chapitre

4

Géraldine HILAIRE-DEBOVE
Orthophoniste, docteur en Sciences du Langage
Chargée d'enseignements ISTR
(Université Claude Bernard, Lyon I)
Chargée de recherche LURCO (ERU 25 et ERU 41)

Correspondance :
debovegeraldine@cegetel.net

Pourquoi et comment évaluer les outils d'évaluation en orthophonie

Why and how to evaluate assessment tools in speech pathology

Résumé

Cet article porte sur un travail de labellisation des outils d'évaluation en orthophonie publiés en France. Cette étude est réalisée par les membres du comité directeur de l'UNADREO, société savante en orthophonie. L'équipe a mis au point une grille de labellisation permettant d'évaluer les outils d'évaluation en orthophonie. Cette grille s'appuie sur les fondements des méthodologies issues de la création de tests dont la psychologie en est le principal fondateur. Aujourd'hui, l'UNADREO a recensé 85 tests publiés entre 2001 et 2016, ces tests sont en cours de labellisation.

mots

Orthophonie, test, labellisation.

Abstract

This article is about a work of labelling of the assessment tools in speech therapy published in France. This study is conducted by the members of the Steering Committee of the UNADREO, learned society in speech therapy. The team has developed a grid of labelling assessment the assessment tools in speech therapy. This grid is based on the foundations of the methodologies from the creation of tests including the psychology is the principal founder. Today, the UNADREO has identified 85 tests published between 2001 and 2016, these tests are currently labelling.

words Speech therapy, test, labelling.

I - De Borel-Maisonny à nos jours : quels outils ?

A – Les premiers outils

Dès ses débuts, l'orthophonie s'organise rapidement autour de la notion de bilan, tout d'abord en s'appuyant sur les observations minutieuses propres aux linguistes et puis rapidement la mise en place de situations permettant de tester le patient sont réalisées. A quand remonte le premier test en orthophonie ? Sans conteste, le premier test de bilan proposé est réalisé par la pionnière de la profession, Suzanne Borel-Maisonny, on est en 1946, date de publication des « tests Borel ».

Les outils développés par S. Borel-Maisonny avaient comme objectif principal d'étayer l'observation minutieuse, et servaient de support à la fois à la description et l'interprétation des comportements verbaux et non verbaux produits au moment de leur manipulation. L'observation de l'utilisation de ces outils auprès d'enfants tout-venant ont permis dès cette période d'avoir une représentation du développement typique et ont ainsi servi à décrire les comportements déviants. On sait aujourd'hui que tous ces outils d'évaluation avaient été créés de façon ingénieuse pour permettre d'être au plus précis de la description des comportements et symptômes. On se rappelle tous des « jouets sonores, flacons emplis de liqueurs

suspectes, perles et jetons de couleur à ranger, lacet de chaussure à enfiler, boîtes à secouer, plaquettes de tailles différentes à classer, élastique servant de diapason, dessin de visage « à l' ?il de travers », canards, indiens et fédéraux à manipuler, girafe à 5 pattes, chat à queue d'écureuil… et toutes ces images séquentielles vieillottes à classer et à raconter… et ces « nostikazofimalé » et autres « rikapé » à répéter… » (Ferrand, 2002) dont certains sont encore dans nos cabinets.

Les premiers pas de l'évaluation en orthophonie ont beaucoup marqué la profession et laissent une trace indélébile. Ces premiers pas vers l'élaboration de tests sont également fortement en lien avec la naissance des diagnostics orthophoniques qui bien que n'étant pas forcément chiffrés, offrent des descriptions précises et argumentées sur lesquelles vont se greffer des traitements efficaces en lien avec les observations. Car ces tests ne sont là que pour rendre saillant ce qui apparaît intuitivement et confortent l'orthophoniste dans le choix du diagnostic et dans les objectifs de sa rééducation. Bien évidemment, le bilan n'est pas forcément normé et étalonné à l'époque, ni tout à fait bien contrôlé mis il a l'avantage de faire naître les repères d'âge qui seront utiles et qui le resteront jusqu'à nos jours.

