nous lui fournissons une ébauche phonétique en lui disant « Et si je te dis que le mot commence par ... ». Lors de cette dernière aide, soit il trouve le mot et obtient 2 dans la catégorie « ébauche phonétique », soit il ne trouve pas et obtient 0 dans cette même catégorie et nous passons à l'image suivante. Une autre option est également envisageable : dans le cas où, lors de la dénomination spontanée, le participant ne produit pas correctement le mot (il produit une déformation phonologique de la cible attendue). Alors nous lui accordons la note de 1 pour la dénomination spontanée et nous lui demandons de répéter après nous ce même mot. Si la répétition est correcte, il obtient 2 pour la répétition, sinon il a 0. Cette tâche est donc sur 66 points pour la dénomination spontanée, l'autre mot, la répétition et l'ébauche phonétique. De plus, lors de cette tâche, nous chronométrons le temps de réponse pour chaque image. En effet, un temps maximum de dénomination de 10 secondes leur est laissé. Si l'enfant ne dénomme pas l'item dans ce laps de temps, il obtient la note de 0. Et on lui demande alors la question « sais-tu comment ça s'appelle ? ».

À propos de la tâche de désignation, 5 images sont affichées simultanément sur l'écran d'ordinateur et le participant doit désigner l'image correspondant au mot qu'il a entendu à travers le casque parmi des distracteurs. Plus précisément, il y aura 2 distracteurs sémantiques (qui partagent une similarité ou une relation sémantique avec le mot cible), un distracteur phonologique qui aura au minimum un phonème ou une syllabe en commun avec le mot cible qui est de structure courte, et un distracteur neutre qui est conçu pour ne pas perturber le processus de désignation, qui ne présente aucun lien sémantique ni phonologique et qui est une image qui n'a rien à voir avec les items à identifier. Plus précisément, il sert à détourner

momentanément l'attention pour tester la capacité de l'enfant à se concentrer sur la tâche principale. Tout comme lors de la tâche de désignation, le participant commence par 3 items d'exemples. Ensuite, le participant doit désigner 33 items. Cette tâche, de même que pour la tâche de dénomination, permet également de chronométrer le temps de réponse pour chaque item. Chaque image correctement désignée, nous lui attribuons la note de 1 et s'il se trompe, nous lui attribuons la note de 0 et nous passons à l'image suivante. Cette tâche est donc sur 33 points et nous devons, en cas d'erreurs, noter combien de distracteurs phonologiques, sémantiques ou neutres, le participant a désigné.

Plus précisément, les recommandations seront suivies pour ces tâches. Dès lors, les items seront tous contrôlés selon l'imageabilité, l'âge d'acquisition, la fréquence, la complexité syllabique et la complexité phonologique (Bragard et al., 2010; Manoiloff et al., 2013).

Notons que les images en désignation et en dénomination ont été sélectionnées selon plusieurs critères. Notamment l'âge d'acquisition, c'est-à-dire l'âge auquel des adultes pensent avoir acquis ce mot. Pour les 2<sup>e</sup> primaires, il se situait entre 6,99 et 9,06 ans et pour les 5<sup>e</sup> primaire, entre 9,10 et 15,45 ans.

Pour la fréquence, c'est-à-dire la fréquence à laquelle l'enfant est exposé au mot, elle se situe entre 35,89 et 60,22 pour les 2<sup>e</sup> primaire. Quant aux enfants de 5<sup>e</sup> primaire, elle se situe entre 20,84 et 51,16. Chaque fréquence a été vérifiée via manulex. Nous avons pris en compte l'indice SFI afin de pouvoir obtenir la fréquence des items. Cet indice est un indice standardisé, centré autour de 50, afin de permettre de comparer les fréquences des mots entre eux plus facilement. Trois classes de fréquence sont proposées : basse fréquence ( $SFI < 45$ ), fréquence moyenne ( $45 \leq SFI \leq 65$ ) et haute fréquence ( $SFI > 65$ ) (Lété et al., 2004 ; Peereman et al., 2007). De manière générale, la fréquence des items des deux épreuves expérimentales est comparable à celle observée dans les principaux outils francophones d'évaluation du lexique chez l'enfant d'âge scolaire. La fréquence des items se situe dans la classe moyenne ou faible. Ferrand, en 1997, expliquait que les différences de fréquence pouvaient donner des informations quant à un certain effet de cette variable. Étant donné que nous voulions avoir une analyse de nos tâches concernant la consistance interne, nous avons utilisé les mêmes degrés de fréquence que ceux utilisés dans l'EVT-3 et PPVT-5.

En ce qui concerne l'imageabilité (c'est-à-dire si l'image est peu représentative (1) à très représentative (7)), elle se situait entre 4,40 et 6,73 pour les 2<sup>e</sup> primaire et entre 2,50 et 6,77 pour les 5<sup>e</sup> primaire.

Lors de la passation de ces épreuves, nous devons toutes les faire passer de la même manière afin que le score obtenu par le participant ne soit pas influencé par des facteurs externes. En effet, lors de la conception de ces deux nouvelles tâches, la création des consignes d'administration et de cotation a été essentielle afin de minimiser la variabilité inter-juges et intra-juge. Nous allons appliquer les mêmes consignes, les coter de la même manière, régler l'intensité sonore et la luminosité de l'ordinateur de la même façon ainsi que de toujours nous positionner à droite du participant. Concernant l'administration, elle se fait via le logiciel « Psychopy ». Ce logiciel nous permettra de recueillir les temps de réaction et la réponse donnée par l'enfant. Nous contrôlerons également la luminosité de l'écran, l'intensité sonore des stimuli lors de la tâche de désignation, l'expérimentateur se tiendra toujours à droite de l'enfant et il contrôlera la cotation de la tâche ainsi que les consignes. Logistiquement, nous utiliserons un casque filaire pour la désignation expérimentale pour permettre une meilleure standardisation (McCauley et Swisher, 1984) ainsi qu'un enregistreur audio pour les deux tâches expérimentales.

Au sujet des tâches non expérimentales, nous utiliserons un ordinateur afin que l'enfant entende les stimuli et qu'il puisse y répondre de manière adéquate ainsi que des cahiers de passation comme le PPVT-5 et l'EVT-3. La luminosité de l'écran et l'intensité sonore seront également contrôlées. L'expérimentateur se situera à droite de l'enfant et contrôlera la cotation des différentes tâches de même que les consignes. Nous utiliserons un enregistreur audio pour l'ensemble des tâches.

## Procédure

Les passations se sont déroulées lors de la fin du deuxième quadrimestre de l'année scolaire 2023-2024 et lors du premier quadrimestre de l'année scolaire 2024-2025. La passation se déroulait en deux séances espacées de minimum 2 jours et maximum 3 semaines. Elles duraient approximativement 45 minutes chacune.

Les principaux lieux d'expérimentation étaient les écoles, les domiciles des enfants et les salles de testing de l'Université de Liège. Avec l'accord des équipes pédagogiques et des tuteurs légaux, nous avons vu les enfants lors des journées scolaires, soit lors des périodes de cours ou pendant les temps de pause (récréations, temps de midi ou garderie).

Par ailleurs, nous pouvions également voir les enfants dans l'enceinte de l'Université de Liège, selon les convenances des familles.

Les épreuves suivantes étaient administrées lors de la première séance :

- Répétition de chiffres (à l'endroit, Evaleo 6-15)
- Dénomination d'images (tâche expérimentale)
- Répétition de pseudomots (Evaleo 6-15)
- Désignation d'images (PPVT-5)
- Discrimination auditive (ELDP 2)
- Répétition de chiffres (à l'envers, Evaleo 6-15)

Lors de la seconde séance, nous avons administré les subtests suivants :

- Désignation d'images (tâche expérimentale)
- Répétition de phrases (CELF 5)
- Dénomination d'images (EVT-3)
- Raisonnement non-verbal (WNV)

## Résultats

Au vu de nos hypothèses et objectifs, nous avons décidé de scinder les résultats par année scolaire. Dès lors, nous présenterons d'abord nos analyses concernant les enfants de 2<sup>e</sup> primaire et ensuite, ceux de 5<sup>e</sup> primaire. Pour notre dernière hypothèse concernant la comparaison des erreurs phonologiques et sémantiques, nous avons repris les deux années dans un même point.

### 2<sup>e</sup> primaire

#### Hypothèse 1 : Consistance interne

Nous espérions que la réussite aux différents items au sein d'une même tâche soit relativement homogène, c'est-à-dire que les réponses aux différents items d'une même tâche sont bien corrélées. De ce fait, chaque item de nos épreuves expérimentales évalue bien ce qu'il doit évaluer à savoir le vocabulaire en réception ou en production respectivement. Nous obtenons

un score KR-20 (Kuder-Richardson) de 0.74 pour la désignation, ce qui indique une consistance interne jugée acceptable, conformément aux recommandations établies (0.7) (Ntumi et al., 2023). Nous n'avons pas utilisé le test de Cronbach, car celui-ci est recommandé pour les données métriques et non dichotomiques comme c'est le cas dans notre travail.