B – Les courants théoriques qui se succèdent

Des débuts de Borel à nos jours, on cite trois courants qui se sont succédé, liés aux concepts théoriques qui se sont développés au cours du dernier siècle. On compte le courant instrumental autour des années 1970 dont les principales représentantes sont Mesdames Borel-Maisonny, Galifi'et-Grangon, Santucci ou Stamback. On cherche alors quelles sont les fonctions qui ne fonctionnent pas bien et les premiers essais de chiffres sous forme de pourcentages apparaissent influencés par les travaux en linguistique. On élabore des batteries qui tentent de « normer » le développement à partir de repères d'âge (voir l'exemple des batteries de Borel-Maisonny sur le langage oral, sur le développement du graphisme et les aptitudes non verbales).

Il y a également le courant dit pédagogique, qui se développe autour des années 1950-1970, on voit apparaître dans le domaine de la lecture la notion de « bon lecteur » et avec elle l'élaboration de test sur l'évaluation de la lecture. Apparaissent durant cette

époque les premières tentatives de chiffrer le retard en vitesse de l'enfant dyslexique. On se met à mesurer le temps de réalisation de tâches précises et on cherche à repérer le déviant à la norme.

Plus proche de nous et toujours d'actualité, on ne peut omettre de citer l'influence du courant dit cognitif/neuropsychologique qui va lancer un nouvel essor à partir des années 90, courant qui aujourd'hui permet d'avoir un langage commun entre professionnels de santé et d'obtenir une vision pluridisciplinaire autour des difficultés du patient.

Un nouveau courant est actuellement en train de s'installer, courant fortement lié au développement des nouvelles technologies, il s'agit depuis quelques années de l'apparition du support numérique sur l'axe bilan et rééducation. Dès 1999, Pierre Ferrand et Jacques Roustit réunissent un groupe d'orthophonistes pour réfléchir à la création d'un logiciel d'aide au bilan : Labo 2002 (Leloup & Roustit, 2002), outil informatisé pour choix de bilan. Depuis, beaucoup d'outils informatisés se sont développés. Nous verrons plus loin que cette tendance est en pleine expansion depuis les années 2000.

II – Labellisation des outils d'évaluation en orthophonie

A – Pourquoi une labellisation des outils d'évaluation en orthophonie

1. Se référer au modèle issu de la psychologie

Si l'on se réfère à d'autres domaines, l'apparition des tests/outils d'évaluation est souvent source de questionnement. Ces outils d'évaluation doivent être rigoureux et répondre à certaines normes. La psychologie publie des tests depuis plus de cent ans et possède davantage d'expériences (Vrignaud, 2000). Cependant, les démarches visant à garantir la qualité de ces outils ne datent eux que de seulement cinquante ans. Pour cela, la Société Française de Psychologie a mis en place une commission des tests ayant comme principale mission la mise en place de démarches garantissant les outils d'évaluation d'un point de vue nationale et internationale (Vrignaud & Loaren, 2015). Cette commission des tests de la SFP a comme principal rôle de donner des informations sur l'utilisation

des tests, de répondre et/ou de recenser les questions posées par les praticiens et/ou par les personnes testées. Cette commission permet aussi de présenter une représentation de la situation française au niveau international, mais également de la diffusion de recommandations (Vrignaud *et al.*, 2003 ; Vrignaud *et al.*, 2007).

On constate que de plus en plus de tests en France sont publiés dans le domaine de l'orthophonie sans pour autant qu'il existe de normes quant à l'élaboration de ces tests. Chaque outil se doit d'avoir un rôle dans le diagnostic et la compréhension de l'origine des troubles du patient, mais il doit aussi être irréprochable auprès d'autres instances, qu'elles soient nationales et/ou internationales, garantissant le bien-fondé de notre profession. C'est pourquoi, l'UNADREO, société savante en orthophonie, en s'appuyant sur le modèle développé par la SFP, s'est donné comme objectif de faire un état des lieux des tests publiés et de les labelliser.