## Hypothèse 2 : Validité concourante

Notre hypothèse étant que nos tâches expérimentales et les tests évaluant les mêmes dimensions sont comparables, nous avons d'abord regardé la normalité des tests standardisés via un test de Shapiro-Wilk, à savoir l'EVT-3 et le PPVT-5. Nous obtenons une normalité pour l'EVT-3 en 2<sup>e</sup> primaire ( $W=0.99$  ;  $p=0.75$ ) et pour la dénomination expérimentale ( $W=0.98$  ;  $p=0.65$ , voir Hypothèse 4). En revanche, alors que le PPVT-5 a une distribution normale ( $W=0.96$  ;  $p=0.05$ ), les données de la désignation expérimentale ne le sont pas ( $W=0.94$  ;  $p=0.01$ , voir Hypothèse 4),

Nous avons donc réalisé une corrélation de Pearson pour comparer les épreuves de dénomination et une corrélation de Spearman pour la désignation. Les corrélations entre l'EVT-3 et la dénomination expérimentale sont supérieures à 0,7 ( $r=0.78$ ). De même que pour les corrélations entre le PPVT-5 et la désignation expérimentale, elles sont supérieures à 0,7 ( $r_s=0.72$ ). Ces résultats attestent d'une validité concourante satisfaisante des mesures de dénomination et de désignation expérimentales avec des épreuves précédemment validées évaluant le même concept.

## Hypothèse 3 : Fidélité inter-juges

Nous espérons que les deux juges qui ont coté la dénomination expérimentale aient un accord élevé. Dès lors, les scores des deux examinateurs peuvent être considérés comme fiables.

Au niveau de la fidélité inter-juges item par item, nous avons utilisé le coefficient Kappa. C'est une mesure statistique utilisée pour évaluer le niveau d'accord entre deux ou plusieurs juges selon les scores 'item'. Il varie de -1 à 1, où 1 indique un accord parfait, 0 indique un accord au niveau du hasard et des valeurs négatives suggèrent un accord inférieur à celui attendu par le hasard. Selon les recommandations littéraires, le coefficient de Kappa est jugé modéré lorsqu'il atteint 0.6.



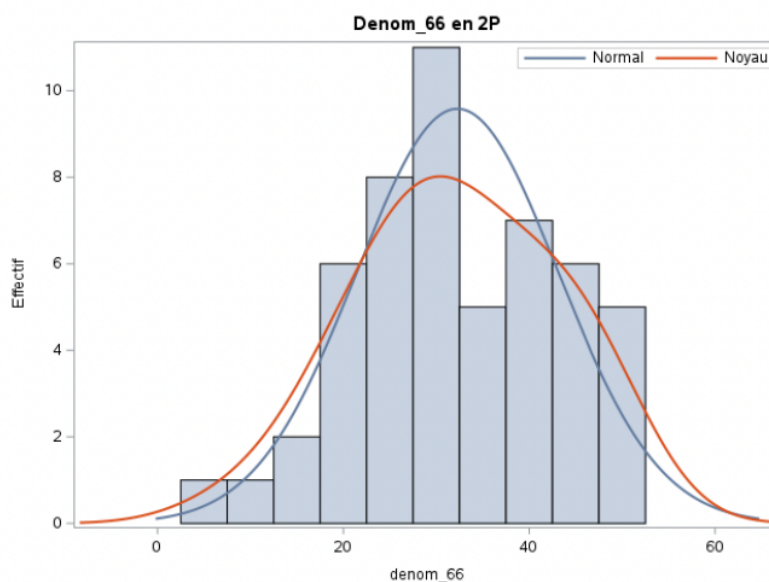
Pour les réponses données par les enfants de 2<sup>e</sup> primaire, pour chaque item, les juges ont systématiquement été en accord sauf pour 1 item (louche). En effet, un des dix sujets n'a pas été évalué de la même manière par les deux juges, ce qui conduit à une estimation du coefficient de Kappa égal à 0.78, ce qui indique que le coefficient de Kappa est jugé correct.

Pour tous les autres items, le coefficient de Kappa est de 1 puisque l'accord est parfait. Notons que pour 3 items, le coefficient de Kappa n'a pu être calculé, car seules de bonnes réponses (« bombe », « loupe ») ou de mauvaises réponses (« lard ») ont été données. Dans ce cas, il y a moins de 2 niveaux de réponses données et cela devient une constante et plus une variable. Bien que, qualitativement, les 2 juges aient été en accord pour tous les sujets. Nous pouvons donc considérer que l'accord est parfait.

#### Hypothèse 4 : Normalité

Nous avons vérifié la normalité de la distribution des résultats des enfants de 2<sup>e</sup> primaire. Nous confirmons cette hypothèse pour la dénomination ( $W=0.98$  ;  $p=0.65$ ), comme en atteste la figure 1.

Figure 1 - Graphique de normalité de la dénomination lexicale pour les enfants de 2e primaire



Afin de faciliter la compréhension et de structurer les informations, plusieurs tableaux sont proposés.

Le tableau 1 reprend les éléments suivants : moyenne, écart-type, score maximum possible, score minimum et maximum observé, aplatissement et asymétrie de la courbe pour la dénomination lexicale chez les enfants de 2<sup>e</sup> primaire.

*Tableau 1 - Éléments de normalité pour la dénomination lexicale pour les enfants de 2e primaire*

Moyenne	Écart-type	Score max. possible	Score min. observé	Score max. observé	Aplatissement de la courbe	Asymétrie de la courbe
32.17	10.91	66	5	52	-0.48 (platykurtique)	-0.18 (asymétrie vers la gauche)

Le tableau 2 présente les percentiles pour la dénomination lexicale chez les enfants de 2<sup>e</sup> primaire.

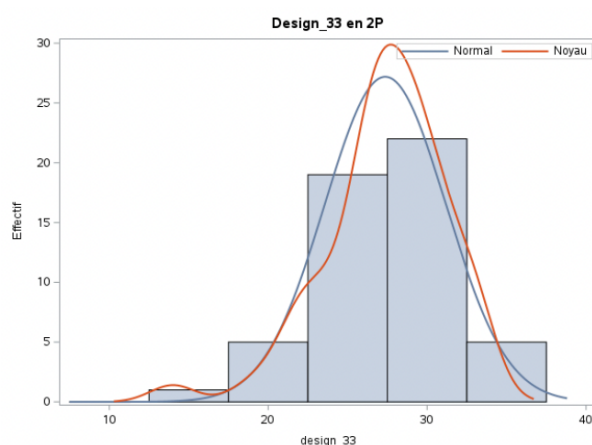
*Tableau 2 -Percentiles pour la dénomination lexicale des enfants de 2e primaire*

Percentiles	0	1	3	5	10	25	50	75	90	95	99	100
2 <sup>e</sup> primaire	5	5	12	14	18	25	32	41	46	50	52	52

De plus, chez les élèves de 2<sup>e</sup> primaire, le temps moyen requis pour dénommer les images, en tenant compte de l'ensemble des réponses, qu'elles soient correctes ou incorrectes, est de 4,02 secondes. L'écart-type associé à cette mesure est de 0,77 seconde. Lorsque l'analyse se limite uniquement aux réponses correctes, le temps moyen de dénomination diminue à 2,99 secondes, avec un écart-type de 0,57 seconde.

Par contre, nous rejetons cette hypothèse pour la désignation ( $W=0.94$  ;  $p=0.01$ ), comme en atteste la figure 2.

Figure 2 - Graphique de normalité de la désignation lexicale pour les enfants de 2<sup>e</sup> primaire



Le tableau 3 reprend les éléments suivants : moyenne, écart-type, score maximum possible, score minimum et maximum observé, aplatissement et asymétrie de la courbe pour la désignation lexicale chez les enfants de 2<sup>e</sup> primaire.

Tableau 3 - Éléments de normalité pour la désignation lexicale pour les enfants de 2<sup>e</sup> primaire

Moyenne	Écart-type	Score max. possible	Score min. observé	Score max. observé	Aplatissement de la courbe	Asymétrie de la courbe
27.365	3.81	33	14	33	1.86 (leptokurtique)	-0.98 (asymétrie sur la gauche)

Le tableau 4 présente les percentiles pour la désignation lexicale chez les enfants de 2<sup>e</sup> primaire. Ces données sont à privilégier au vu de la non normalité de la distribution.

Tableau 4 - Percentiles pour la désignation lexicale pour les enfants de 2<sup>e</sup> primaire

Percentiles	0	1	3	5	10	25	50	75	90	95	99	100
2 <sup>e</sup> primaire	14	14	19	21	22	26	28	30	32	33	33	33

La présence de valeurs extrêmes (boxplot et kurtosis = 1.86), l'asymétrie sur la gauche (boxplot et skewness = -0.98), la p-valeur du test de Shapiro-Wilk ( $W=0.94$  ;  $p=0.008$ ) conduisent à

rejeter la normalité des données. Ces moyennes et écarts-types ne pourront donc normalement pas être utilisés dans le calcul d'un score standard puisque la distribution n'est pas normale.

Enfin, en 2<sup>e</sup> primaire, les enfants mettent en moyenne 4,9 secondes, avec un écart-type de 1,15, pour désigner les images lorsque toutes les réponses, qu'elles soient correctes ou non, sont prises en compte. Lorsqu'on se limite aux réponses correctes, le temps moyen de désignation est de 4,54 secondes, avec un écart-type de 1,11.

### Hypothèse 5 : Comparaison dénomination et désignation

Notre hypothèse était que la tâche de dénomination serait plus compliquée que l'épreuve de désignation pour l'ensemble des enfants de 2<sup>e</sup> primaire. Effectivement, l'étendue du lexique compris est supérieure à l'étendue du lexique produit (Davialt, 2011).

Au préalable, nous avons dû remettre le score de dénomination sur 33 au lieu de 66 pour pouvoir la comparer avec la désignation (/33). En effet, la cotation de la dénomination était de 0 si l'enfant ne produisait rien ou s'il commettait une erreur sémantique ou sans lien avec la cible, de 1 s'il commettait une erreur phonologique lors de la production de la cible et 2 lors des productions correctes. Pour notre ajustement, nous avons décidé de réaliser une cotation binaire où l'enfant a 1 s'il produit la cible avec ou sans erreur phonologique et 0 dans les autres cas, car notre but est d'évaluer l'étendue du lexique de l'enfant et non ses productions phonologiques.

Nous allons donc réaliser une comparaison des scores (VD) aux tâches de désignation et de dénomination (VI). Nous réalisons donc un test T pour données paires (si la normalité des scores de différence est vérifiée). Autrement, nous réalisons un test de Wilcoxon pour données paires.

Dans l'échantillon des 52 élèves de 2<sup>e</sup> primaire, la différence des scores aux épreuves de désignation et de dénomination est en moyenne de 11.2 points (/33) et de 3.89 pour l'écart-type. Le test de Shapiro-Wilk conduit à tolérer la normalité de ces scores de différences ( $W=0.97$  ;  $p=0.26$ ), ce qui permet d'utiliser le test t de Student pour tester leur significativité, qui, en l'occurrence, est attestée ( $t=20.80$  ;  $p<0.0001^{***}$ ), ainsi qu'un intervalle de confiance à 95% pour la moyenne des scores de différences dans la population des 2P : [10.15 ; 12.31].

Notre hypothèse est donc vérifiée : l'épreuve de désignation est significativement mieux réussie que la dénomination.

## 5<sup>e</sup> primaire

### Hypothèse 1 : Consistance interne

Nous espérons que la réussite aux différents items au sein d'une même tâche soit relativement homogène, c'est-à-dire que les réponses aux différents items d'une même tâche sont bien corrélées. De ce fait, chaque item de nos épreuves expérimentales évalue bien ce qu'il doit évaluer à savoir le vocabulaire en réception ou en production respectivement. Nous obtenons un score KR-20 (Kuder-Richardson) de 0.78 pour la désignation. Nous atteignons les recommandations littéraires à savoir 0.7 (Ntumi et al., 2023). Nous n'avons pas utilisé le test de Cronbach, comme pour les enfants de 2<sup>e</sup> primaire, car celui-ci est recommandé pour les données métriques et non dichotomiques comme c'est le cas dans notre travail.

### Hypothèse 2 : Validité concourante

Notre hypothèse étant que nos tâches expérimentales et les tests évaluant les mêmes dimensions sont comparables, nous avons d'abord regardé la normalité des tests standardisés, à savoir l'EVT-3 et le PPVT-5. Nous avons donc réalisé un test de Shapiro-wilk pour la tester. Nous obtenons une normalité pour l'EVT-3 en 5<sup>e</sup> primaire ( $W=0.975$  ;  $p=0.35$ ) et pour la dénomination expérimentale ( $W=0.98$  ;  $p=0.74$ , voir Hypothèse 4). En revanche, alors que la désignation expérimentale a une distribution normale ( $W=0.97$  ;  $p=0.16$ , voir Hypothèse 4), les données du PPVT-5 ne le sont pas ( $W=0.94$  ;  $p=0.02$ ). Nous avons donc réalisé une corrélation de Pearson pour comparer les épreuves de dénomination et une corrélation de Spearman pour la désignation.

Les corrélations entre l'EVT-3 et la dénomination expérimentale sont supérieures à 0,7 ( $r=0.8$ ). Il y a donc une bonne validité concourante pour notre dénomination expérimentale. En revanche, nous n'avons pas une validité concourante suffisante entre le PPVT-5 et la désignation expérimentale ( $r_s=0.57$ ).

### Hypothèse 3 : Fidélité inter-juges

Comme pour les enfants de 2<sup>e</sup> primaire, nous avons calculé le coefficient Kappa pour les scores 'item' pour les enfants de 5<sup>e</sup> primaire. Nous espérons que celui-ci atteigne au minimum 0.6 comme le recommande la littérature (Landis & Koch, 1977).

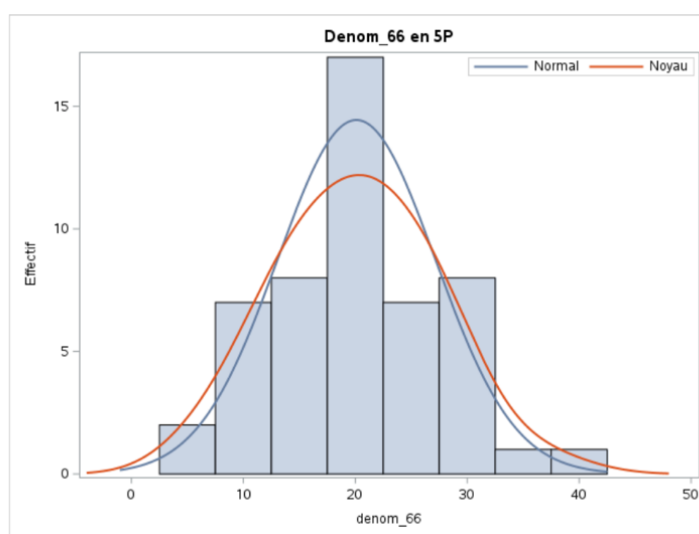
Les juges ont été en accord sur tous les sujets pour chaque item sauf pour 3 où 1 des 10 sujets n'a pas été évalué de la même manière par les 2 juges : « sabre » ( $\kappa = 0.8$ ) et « surf » ( $\kappa = 0.78$ ). L'item « guêtre » n'a pas pu être réalisé, car le juge 2 n'a attribué que des '0' (moins de deux niveaux).

Pour les autres items, la valeur du kappa est égale à 1 puisque l'accord est parfait. Notons cependant que pour les items suivants, kappa n'a pas pu être calculé (car moins de 2 niveaux), bien que, qualitativement, les 2 juges aient été en accord pour attribuer un « 0 » à tous les sujets pour ces items : « cime », « douille », « germe », « jauge », « meule », « sciure », « sonde », « vanne ».

### Hypothèse 4 : Normalité

Nous avons vérifié la normalité de la distribution des résultats des enfants de 5<sup>e</sup> primaire. Nous confirmons cette hypothèse pour la dénomination ( $W=0.98$  ;  $p=0.74$ ), comme en atteste la figure 3.

Figure 3 - Graphique de normalité de la dénomination lexicale pour les enfants de 5e primaire



Afin de faciliter la compréhension et de structurer les informations, plusieurs tableaux sont proposés.

Le tableau 5 reprend les éléments suivants : moyenne, écart-type, score maximum possible, score minimum et maximum observé, aplatissement et asymétrie de la courbe pour la dénomination lexicale chez les enfants de 5<sup>e</sup> primaire.

*Tableau 5 - Éléments de normalité pour la dénomination lexicale pour les enfants de 5e primaire*

Moyenne	Écart-type	Score max. possible	Score min. observé	Score max. observé	Aplatissement de la courbe	Asymétrie de la courbe
20.10	7.04	66	6	38	-0.20 (platykurtique)	0.11 (asymétrie sur la droite)

Le tableau 6 présente les percentiles pour la dénomination lexicale chez les enfants de 5<sup>e</sup> primaire.

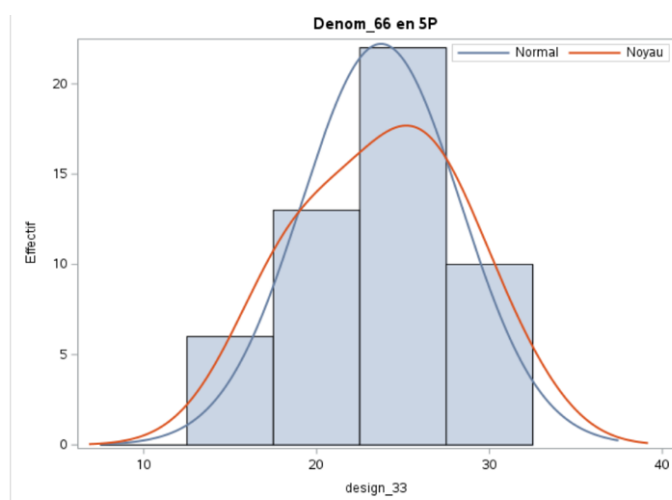
*Tableau 6 - Percentiles pour la dénomination lexicale pour les enfants de 5e primaire*

Percentiles	0	1	3	5	10	25	50	75	90	95	99	100
5 <sup>e</sup> primaire	6	6	6	8	12	14	20	25	28	30	38	38

De plus, les élèves de 5<sup>e</sup> primaire présentent un temps moyen de dénomination de 5,06 secondes (écart-type : 0,95) lorsque l'on considère l'ensemble des items, qu'ils soient correctement ou incorrectement nommés. En restreignant l'analyse aux seuls items correctement dénommés, le temps moyen de réponse s'élève à 3,41 secondes, avec un écart-type de 0,73.

Nous obtenons également une tolérance de l'hypothèse de normalité pour la désignation lexicale ( $W=0.97$  ;  $p=0.16$ ), comme en atteste la figure 4.