2. Du côté de l'orthophonie

L'UNADREO prend appuie sur cette commission pour répondre aux mêmes questionnements mais cette fois-ci dans le domaine de l'orthophonie. En effet, au cours des vingt dernières années, de nombreux outils d'évaluation en orthophonie ont été développés en France et il est évident aujourd'hui que ces tests doivent répondre à une rigueur scientifique afin d'aider au mieux le praticien, de permettre d'objectiver le diagnostic. C'est pourquoi le groupe de travail a été mis en place depuis 2016 pour examiner les tests publiés en orthophonie. Comme le souligne T. Rousseau (2016), président de l'UNADREO : « L'objectif n'est pas de dire qu'un test est bon ou pas, sachant, par exemple, qu'un test même non standardisé peut être utile et intéressant pour la pratique orthophonique, pour faire le suivi d'un patient, pour évoluer son évolution dans le temps, juste par rapport à lui-même. Il ne s'agira pas de répondre par oui ou non sur l'utilité, la recommandation, l'intérêt des tests mais de donner une « appréciation » selon un certain nombre de critères scientifiques et aussi d'intérêt clinique. »

Le rôle de l'UNADREO est donc, comme celui de la SFP, de garantir la qualité des tests d'un point de vue national mais également international et d'informer les praticiens quant à cette qualité.

Le groupe de travail a commencé tout d'abord à réfléchir sur les critères à retenir pour étudier et labelliser les outils d'évaluation en orthophonie. Ces critères seront développés dans les paragraphes suivants. Une grille a été élaborée après plusieurs remaniements, elle est présentée dans la seconde partie.

Le groupe de travail constitué de l'ensemble des membres du comité directeur de l'UNADREO a commencé par recenser tous les outils publiés au cours des quinze dernières années. L'inventaire de ces outils est présenté en troisième partie.

La labellisation de ces tests est en cours de réalisation.

B – Définition, fonction et élaboration d'un outil d'évaluation

1. Le rôle des outils d'évaluation

La première question à laquelle nous souhaiterions répondre est la suivante : Pourquoi élabore-t-on des outils d'évaluation ?

Dans la grande majorité des cas, la création de nouveaux outils est souvent motivée par le constat d'un manque d'outil, outil nécessaire à l'argument diagnostic, et combler ce manque permet ensuite que ce nouvel outil basé sur un support théorique puisse rendre service en clinique au moment du diagnostic.

En résumé, ces outils d'évaluation permettent :

- D'aider le clinicien à faire le point sur ce qui ne fonctionne pas, et ainsi le conforter dans son intuition issue de l'anamnèse et donner des arguments en faveur de son hypothèse de diagnostic ;
- De permettre, à partir d'épreuves de bilan, d'élaborer un diagnostic argumenté qui sera ensuite transmis au patient et au médecin prescripteur. Ces outils deviennent de plus en plus nécessaires en raison de l'évolution de nos champ et décret de compétences ;
- D'élaborer de nouveaux tests et de nouvelles épreuves pour rester au plus près de l'actualité scientifique et nous permettre, nous orthophonistes, de rester compétents en matière de diagnostic dans le domaine pathologique ;

Il faut également reconnaître qu'avec l'avancée de la recherche et des besoins sociaux, il est important que nous puissions étoffer nos bilans orthophoniques pour rester le plus crédible possible au

vu des autres instances, de l'avancée scientifique en France mais également à l'étranger et d'être toujours dans une perspective de rester experts dans notre domaine et notre champ de compétence.