Figure 4 -Graphique de normalité de la désignation lexicale pour les enfants de 5e primaire



Le tableau 7 reprend les éléments suivants : moyenne, écart-type, score maximum possible, score minimum et maximum observé, aplatissement et asymétrie de la courbe pour la désignation lexicale chez les enfants de 5<sup>e</sup> primaire.

Tableau 7 - Éléments de normalité pour la désignation lexicale pour les enfants de 5e primaire

Moyenne	Écart-type	Score max. possible	Score min. observé	Score max. observé	Aplatissement de la courbe	Asymétrie de la courbe
23.725	4.58	33	14	32	-0.83 (platykurtique)	-0.11 (asymétrie sur la gauche)

Le tableau 8 présente les percentiles pour la désignation lexicale chez les enfants de 5<sup>e</sup> primaire.

Tableau 8 - Percentiles pour la désignation lexicale pour les enfants de 5e primaire

Percentiles	0	1	3	5	10	25	50	75	90	95	99	100
5 <sup>e</sup> primaire	14	14	16	17	17	19	24	27	29	31	32	32

Enfin, ils mettent en moyenne 5,44 secondes avec un écart-type s'élevant à 1,18 seconde pour désigner les images lorsque toutes les réponses, correctes ou incorrectes, sont prises en compte.



Quand on ne prend en compte que les réponses correctes, le temps moyen de désignation est de 4,76 secondes avec un écart-type de 0,93.

#### Hypothèse 5 : Comparaison dénomination et désignation

Notre hypothèse était que la tâche de dénomination serait plus compliquée que l'épreuve de désignation pour l'ensemble des enfants de 5<sup>e</sup> primaire. Effectivement, l'étendue du lexique compris est supérieure à l'étendue du lexique produit (Daviault, 2011).

Au préalable, nous avons dû remettre le score de dénomination sur 33 au lieu de 66 pour pouvoir la comparer avec la désignation (/33). En effet, la cotation de la dénomination était de 0 si l'enfant ne produisait rien ou s'il commettait une erreur sémantique ou sans lien avec la cible, de 1 s'il commettait une erreur phonologique lors de la production de la cible et 2 lors des productions correctes. Pour notre ajustement, nous avons décidé de réaliser une cotation binaire où l'enfant a 1 s'il produit la cible avec ou sans erreur phonologique et 0 dans les autres cas, car notre but est d'évaluer l'étendue du lexique de l'enfant et non ses productions phonologiques.

Nous allons donc réaliser la même procédure pour le choix des tests que les enfants de 2<sup>e</sup> primaire.

Dans l'échantillon des 51 élèves de 5<sup>e</sup> primaire, la différence des scores aux épreuves de désignation et de dénomination est en moyenne de 13.7 points (/33) avec un écart-type de 3,59. Si le caractère significatif des différences entre ces 2 types d'épreuves est attesté par le test de Wilcoxon ( $S=663$  ;  $p < 0,0001$ ) vu que le test de Shapiro-Wilk conduit ici à rejeter la normalité des scores de différences au seuil 0.05.

Notre hypothèse est donc vérifiée : l'épreuve de désignation est significativement mieux réussie que la dénomination.

## Comparaison 2<sup>e</sup> primaire et 5<sup>e</sup> primaire

### Hypothèse 6 : Comparaison erreurs sémantiques et phonologiques pour la tâche de désignation

Notre hypothèse est que nous estimons que les enfants vont davantage commettre des erreurs de type sémantique en 5<sup>e</sup> primaire et réaliser plus d'erreurs de type phonologique en 2<sup>e</sup> primaire.

Selon la littérature, plus les enfants sont âgés, plus ils commettront des erreurs sémantiques (et moins d'erreurs phonologiques) (Bragard et al., 2010; Sheng & McGregor, 2010). Nous avons donc voulu connaître les types d'erreurs commises selon l'année scolaire, en comparant les moyennes de chacune (selon l'année). Pour ce faire, nous avons relevé le nombre d'erreurs sémantiques commises et le nombre d'erreurs phonologiques commises à la tâche de désignation.

Pour chaque année scolaire, le tableau 9 expose les données statistiques concernant les deux types d'erreurs.

*Tableau 9 - Données statistiques concernant les deux types d'erreurs réalisées*

Année scolaire	Nombre d'observations	Variables	Moyenne	Écart-type	Maximum	Minimum
2P	52	Erreurs sémantiques	2,96	2,26	8	0
		Erreurs phonologiques	1,75	1,81	6	0
5P	51	Erreurs sémantiques	3,47	2,52	10	0
		Erreurs phonologiques	5,27	2,63	10	0

Nous avons donc comparé ces deux types d'erreurs, d'abord, selon l'année à l'aide d'une ANOVA à mesures répétées. Effectivement, si l'erreur de l'enfant n'est pas phonologique, elle sera sémantique ou neutre. On remarque un effet de l'année scolaire sur le nombre d'erreurs total, les enfants de 5P commettant davantage d'erreurs que les enfants de 2P ( $F=30.04$  ;  $p<0.0001$ ). Cela fait écho aux moyennes décrites plus haut et peut s'expliquer par le fait que les items proposés dans les deux tranches d'âge ne sont pas les mêmes. De plus, les résultats de

la comparaison du nombre d'erreurs selon le type (phonologiques VS sémantiques) ne sont pas significativement différents ( $F < 0.28$ ).

Ensuite, nous avons comparé les types d'erreurs avec leurs moyennes respectives, selon l'année scolaire. On a donc d'abord réalisé un test de Shapiro-wilk pour tester la normalité du nombre d'erreurs (selon le type et l'année) et la différence entre ceux-ci. Le tableau 10 exprime ces résultats.

*Tableau 10 - Test de Shapiro-Wilk évaluant la normalité des deux types d'erreurs*

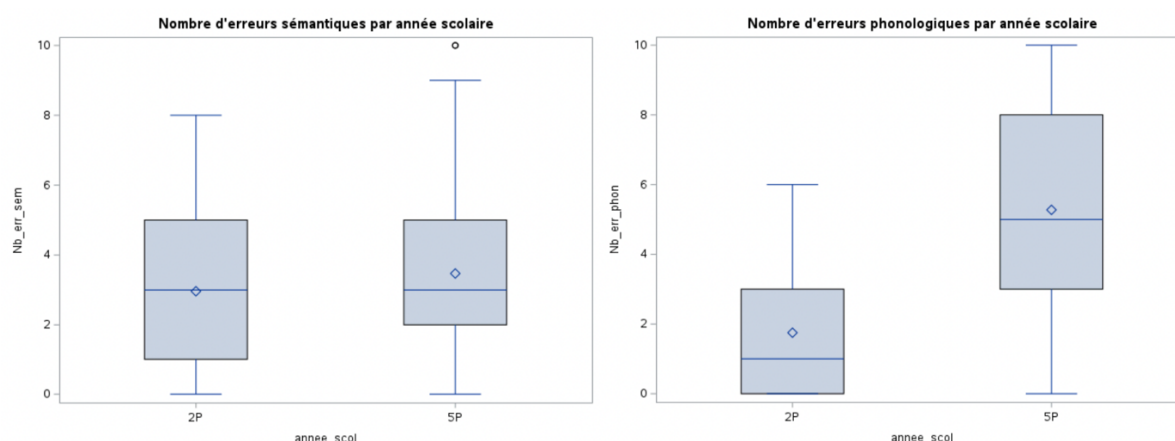
Test de Shapiro-Wilk	2P		5P	
	<i>W</i>	<i>p</i>	<i>W</i>	<i>p</i>
Nombre d'erreurs sémantiques	0.940424	0.0117*	0.966323	0.1550
Nombre d'erreurs phonologiques	0.927049	0.0035**	0.914242	0.0013**

Dès lors, malgré que l'ANOVA soit réputée pour être robuste quant aux violations de normalité, il nous semble toutefois préférable d'opter pour une approche non paramétrique pour les comparaisons 2 à 2.

Nous avons réalisé un test de Wilcoxon pour échantillons indépendants pour comparer les enfants des deux années scolaires séparément pour le nombre d'erreurs sémantiques et phonologiques.

En ce qui concerne les erreurs sémantiques, nous remarquons une absence de différence significative entre les deux années ( $W=2782.000$  ;  $p=0.39$ ). En revanche, nous obtenons une différence significative pour les erreurs phonologiques commises ( $W=3594.000$  ;  $p<0.0001$ ). On constate, comme l'illustre le graphe 5 ci-dessous, que les enfants de 5<sup>e</sup> primaire commettent plus d'erreurs phonologiques que les enfants de 2<sup>e</sup> primaire.

Figure 5 - Graphiques de comparaison des erreurs pour les enfants de 2e et 5e primaires



Enfin, nous avons voulu connaître la moyenne du type d’erreurs (sémantique VS phonologique) par année scolaire séparément.

Dans l’échantillon des 2P, la différence du nombre d’erreurs commises (sémantiques – phonologiques) est en moyenne égale à 1.21 (ET=2,54). Chez les 5P, la différence du nombre d’erreurs commises (sémantiques – phonologiques) est en moyenne égale à -1.80 (ET=2,98).

Nous concluons donc sur une différence significative du nombre d’erreurs phonologiques entre les deux années, mais une absence de significativité pour les erreurs sémantiques entre les deux années.