Grâce au développement des théories au cours des années 90 et des travaux réalisés les années précédentes, le clinicien a une meilleure connaissance des symptômes, les tests et le choix des items qui les régissent, vont se focaliser sur ces appuis scientifiques. Ainsi Coquet (2002) rappelle les travaux qui ont été à l'origine de création des premiers tests normés : par exemple, Olswang *et al.* (1998) distingue deux types d'indicateurs pour le diagnostic de retard de langage et/ou de parole chez le petit avant 3 ans, d'une part les facteurs de risques qui seront repérés lors de l'anamnèse et d'autre part les comportements langagiers. Les comportements langagiers et non verbaux présentant des scores faibles tel le vocabulaire réduit et peu diversifié (en particulier pour les verbes) ; le décalage compréhension / production et retard de compréhension ; des difficultés d'ordre phonologique (structure syllabique non respectée, erreurs sur les voyelles, limitation du nombre de consonnes et de leur variété et erreurs dans leur réalisation) etc. qui vont pouvoir être testés à partir de tests étalonnés.

On voit aussi que chaque époque se trouve influencée par les travaux scientifiques d'autres disciplines telles que la linguistique, la psychologie, la neuropsychologie etc. On voit des épreuves ajoutées aux batteries en fonction des nouvelles avancées scientifiques. Le test de bilan suit la connaissance scientifique et ne peut s'en exclure. Ce qui implique également que tout test peut devenir obsolète dès qu'un nouveau courant scientifique apparaît et remet en question les connaissances scientifiques sur lesquelles ont été basées ces tests. Il est donc important de garder en mémoire, qu'un test reste un outil pouvant être supplanté par un autre et il est du devoir de l'orthophoniste de se tenir au courant des avancées scientifiques dans ses domaines de prédilections.

2. Définition et construction d'un test

Nous reprendrons la définition proposée dans le livre blanc « la méthode des tests » éditée par les éditions ECPA (2015). Un test est selon la définition de Pierre Pichot « une situation expérimentale standardisée servant de stimulus à un comportement ; ce comportement est évalué par comparaison statistique avec celui d'autres

individus placés dans la même situation permettant ainsi de classer le sujet examiné soit quantitativement soit typologiquement ».

Selon cette définition, le sujet exposé au test se trouve dans une situation expérimentale ; donc non naturelle. Il convient donc que la passation auprès des sujets soit réalisée selon des conditions rigoureusement identiques et ce à chaque application et nécessite donc une standardisation.

Selon Laveault & Grégoire (2014), la construction d'un test est un processus long comprenant cinq étapes principales. La première étape consiste à déterminer à quoi va servir ce test.

Il existe deux catégories de tests :

- les tests normés, dont la principale fonction est de discriminer les sujets appartenant à un groupe pour lequel est construit le test ;
- les tests critériés, dont la fonction est d'évaluer « si un sujet possède ou non certaines caractéristiques prises comme référence », par exemple si telle ou telle compétence est acquise.

Le choix de l'un des deux types, critérié ou normé, influence fortement la méthodologie utilisée pour l'élaboration du test.

Il existe également une autre distinction que celle que nous venons de citer. En effet, on différencie le « test certificatif » du « test diagnostic ». Le test certificatif est basé sur les compétences d'un sujet et consiste à cerner si un enfant par exemple a acquis telle ou telle compétence en fin de cycle pour pouvoir aborder le cycle suivant. Ce type de test est plutôt utilisé dans le domaine éducatif. L'autre type de test, qui nous intéresse plus particulier, est le test diagnostique qui contrairement au précédent est beaucoup plus ciblé. Son but n'est pas de montrer qu'une compétence est ou non acquise mais plutôt de comprendre le sens d'une performance. Contrairement au test certificatif qui par exemple montrera qu'un sujet est en difficulté pour réaliser un type de tâche, le test diagnostique aura pour fonction de comprendre pourquoi cette tâche est échouée par le patient, on est au-delà de la performance puisque l'on tente de cerner les capacités cognitives sous-jacentes. Le test sera basé alors sur les processus mis en œuvre pour réaliser une tâche, par exemple on recherchera dans les processus de lecture ce qui n'est pas efficient contrairement à la norme si l'on évalue le langage écrit.