## Discussion

### Réponses à nos hypothèses

Ce mémoire avait pour objectif de valider et normaliser une tâche de désignation lexicale et une tâche de dénomination lexicale dans le cadre du diagnostic du TDL. Dans ce cadre, nous avons émis différentes hypothèses :

Premièrement, nous évaluons si chaque épreuve mesurait bien un seul et même concept. Il s’agit donc de la vérification de la consistance interne via le coefficient KR-20. Pour les enfants de 2<sup>e</sup> primaire, nous avons atteint le coefficient KR-20 acceptable pour la littérature, qui est de 0.7 (Ntumi et al., 2023), de même que pour les enfants de 5<sup>e</sup> primaire. Les items sont donc assez cohérents entre eux.

Deuxièmement, nous voulions voir si nos épreuves expérimentales mesuraient bien ce qu'elles sont censées mesurer (Friberg, 2010; McCauley & Swisher, 1984). Dès lors, nous avons réalisé une corrélation entre l'épreuve de dénomination expérimentale et l'EVT-3 ainsi qu'entre la désignation expérimentale et le PPVT-5. Pour les enfants de 2<sup>e</sup> primaire, nos épreuves expérimentales attestent d'une bonne validité concourante. Pour les enfants de 5<sup>e</sup> primaire, la tâche expérimentale de dénomination présente également une bonne validité concourante, par contre, la tâche expérimentale de désignation n'a pas une validité concourante suffisante selon la littérature. Nous pourrions penser à l'hétérogénéité de notre échantillon. De plus, notre tâche expérimentale ne contient que des noms, comparativement au PPVT-5 qui contient des verbes, des substantifs et des adjectifs qualitatifs. Des facteurs parasites peuvent également avoir contribué à ce résultat comme l'attention ou encore le niveau socio-culturel.

Troisièmement, nous avons analysé si nos épreuves permettaient de donner des scores fiables, quel que soit l'évaluateur qui les administre via le coefficient de Kappa (Landis & Koch, 1977). Pour les enfants de 2<sup>e</sup> primaire, un des dix sujets n'a pas été évalué de la même manière par les deux juges, pour le reste, l'accord est parfait puisque le coefficient de Kappa est de 1. Pour les enfants de 5<sup>e</sup> primaire, les juges ont été en accord pour tous les sujets pour chaque item sauf pour 3 où 1 des sujets n'a pas été évalué de la même manière par les 2 juges, sinon pour les autres items, l'accord est parfait étant donné que le coefficient de Kappa est de 1. La standardisation a globalement été respectée entre les évaluateurs (cotation). Nous avons des consignes de passation et de cotation claires, pour nous permettre de maximiser cette fidélité.

Quatrièmement, nous avons vérifié que nos données récoltées ne présentaient pas d'effet plancher ou plafond. Nous avons également vérifié la distribution des données (Spaulding et al., 2012). En 2<sup>e</sup> primaire, nous confirmons que la distribution des données est normale pour la tâche de dénomination. Quant à la tâche de désignation, la distribution des données est non normale. Pour cette dernière, nous obtenons une asymétrie gauche. Nous avons donc obtenu des données avec un effet plafond. L'épreuve serait donc trop facile. Ceci pourrait être dû à des items trop simples pour les enfants ou encore, à nouveau, par l'hétérogénéité de l'échantillon (les enfants testés en fin d'année scolaire avaient un bagage langagier supplémentaire que les enfants entrant en 2<sup>e</sup> primaire). Nous pourrions revoir les items de l'épreuve en prenant des items ayant une fréquence moins élevée. En 5<sup>e</sup> primaire, la distribution des données est normale tant pour la tâche de dénomination que de désignation. Cela confirme donc notre hypothèse pour ceux-ci.

Cinquièmement, nous avons supposé que les enfants, en 2<sup>e</sup> comme en 5<sup>e</sup> primaire, auront plus de facilités lors de la tâche de compréhension lexicale que lors de la tâche de production lexicale. Pour les deux années scolaires, la tâche de désignation est significativement mieux réussie que la tâche de dénomination. Ceci est en concordance avec les données de la littérature (Daviault, 2011). En 2018, Zhong évoquait également le fait que les personnes comprennent davantage de mots qu’elles n’en produisent.

Et enfin, nous estimions que les enfants commettraient davantage d’erreurs de type sémantique en 5<sup>e</sup> primaire et réaliseraient plus d’erreurs de type phonologique en 2<sup>e</sup> primaire (Bragard et al., 2010). Contre toute attente, nos résultats vont à l’encontre de notre hypothèse. En effet, les enfants de 2<sup>e</sup> primaire ont commis plus d’erreurs sémantiques que phonologiques et inversement pour les enfants de 5<sup>e</sup> primaire. Même si nous avons contrôlés les effets environnementaux, le volume de présentation des items dans le casque a pu influencer la perception de ceux-ci pour les enfants de 5<sup>e</sup> primaire. Dès lors, les enfants auraient donc moins perçu les traits phonologiques des cibles. L’attention des enfants a pu également jouer un rôle dans nos résultats. Étant donné que l’épreuve de désignation a été proposée en premier lieu lors de la 2<sup>e</sup> séance, certains enfants avaient besoin de plus de temps pour se focaliser sur la tâche. De même, selon le moment de testing (heures scolaires, lors des récréations, en soirée), la concentration des enfants variait. Enfin, l’échantillon pourrait ne pas être assez représentatif de la population d’enfants tout-venant.

## Forces de l’étude

Ces deux épreuves expérimentales créées ont permis, pour les enfants de 3<sup>e</sup> maternelle à la 5<sup>e</sup> primaire (de 2<sup>e</sup> primaire et 5<sup>e</sup> primaire dans le cadre de ce mémoire), d’avoir des normes récentes, en percentiles ou encore en moyennes (écarts-types) (selon la normalité) pour chaque année scolaire. Nos échantillons sont équivalents au niveau du sexe et du nombre de participants. Nous avons globalement un échantillon de 50 participants pour chaque année scolaire. Ceci améliore la puissance statistique (Cohen, 1992). Cependant, 100 participants par niveau est recommandé (Friberg, 2010). Une des critiques de l’EVT-3 et du PPVT-5, était le nombre trop important d’items et donc un effet de lassitude. Cependant, il s’agissait des tests ayant les normes les plus récentes (Dunn, 2023; Williams, 2023). Nos épreuves sont donc un bon compromis avec des normes récentes et une passation qui n’est pas trop longue. De plus,

les deux épreuves évaluent les mêmes cibles dans les deux versants (compréhension et production lexicales). Ceci permet une meilleure comparaison des capacités concernant le lexique passif et actif.

## Limites

Dans nos épreuves, il aurait été intéressant de varier les classes grammaticales afin d'en évaluer plusieurs. Nous savons que les enfants ayant un TDL ont davantage de difficultés pour les verbes (Bragard & Schelstraete, 2006). Dès lors, l'introduction de verbes dans nos épreuves aurait été judicieuse. La fréquence des items au sein de nos épreuves expérimentales auraient pu être plus élargie pour pouvoir observer davantage un effet de fréquence (Ferrand, 1997). Une autre limite de cette étude concerne le fait que certains enfants n'ont pas été évalués par le même évaluateur lors des deux séances. Ce changement au niveau de l'évaluateur peut avoir introduit une variabilité au niveau des résultats, notamment par rapport aux différences dans la manière de présenter les consignes, de gérer la relation avec l'enfant ou d'attribuer les points pour la cotation. Il aurait été préférable que chaque enfant soit évalué par la même personne lors des deux séances pour garantir une plus grande homogénéité dans la passation et de limiter les biais liés à l'évaluateur. Nous aurions également pu prendre plus en compte le niveau socio-culturel de la famille, les données développementales ou encore les informations relatives aux parents (études, profession, ...) au niveau des résultats (Hoff, 2003). Enfin, comme dit plus haut, notre échantillon par année scolaire aurait pu être plus élevé afin d'augmenter davantage la puissance statistique. Nous aurions pu voir si les effets observés auraient toujours été présents avec un échantillon plus grand. La littérature nous recommande 100 participants par niveau (Friberg, 2010).

## Conclusions

Le but de ce travail était de normaliser et de valider une nouvelle épreuve lexicale pour des enfants de 2<sup>e</sup> et 5<sup>e</sup> primaires. Ceci s'inscrit dans la thèse de Madame Menjot. Le but de celle-ci est de créer une épreuve d'évaluation dynamique pour les enfants de classes primaires ayant un trouble du langage. Nous avons remarqué que les principaux tests évaluant le lexique chez les enfants de primaire ne remplissent pas les critères psychométriques recommandés dans la littérature pour un diagnostic de trouble du langage. Comme utiliser les mêmes items lors de la

tâche de compréhension et la tâche de production ou encore avoir des normes récentes. Dès lors, une tâche de dénomination lexicale et une tâche de désignation lexicale ont été créées. Nous les avons proposées à 52 enfants de 2<sup>e</sup> primaire et 51 enfants de 5<sup>e</sup> primaire.

Nous avons pu remarquer que les deux tâches expérimentales avaient une bonne fidélité inter-juges ainsi qu'une validité concourante satisfaisante pour les enfants de 2<sup>e</sup> primaire. Au niveau de la normalité, la distribution des données est normale pour la tâche de dénomination mais est non-normale pour la tâche de désignation. La consistance interne est jugée acceptable car elle atteint la recommandation littéraire qui est de 0.7. Enfin, lorsque nous comparons l'épreuve de dénomination et l'épreuve de désignation, les résultats montrent que l'épreuve de désignation a été significativement mieux réussie que celle de dénomination, ce qui suit notre hypothèse.

Chez les enfants de 5<sup>e</sup> primaire, nous avons pu constater qu'il y a une bonne fidélité inter-juges ainsi qu'une bonne validité concourante pour la tâche expérimentale de dénomination. Par contre, il n'y a pas une validité concourante suffisante entre le PPVT-5 et la tâche expérimentale de désignation. Pour la normalité, la distribution est normale tant pour l'épreuve de dénomination que celle de désignation. La consistance interne est également jugée acceptable car elle atteint la recommandation littéraire qui est de 0.7. Enfin, tout comme pour les enfants de 2<sup>e</sup> primaire, l'épreuve de désignation est significativement mieux réussie que celle de dénomination.

Nous avons également comparé les enfants de 2<sup>e</sup> primaire et les enfants de 5<sup>e</sup> primaire au niveau des erreurs phonologiques et sémantiques pour la tâche de désignation. Contre toute attente, les résultats vont à l'encontre de notre hypothèse qui était que les enfants de 5<sup>e</sup> primaire vont davantage faire d'erreurs de type sémantique alors que les enfants de 2<sup>e</sup> primaire vont davantage faire d'erreurs de type phonologique.

Une suite est envisagée pour ce travail étant donné que la thèse de Madame Menjot se concentre sur les enfants TDL et sur l'élaboration d'une tâche dynamique. D'autres perspectives pourraient d'être de modifier les critères de sélection des participants en incluant davantage d'enfants bilingues ou de différents milieux socio-économiques, d'envisager une plus large variabilité de fréquence pour les items de la tâche ou encore de varier les classes grammaticales.



## BIBLIOGRAPHIE

Arias-Trejo, N., Angulo-Chavira, A. Q., & Plunkett, K. (2025). The effects of phonological and semantic similarity on early referent identification. *Journal of Experimental Child Psychology*, 249, 106112. <https://doi.org/10.1016/j.jecp.2024.106112>

Best, W. (2005). Investigation of a new intervention for children with word-finding problems. *International Journal of Language & Communication Disorders*, 40(3), 279-318. <https://doi.org/10.1080/13682820410001734154>

Bianco, M. (2021, February). *La fluence en lecture au cycle 3* [Vidéo]. YouTube. <https://www.youtube.com/watch?v=KVTKLYDQAmo>

Bishop, D. V. M., & McDonald, D. (2009). Identifying language impairment in children : Combining language test scores with parental report. *International Journal of Language & Communication Disorders*, 44(5), 600-615. <https://doi.org/10.1080/13682820802259662>

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & the CATALISE-2 consortium. (2017). *Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology*. *Journal of Child Psychology and Psychiatry*, 58(10), 1068–1080. <https://doi.org/10.1111/jcpp.12721>

Bloom, P. (2001). Précis of how children learn the meanings of words. *Behavioral and Brain Sciences*, 24(6), 1095–1103. <https://doi.org/10.1017/S0140525X01000139>

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A Quasi-Universal Nonword Repetition Task as a Diagnostic Tool for Bilingual Children Learning Dutch as a Second Language. *Journal of Speech, Language, and Hearing Research*, 58(6), 1747-1760. [https://doi.org/10.1044/2015\\_JSLHR-L-15-0058](https://doi.org/10.1044/2015_JSLHR-L-15-0058)

Bogliotti, C. (2012). Les troubles de la dénomination. *Langue française*, 174(2), 95–110. <https://doi.org/10.3917/lf.174.0095>

Bonin, P., Laroche, B., & Méot, A. (2021). Psycholinguistic norms for a set of 506 French

compound words. *Behavior Research Methods*, 54(1), 393-413. <https://doi.org/10.3758/s13428-021-01601-w>

Borovsky, A. (2022). Lexico-semantic structure in vocabulary and its links to lexical processing in toddlerhood and language outcomes at age three. *Developmental Psychology*, 58(4), 607-630. <https://doi.org/10.1037/dev0001291>

Bourque, J., Doucet, D., LeBlanc, J., Dupuis, J., & Nadeau, J. (2020). L'alpha de Cronbach est l'un des pires estimateurs de la consistance interne : Une étude de simulation. *Revue des Sciences de l'Éducation*, 45(2), 78–99. <https://doi.org/10.7202/1067534ar>

Bragard, A., & Schelstraete, M.-A. (2006). Le manque du mot dans les troubles spécifiques du langage chez l'enfant. *L'Année Psychologique*, 106(4), 633–661. <https://doi.org/10.4074/S0003503306004064>

Bragard, A., Schelstraete, M.-A., Collette, E., & Grégoire, J. (2010). Évaluation du manque du mot chez l'enfant : Données développementales récoltées auprès d'enfants francophones de sept à 12 ans. *European Review of Applied Psychology*, 60(2), 113-127. <https://doi.org/10.1016/j.erap.2009.11.003>

Bussy, G., Rigard, C., & Des Portes, V. (2014). Impact d'un entraînement de la mémoire à court terme verbale sur le langage d'enfants ayant une déficience intellectuelle. *Revue Francophone de la Déficience Intellectuelle*, 24, 141–151. <https://doi.org/10.7202/1021269ar>

Caesar, L. G., & Kohler, P. D. (2009). Tools clinicians use: A survey of language assessment procedures used by school-based speech-language pathologists. *Communication Disorders Quarterly*, 30(4), 226–236. <https://doi.org/10.1177/1525740108326334>

Cain, K., & Oakhill, J. (2011). Matthew effects in young readers: Reading comprehension and reading experience aid vocabulary development. *Journal of Learning Disabilities*, 44(5), 431–443. <https://doi.org/10.1177/0022219411410042>

Cattini, J., & Lafay, A. (2024). Analyse critique des qualités psychométriques des batteries de performances évaluant la cognition mathématique. *ANAE*, 190, 280–290.

Chevrié-Muller, C., Maillart, C., & Simon, A. M. (2011). *L2MA-2 : Langage oral, langage écrit, mémoire, attention* [Oral language, written language, memory, attention]. ECPA.

Chevrié-Müller, C., & Plaza, M. (2001). *Nouvelles épreuves pour l'évaluation du langage oral*. Les Éditions du Centre de Psychologie Appliquée.

Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2008). Bootstrapping lexical and syntactic acquisition. *Language and Speech*, 51(1–2), 61–75. <https://doi.org/10.1177/00238309080510010501>

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. <https://doi.org/10.1037/0033-2909.112.1.155>

Colunga, E., & Sims, C. E. (2017). Not only size matters: Early-talker and late-talker vocabularies support different word-learning biases in babies and networks. *Cognitive Science*, 41(S1), 73–95. <https://doi.org/10.1111/cogs.12409>

Coquet, F., Ferrand, P., & Roustit, J. (2009). *EVALO 2-6*.

Cronin, V. S. (2002). The syntagmatic–paradigmatic shift and reading development. *Journal of Child Language*, 29(1), 189–204. <https://doi.org/10.1017/S0305000901004998>

Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school readiness and early reasoning. *Journal of Educational Psychology*, 106(3), 639–650. <https://doi.org/10.1037/a0035756>

Daviault, D. (2011a). L'acquisition du lexique. In *L'émergence et le développement du langage chez l'enfant* (pp. 79–102). Chenelière Éducation.

Daviault, D. (2011b). Les concepts fondamentaux et les grandes théories. In *L'émergence et le développement du langage chez l'enfant* (pp. 1–8). Chenelière Éducation.

Dent, K., Johnston, R. A., & Humphreys, G. W. (2008). Age of acquisition and word frequency

effects in picture naming: A dual-task investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(2), 282–301. <https://doi.org/10.1037/0278-7393.34.2.282>

Duff, F. J., Reen, G., Plunkett, K., & Nation, K. (2015). Do infant vocabulary skills predict school-age language and literacy outcomes? *Journal of Child Psychology and Psychiatry*, 56(8), 848–856. <https://doi.org/10.1111/jcpp.12378>

Dujardin, E., Auphan, P., Bailloud, N., Ecalle, J., & Magnan, A. (2021). Tools and teaching strategies for vocabulary assessment and instruction: A review. *Social Education Research*, 3, 34–66. <https://doi.org/10.37256/ser.3120221044>

Dunn, D. M. (2023). *Peabody Picture Vocabulary Test* (5e éd.). NCS Pearson.