Ces tests s'appuient sur les données issues de la recherche notamment en matière de fonctionnement cognitif.

De par leurs objectifs différents, ces tests ne sont pas conçus avec la même méthodologie et nous permettent de différencier les outils utilisés par les enseignants en milieu scolaire des outils d'évaluation relevant du médical et donc réservés aux orthophonistes dans le cadre du bilan, à savoir une fonction de diagnostic à des fins de rééducation.

Lors de l'élaboration d'un test, dès que l'étape 1 a été définie, il va falloir ensuite réaliser les étapes suivantes qui sont :

Etape 2 : définir ce que l'on souhaite mesurer ;

Etape 3 : créer les items qui seront ensuite évalués lors de l'étape 4 ;

Etape 5 : consiste à déterminer les propriétés métriques du test définitif.

Il s'agit donc d'un long travail qui nécessite des étapes qui permettent, lorsqu'elles sont rigoureusement respectées, de garantir qu'un test est sensible, fiable et valide.

Enfin, selon Huteau & Lautey (1997) cité par Marin-Curtoud *et al.* (2010), un test reste un dispositif présentant quatre propriétés : il doit être standardisé, il doit permettre de situer le comportement de chaque individu en fonction d'un groupe de référence, le degré de précision des mesures qu'il permet doit être évalué (fidélité) et enfin la signification théorique ou pratique des mesures doit être précisée (validité).

3. Test et rigueur scientifique

La grille d'évaluation mise en place par l'UNADREO s'appuie sur les données issues de la littérature et les critères en matière d'élaboration des tests (Laveault & Gregoire, 2014). De manière générale, quatre critères garantissant l'outil sont généralement mentionnés, il s'agit de la standardisation, la sensibilité d'un test, sa fidélité et sa validité.

La standardisation permet de réduire les biais concernant la passation.

Les autres mesures, sensibilité, validité et fidélité garantissent la qualité du test, on parle d'ailleurs de mesures dites « indices de qualité » et concernent davantage la construction de l'outil.

- Standardisation

On entend par standardisation le fait de réduire le biais lié à l'observateur. En effet, la standardisation permet de s'assurer que les scores obtenus par un sujet sont imputables à des différences individuelles et non pas à des variations liées à la situation de passation. Pour réduire ce biais, un test est dit standardisé dès qu'il possède une bonne présentation du matériel, des consignes de passation claires et précises et un mode de notation détaillé.

Extrait « la méthode des test », ECPA (2015)

Cette standardisation doit permettre à l'usager de s'assurer que l'étalonnage est représentatif de la population de référence et doit rendre compte de comment le test a été élaboré (stratification). Il peut également présenter et expliquer le choix des items sélectionnés par exemple (sur quelle(s) base(s), sur quels outils de références, choix des variables linguistiques, choix du support de passation – dessin au trait/image/photo/écran, etc.).

- Sensibilité

La sensibilité d'un test fait référence au pouvoir de discrimination d'un test, à savoir sa capacité à distinguer les individus les uns des autres.

- Fidélité

La fidélité s'appuie sur la théorie classique des scores dont la forme actuelle est due principalement aux travaux de Gulliksen (1950), Magnusson (1967) et de Lord & Novick (1968) dont les fondements ont été publiés par Spearman (1907).

Les mesures de fidélité assurent la stabilité d'un test. Un test doit présenter une bonne stabilité temporelle et une bonne consistance interne. On peut par exemple démontrer la fidélité inter-ob-

servateurs, c'est-à-dire contrôler l'influence de l'examinateur lors de la passation d'une tâche et son influence sur la réussite ou non des items. On parlera de fidélité test-retest ou fidélité inter-observations pour référer à la constance des résultats obtenus à partir d'une tâche à travers le temps par exemple ou lorsque l'on utilise un test réalisé sous deux versions différentes. Un test fidèle permet également de mesurer le changement d'un individu.

Il est possible de mesurer la fidélité d'un test grâce au coefficient de consistance interne, on utilise l'alpha de Cronbach qui est basé sur le rapport entre la part de variance de chaque item et la variance de l'ensemble des scores de l'échelle. Selon Beech & Harding (1994) cité par Marin-Curtoud *et al.* (2010). Le coefficient de consistance interne du test, d'une valeur minimale de 0.70 devrait être obtenu à partir d'un échantillon de 100 sujets au moins.

Quant à la fidélité test-retest, elle devrait être égale à 1.00, elle reste toutefois satisfaisant à partir de 0.70. L'administration test-retest doit, selon ces mêmes auteurs, être réalisée entre un et trois mois.

- **Validité**

Le concept de validité a beaucoup évolué depuis les cinquante dernières années. Ce concept est régulièrement redéfini en fonction des courants, sa définition est publiée dans « standards for Educational and Psychological Testing ». La dernière version date de 1999, cette version souligne que la procédure de validation est une « définition du cadre conceptuel du test. Quel concept le test vise-t-il et en quoi ce concept se distingue-t-il de concepts voisins ? Le cadre conceptuel est, pour une part, défini par l'usage prévu des scores aux tests » (Laveault & Gregoire, 2014).

On parle de validité de construit lorsqu'un test repose sur une connaissance étendue des modèles théoriques ainsi que sur les données scientifiques récentes concernant l'habileté que le test est censé évaluer.

Selon de nombreux auteurs, la validité d'un test renvoie à sa pertinence, aux affirmations que l'on peut faire à partir des scores obtenus mais aussi aux éléments dont on dispose qui permettent de justifier les inférences que le testeur réalise à partir des scores.

Selon le « standards for Educational and Psychological Testing », la validité relève des concepteurs et chercheurs qui se doivent de

réaliser des études sur la validité des outils qu'ils construisent. Ils s'engagent, et cela relève de leur responsabilité, à vérifier si leurs interprétations des scores possèdent une validité suffisante, ces preuves sont rassemblées autour de cinq catégories (tableau 1).

Tableau 1. Catégories de preuves de validités selon le « standards for Educational and Psychological Testing » (1999) issu de Laveault et Gregoire (2014).

Types de preuves, basées sur…	Caractéristiques
Le contenu	Evaluation formalisée par des experts de l'ensemble des caractéristiques des items en référence à ce que le test prétend mesurer
Les processus de réponse	Evaluation de l'adéquation entre les caractéristiques visées par le test et de celles qui sont effectivement mises en œuvre par les répondants
La structure interne	Evaluation du degré de relation entre les items et les composantes du test définies par le modèle
Les relations avec d'autres variables	Evaluation du degré de liaison des scores au test avec d'autres mesures externes au test
Les conséquences du testing	Evaluation des conséquences non souhaitées de l'application du test et de l'utilisation des scores

Les preuves basées sur le contenu du test proviennent dans la majorité des cas de l'avis d'experts consultés pour évaluer les items d'un test et vérifier que les items choisis sont représentatifs du concept ou domaine visé. Il s'agit de jugements subjectifs mais qui respectent une méthodologie rigoureuse s'appuyant sur les connaissances scientifiques. Dans bien des cas, les conclusions issues de ces consultations d'experts sont confortées par le résultat de recherches réalisées ultérieurement. Elles sont donc reconnues comme solides. On parle dans ce cas de validité du contenu, qui permet de conclure que théoriquement chaque élément du test mesure bien ce qu'il est censé mesurer.