Dymarska, A., Connell, L., & Banks, B. (2023). Weaker than you might imagine: Determining imageability effects on word recognition. *Journal of Memory and Language*, 129, Article 104398. <https://doi.org/10.1016/j.jml.2022.104398>

Ecalle, J., Bailloud, N., Dujardin, E., & Magnan, A. (2021). Évaluation informatisée du vocabulaire chez les enfants de 8 à 11 ans. *Canadian Journal of Learning and Technology*, 47(2). <https://doi.org/10.21432/cjlt28025>

Feldman, H. M. (2019). How young children learn language and speech. *Pediatrics in Review*, 40(8), 398–411. <https://doi.org/10.1542/pir.2017-0325>

Ferrand, L. (1997). La dénomination d'objets : théories et données. *L'année psychologique*, 97(1), 113–146. <https://doi.org/10.3406/psy.1997.28939>

Florin, A. (2010). Le développement du lexique et l'aide aux apprentissages. *Enfances & Psy*, 47(2), 30–41. <https://doi.org/10.3917/ep.047.0030>

Frewin, K. L., Gerson, S. A., Vanderwert, R. E., & Gambi, C. (2025). Parent-reported relations between vocabulary and motor development in infancy: Differences between verbs and nouns. *Infancy*, 30(1), e12638. <https://doi.org/10.1111/infa.12638>

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

Fu, N. C., Chan, A., Chen, S., Polišenská, K., & Chiat, S. (2024). Revisiting nonword repetition as a clinical marker of developmental language disorder: Evidence from monolingual and bilingual L2 Cantonese. *Brain and Language*, 257, 105450. <https://doi.org/10.1016/j.bandl.2024.105450>

Gierut, J. A. (2007). Phonological complexity and language learnability. *American Journal of Speech-Language Pathology*, 16(1), 6–17. [https://doi.org/10.1044/1058-0360\(2007/003\)](https://doi.org/10.1044/1058-0360(2007/003))

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, 41(4), 481–505. <https://doi.org/10.1093/applin/amy057>

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Center for Educator Compensation Reform. <https://files.eric.ed.gov/fulltext/ED532068.pdf>

Hadley, E. B., & Dickinson, D. K. (2020). Measuring young children's word knowledge: A conceptual review. *Journal of Early Childhood Literacy*, 20(2), 223–251. <https://doi.org/10.1177/1468798417753713>

Hamann, C., & Abed Ibrahim, L. (2017). Methods for identifying specific language impairment in bilingual populations in Germany. *Frontiers in Communication*, 2, 16. <https://doi.org/10.3389/fcomm.2017.00016>

Hashimoto, B. J. (2021). Is frequency enough? The frequency model in vocabulary size testing. *Language Assessment Quarterly*, 18(2), 171–187. <https://doi.org/10.1080/15434303.2020.1860058>

Helloin, M.-C., Thibault, M.-P. & Croteau, B. (2010). Batterie informatisée pour l'examen du langage oral et écrit pour les les enfants âgés de 5 à 8 ans : Exalang 5-8. Orthomotus.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, 74(5), 1368–1378. <https://doi.org/10.1111/1467-8624.00612>

Jullien, S. (2008). Constructions syntaxiques et discours : Les introductions de référents dans des narrations produites par des enfants présentant des troubles spécifiques du développement du langage oral et des enfants tout-venant. *Travaux neuchâtelois de linguistique*, 48, 7–24. <https://doi.org/10.26034/tranel.2008.2806>

Kazanina, N., Bowers, J. S., & Idsardi, W. (2018). Phonemes: Lexical access and beyond. *Psychonomic Bulletin & Review*, 25(2), 560–585. <https://doi.org/10.3758/s13423-017-1362-0>

Kern, S. (2019a). Des premiers mots à l'émergence de la grammaire. In *Le développement du langage chez le jeune enfant* (pp. 85–112). De Boeck Supérieur.

Kern, S. (2019b). Introduction. In *Le développement du langage chez le jeune enfant : Théorie, clinique, pratique* (pp. 7–12). De Boeck Supérieur.

Khomsî, A. (2001). *Évaluation du langage oral*. ECPS.

Krzemien, M., Dauvister, E., & Maillart, C. (2020). La généralisation chez les enfants avec un trouble développemental du langage : S'intéresser au niveau lexico-sémantique. *ANAE: Approche Neuropsychologique des Apprentissages chez l'Enfant*, 32. <https://hdl.handle.net/2268/248708>

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>

Launay, L., Maeder, Ch., Roustit, J., & Touzin, M. (2018). *Evaléo 6-15 : Évaluation du langage écrit et du langage oral 6-15 ans*.

Leonard, L. B., Ellis Weismer, S., Miller, C. A., Francis, D. J., Tomblin, J. B., & Kail, R. V.

(2007). Speed of Processing, Working Memory, and Language Impairment in Children. *Journal of Speech, Language, and Hearing Research*, 50(2), 408-428. [https://doi.org/10.1044/1092-4388\(2007/029\)](https://doi.org/10.1044/1092-4388(2007/029))

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX : A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers*, 36(1), 156–166. <https://doi.org/10.3758/BF03195560>

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 à 11,6 ans*. <https://medecine.univ-lille.fr/orthophonie/reseau-professionnel>

MacRoy-Higgins, M., Shafer, V. L., Fahey, K. J., & Kaden, E. R. (2016). Vocabulary of Toddlers Who Are Late Talkers. *Journal of Early Intervention*, 38(2), 118-129. <https://doi.org/10.1177/1053815116637620>

Maillart, C. (2022). Introduction. Le trouble développemental du langage : Enjeux actuels. *Enfance*, 1(1), 5–23. <https://doi.org/10.3917/enf2.221.0005>

Maillart, C. (2019). *Le profil professionnel (et de compétences) des logopèdes*. [https://www.fplse.uliege.be/upload/docs/application/pdf/2023-08/referentiel\\_de\\_compences\\_du\\_logopede\\_-\\_maillart\\_2019.pdf](https://www.fplse.uliege.be/upload/docs/application/pdf/2023-08/referentiel_de_compences_du_logopede_-_maillart_2019.pdf)

Manoiloff, L., Segui, J., & Hallé, P. (2013). L'effet de fréquence dans l'accès aux propriétés phonologiques des noms d'objets. *L'Année Psychologique*, 113(3), 335-348. <https://doi.org/10.3917/anpsy.133.0335>

McCauley, R. J., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders*, 49(1), 34–42. <https://doi.org/10.1044/jshd.4901.34>

Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1), 116–133. <https://doi.org/10.1037/0278-7393.21.1.116>

Nicollet, J., Guillen, C. J., Jouhar, A.-C., & Rossier, J. (2009). Performance on intelligence tests: Toward a reversal of the Flynn effect? *L'Orientation scolaire et professionnelle*, 38(3), 353–368. <https://doi.org/10.4000/osp.2109>

Peereman, R., Lété, B., & Sprenger-Charolles, L. (2007). Manulex-infra: Distributional characteristics of grapheme–phoneme mappings, and infralexical and lexical units in child-directed written material. *Behavior Research Methods*, 39(3), 579–589. <https://doi.org/10.3758/BF03193029>

Perraudin, S., & Mounoud, P. (2009). Contribution of the priming paradigm to the understanding of the conceptual developmental shift from 5 to 9 years of age. *Developmental Science*, 12(6), 956–977. <https://doi.org/10.1111/j.1467-7687.2009.00847.x>

Perry, L. K., & Samuelson, L. K. (2011). The shape of the vocabulary predicts the shape of the bias. *Frontiers in Psychology*, 2, Article 345. <https://doi.org/10.3389/fpsyg.2011.00345>

Rakison, D. H., & Yermolayeva, Y. (2010). Infant categorization. *WIREs Cognitive Science*, 1(6), 894–905. <https://doi.org/10.1002/wcs.81>

Raman, I., Raman, E., & Mertan, B. (2014). A standardized set of 260 pictures for Turkish: Norms of name and image agreement, age of acquisition, visual complexity, and conceptual familiarity. *Behavior Research Methods*, 46(2), 588–595. <https://doi.org/10.3758/s13428-013-0376-4>

Rohde, A., McCracken, M., Worrall, L., Farrell, A., O'Halloran, R., Godecke, E., David, M., & Doi, S. A. (2022). Inter-rater reliability, intra-rater reliability and internal consistency of the Brisbane Evidence-Based Language Test. *Disability and Rehabilitation*, 44(4), 637–645. <https://doi.org/10.1080/09638288.2020.1776774>

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3), 382–439. [https://doi.org/10.1016/0010-0285\(76\)90013-X](https://doi.org/10.1016/0010-0285(76)90013-X)

Rujas, I., Mariscal, S., Murillo, E., & Lázaro, M. (2021). Sentence repetition tasks to detect and prevent language difficulties: A scoping review. *Children*, 8(7),



578. <https://doi.org/10.3390/children8070578>

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, 12(4), 110–114. <https://doi.org/10.1111/1467-8721.01243>

Schwob, S. (2020). Défi de l'évaluation langagière des enfants bilingues: État des lieux des ressources disponibles en logopédie. *Travaux neuchâtelois de linguistique*, 72, 63–87. <https://doi.org/10.26034/tranel.2020.2908>

Sheng, L., & McGregor, K. K. (2010). Lexical–semantic organization in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 53(1), 146–159. [https://doi.org/10.1044/1092-4388\(2009/08-0160\)](https://doi.org/10.1044/1092-4388(2009/08-0160))

Souza, A. C. D., Alexandre, N. M. C., & Guirardello, E. D. B. (2017). Propriedades psicométricas na avaliação de instrumentos: Avaliação da confiabilidade e da validade. *Epidemiologia e Serviços de Saúde*, 26(3), 649–659. <https://doi.org/10.5123/S1679-49742017000300022>

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools*, 37(1), 61–72. [https://doi.org/10.1044/0161-1461\(2006/007\)](https://doi.org/10.1044/0161-1461(2006/007))

Spaulding, T. J., Swartwout Szulga, M., & Figueroa, C. (2012). Using norm-referenced tests to determine severity of language impairment in children: Disconnect between U.S. policy makers and test developers. *Language, Speech, and Hearing Services in Schools*, 43(2), 176–190. [https://doi.org/10.1044/0161-1461\(2011/10-0103\)](https://doi.org/10.1044/0161-1461(2011/10-0103))

Steele, S. C., & Gibbons, L. G. (2023). Vocabulary assessment techniques: Perspectives of school-based speech-language pathologists. *Communication Disorders Quarterly*, 44(2), 79–88. <https://doi.org/10.1177/15257401221088241>

Tabouret-Keller, A. (2021). L'acquisition du langage parlé chez un petit enfant en milieu bilingue. *Cahiers du plurilinguisme européen*, 13. <https://doi.org/10.57086/cpe.1362>

Thordardottir, E., Kehayia, E., Mazer, B., Lessard, N., Majnemer, A., Sutton, A., Trudeau, N., & Chilingaryan, G. (2011). Sensitivity and specificity of French language and processing measures for the identification of primary language impairment at age 5. *Journal of Speech, Language, and Hearing Research*, 54(2), 580–597. [https://doi.org/10.1044/1092-4388\(2010/09-0196\)](https://doi.org/10.1044/1092-4388(2010/09-0196))

Wechsler, D., & Naglieri, J. (2009). *WNV : Échelle non verbale d'intelligence de Wechsler et Naglieri*. Pearson.