Il ne faut pas confondre validité du contenu et validité de surface ou apparente. La validité de surface s'appuie sur le jugement réalisé par des non-experts et ne se fonde pas sur une méthodologie particulière. Les juges se contentent simplement de dire si les items semblent ou non mesurer ce qu'ils sont censés mesurer. Selon Laveault & Gregoire (2014), la validité de surface permet « de créer des tests plus crédibles et mieux acceptés par les utilisateurs, car leur contenu apparaît plus légitime à ces derniers ».

La validité du contenu et la validité de surface contribuent à renforcer la validité du construit.

Pour qu'un test soit valide, il faut également vérifier « si les démarches mises en œuvre par les sujets pour produire leurs réponses correspondent bien à ce qui est prévu dans le cadre conceptuel qui sous-tend le test » (Laveault & Gregoire, 2014). Les preuves basées sur les processus de réponse s'appuient en grande partie sur une analyse détaillée des réponses individuelles. Elles consistent par exemple à un entretien post-test avec les sujets pour comprendre comment ils ont procédé pour trouver une réponse, elles peuvent également se baser sur la mesure du temps de réponses ou provenir d'éléments relevés à partir de l'analyse d'enregistrements vidéo.

L'analyse des réponses a pour objectif de comprendre ce que sous-tend une réponse en s'appuyant sur le modèle théorique sur lequel a été fondé le test. Par exemple, un test élaboré à partir du modèle de lecture à deux voies de Coltheart *et al.* (2001) prend comme postulat l'existence de deux types de voie de lecture que l'on va pouvoir tester à partir d'items particuliers : la voie de lecture dite d'assemblage qui intervient lorsque l'on décode pour la première fois un mot non connu pourra être testée à partir de mots réguliers ou des non-mots, la seconde voie, dite d'adressage permettra la lecture de mots connus déjà stockés en mémoire à long terme et que l'on active au moment de la reconnaissance, on pourra la tester notamment à partir de mots irréguliers. Les listes de mots réguliers et irréguliers sont couramment utilisées dans le modèle comme testant les deux voies de lecture, ainsi un patient qui réussit la lecture de mots réguliers et échoue sur les mots irréguliers permet d'inférer que la voie de lecture dite d'adressage n'est pas efficiente. De même si le temps mis pour lire une liste sort de la norme, on pourra éga-

lement aboutir à la même conclusion. Il convient donc dans ce cas que chaque item proposé ait bien été contrôlé (structure, fréquence, type).

Il existe également des preuves basées sur la structure interne d'un test. L'analyse de la structure interne d'un test consiste à vérifier la consistance entre les différents éléments du test et ce qu'ils évaluent en lien avec le modèle dont est issu le test. Ainsi dans de nombreux tests et notamment dans le domaine de l'orthophonie, la structure du test repose sur un modèle plus large comprenant différentes composantes qui conduisent au calcul de plusieurs scores composites mais rarement d'un score global comme cela est le cas par exemple dans les échelles d'intelligence de Weschler. Par exemple, dans la majorité des tests évaluant le langage oral, on trouve des scores sur les différentes composantes du langage (phonologie, lexique, syntaxe, pragmatique, scores sur les versants production vs compréhension) issues de modèles ayant une conception assez modulaire du langage. Malheureusement dans bien des cas il n'existe pas de score global alors que l'ensemble des épreuves permettrait d'obtenir une mesure globale du langage et de la communication.

En psychologie, il existe plusieurs types de calculs statistiques permettant de vérifier les relations entre les épreuves et le fait qu'elles soient conformes au modèle théorique, il s'agit de l'analyse factorielle, des modèles structuraux d'équations et du scalogramme de Guttman.

Le quatrième point évoqué par Laveault & Gregoire (2014), concernent les preuves basées sur les relations avec d'autres variables. Dans ce cadre, le chercheur examine les corrélations entre les scores au test et d'autres mesures à partir d'autres tests déjà existants. On parle alors de validité externe. On peut également examiner les corrélations entre les scores au test et les résultats d'examen ou les jugements d'experts. Lorsque la corrélation existe, on parlera de preuves de convergence.