Wiig, E., Semel, E., & Secord, W. (2019). *CELF 5 : Batterie d'évaluation des fonctions langagières et de communication*.

Williams, K. T. (2023). *EVT-3 : Évaluation du vocabulaire expressif*.

Wing, C. S. (1990). A preliminary investigation of generalization to untrained words following two treatments of children's word-finding problems [Research article]. *ASHA Wire; American Speech-Language-Hearing Association*. <https://doi.org/10.1044/0161-1461.2103.151>

Youngstrom, E. A., Van Meter, A., Frazier, T. W., Hunsley, J., Prinstein, M. J., Ong, M.-L., & Youngstrom, J. K. (2017). Evidence-based assessment as an integrative model for applying psychological science to guide the voyage of treatment. *Clinical Psychology: Science and Practice*, 24(4), 331–363. <https://doi.org/10.1111/cpsp.12207>

Zhang, J., Wang, W., Huen, J. M. Y., Lyu, J., & Lew, B. (2022). Towards higher psychometric properties: Testing the psychological strain scales with larger samples. *Archives of Suicide Research*, 26(2), 912–927. <https://doi.org/10.1080/13811118.2020.1847707>

Zhong, H. F. (2018). The relationship between receptive and productive vocabulary knowledge: A perspective from vocabulary use in sentence writing. *The Language Learning Journal*, 46(4), 357–370.

## ANNEXES

### Annexe 1 - Récapitulatif des caractéristiques psychométriques des principaux tests francophones d'évaluation du lexique

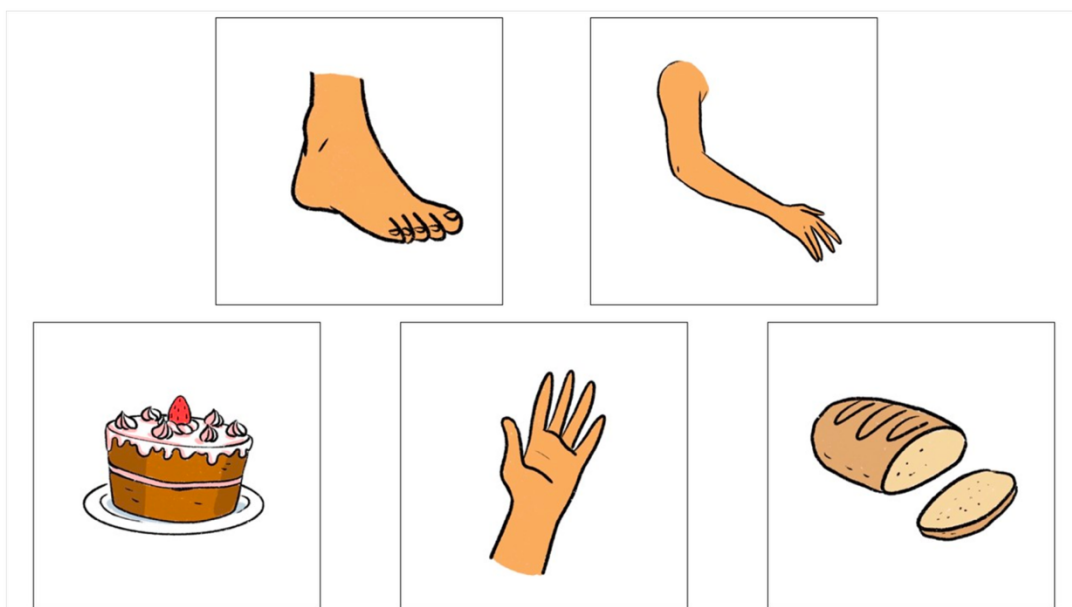
	Distribution des données	Erreur standard de mesure	Validité prédictive	Validité discriminante	Validité théorique	Validité concourante	Fidélité test- retest	Fidélité interjuge	Consistance interne	Structure interne
ELO	Non normale	Non- calculable	X	X	OK	X	X	X	X	X
L2MA2	X	Non- calculable	X	OK	OK	X	OK	X	X	X
NEE-L	X	Calculable	X	X	OK	X	Insatisfaisante	X	X	X
Exalang 5-8	X	Calculable	X	X	OK	X	Insatisfaisante	X	X	X
Evaléo 6-15	Non normale	Calculable	X	X	OK	X	X	X	Inadéquate	OK
Evalo 2-6	X	Calculable	X	X	OK	X	X	Adéquate	X	X
EVT-3	Non normale	Calculable	X	X	OK	OK	Insatisfaisante	X	X	X
PPVT-5	Non normale	Calculable	X	X	OK	OK	Insatisfaisante	X	X	X

X = Aucune donnée

Annexe 2 - Capture d'écran d'un exemple de représentation de la tâche de dénomination  
lexicale expérimentale



Annexe 3 - Capture d'écran d'un exemple de représentation de la tâche de désignation  
lexicale expérimentale



## RÉSUMÉ

Les compétences lexicales d'un individu sont essentielles dans le développement du langage oral (Ebbels et al., 2022). De nombreux tests sont à la disposition des logopèdes pour l'évaluer et diagnostiquer un éventuel Trouble Développementale du Langage (TDL) (Bishop et al., 2017). En revanche, nombre d'entre eux ne respectent pas les recommandations concernant les variables et caractéristiques psychométriques (Bragard et al., 2010; McCauley & Swisher, 1984). Face à ces observations, notre mémoire s'inscrit dans une thèse. Celle-ci, dirigée par Madame Pauline Menjot, cherche à créer, normer et valider une tâche lexicale dynamique pour l'évaluation des enfants de la 3<sup>e</sup> maternelle à la 5<sup>e</sup> primaire et participer à l'élaboration du diagnostic de TDL. Notre tâche était donc de normer et valider la tâche pour des élèves de 2<sup>e</sup> primaire et 5<sup>e</sup> primaire.

Au niveau de la méthodologie, 52 enfants de 2<sup>e</sup> primaire et 51 enfants de 5<sup>e</sup> primaire ont été recrutés suivant des critères d'inclusion et d'exclusion. Plusieurs tests standardisés ont été administrés afin de vérifier que les enfants recrutés présentaient un développement langagier dans la norme, d'évaluer la validité concourante des épreuves expérimentales, et d'obtenir des informations supplémentaires sur leur fonctionnement cognitif. De plus, deux épreuves expérimentales, une tâche de dénomination et de désignation, ont été créées et administrées via le logiciel « Psychopy » à l'ensemble des participants. Chaque épreuve comprend les mêmes 33 items, contrôlés selon plusieurs paramètres : imageabilité, âge d'acquisition, fréquence lexicale, complexité syllabique et complexité phonologique. (Bragard et al., 2010; Manoiloff et al., 2013).

En 2<sup>e</sup> primaire, les résultats indiquent une consistance interne acceptable, ainsi qu'une validité concourante satisfaisante pour les deux épreuves expérimentales. Une fidélité inter-juges jugée correcte et une distribution normale pour la tâche de dénomination mais une distribution non normale pour la tâche de désignation. De plus, l'épreuve de désignation est significativement mieux réussie que l'épreuve de dénomination. En 5<sup>e</sup> primaire, la consistance interne reste acceptable. La validité concourante est bonne pour l'épreuve de dénomination, mais s'avère insuffisante pour l'épreuve de désignation. La fidélité inter-juges se révèle acceptable et la distribution des données est normale pour les deux épreuves expérimentales. Comme en 2<sup>e</sup>

primaire, l'épreuve de désignation est significativement mieux réussie que l'épreuve de dénomination.

Enfin, une comparaison des erreurs sémantiques et phonologiques produites à la tâche de désignation entre les enfants de 2<sup>e</sup> et 5<sup>e</sup> primaire a été réalisée. Contrairement à notre hypothèse de départ, les enfants de 2<sup>e</sup> primaire produisent plus d'erreurs sémantiques, tandis que ceux de 5<sup>e</sup> primaire présentent un plus grand nombre d'erreurs phonologiques.

En conclusion, ce travail va être poursuivi dans le cadre de la thèse de Madame Menjot pour l'évaluation des enfants de la 3<sup>e</sup> maternelle à la 5<sup>e</sup> primaire dans le cadre du diagnostic de TDL. Il n'empêche que nous pourrions améliorer la sélection de participants ou des items au sein de notre tâche.