Les preuves de validité peuvent être obtenues différemment : soit les deux mesures sont réalisées simultanément, on parle alors de validité concomitante, soit les mesures au test permettent de prédire des résultats qui seront obtenus ultérieurement, dans ce cas on parlera de validité prédictive. On peut utiliser comme me-

sure le coefficient de Bravais-Pearson (r>=0,40) dans le cas d'une validité concurrente ou prédictive.

Nous retiendrons que :

- La validité interne notamment lors de la construction du test, s'appuie sur des études statistiques effectuées pendant la construction du test ou sur les preuves publiées par la suite sur le test lui-même ;
- La validité externe, il s'agit de comparer le test avec d'autres tests déjà existants par exemple. Bien souvent, cela n'est pas possible puisque l'une des motivations de la création d'outil est justement le manque d'outil d'évaluation.

Enfin, pour conclure ce paragraphe, un test est dit valide s'il mesure bien ce qu'il est censé mesurer.

Si l'on se replonge dans l'usage que les orthophonistes font des tests, il faut garder à l'esprit que le principal objectif est de diagnostiquer une pathologie (validité diagnostic), on dira alors que le test a une validité empirique s'il permet effectivement d'identifier un trouble et d'en préciser sa nature, bien évidemment le diagnostic n'est possible qu'en référence à une théorie sous-jacente.

Enfin, on parle de validité écologique pour référer au postulat que les comportements observés au cours de la passation d'un test reflètent les comportements produits en situation naturelle dans le milieu naturel. Pour cela, il est souhaitable que les conditions de passation soient proches de la réalité, du milieu de vie pour permettre la généralisation. On dit que la validité écologique est grande lorsque le chercheur parvient à reproduire les conditions où l'on pourra observer le plus fidèlement les comportements réels, la validité écologique est moindre dès lors que les individus ont moins de chance de se comporter en réalité comme les sujets observés lors de l'étude.

C – Présentation de la grille « Evaluation Test en orthophonie »

1. Introduction

La grille proposée par l'UNADREO a été élaborée par l'équipe du comité directeur et s'appuie sur les critères mentionnés dans la littérature garantissant la qualité et représentativité d'un outil

d'évaluation. Il est clair pour l'ensemble de l'équipe que tous les critères d'évaluation ne peuvent à ce jour être validés en intégralité, le but n'étant pas de ne pas valider un test mais de souligner quelles sont les parties du test qui garantissent la validité et fiabilité du test de celles qui nécessiteraient des études afin d'améliorer la garantie de l'outil auprès des professionnels et des chercheurs.

Cette grille se compose de quatre parties : standardisation, sensibilité, fidélité et validité.

2. Partie « Standardisation »

Dans la partie standardisation, un intérêt particulier est porté sur :

- la présence et présentation des étalonnages,
- la stratification du test, à savoir la présence des explications concernant la construction du test,
- la présence des variables linguistiques, à savoir comment les choix ont été faits, pourquoi ces choix, ont-ils été révisés suite à une première passation auprès d'un groupe prétest,
- la représentativité de la population testée (nombre, région, catégorie socio-culturelle/niveau d'éducation, sexe, âge, critère d'inclusion), présence de groupes témoins, à savoir existe-il une étude sur une population pathologique versus groupe témoins ?

3. Partie « sensibilité »

La seconde partie concerne la sensibilité du test, à savoir si le test élaboré et les épreuves proposées permettent effectivement de situer le résultat de chaque patient dans un groupe de référence et de discriminer finement les sujets dans leur score vis-à-vis du groupe de référence. Ce test est-il présenté avec étude de cas, a-t-il été testé pour diagnostiquer des sujets pathologiques versus non-pathologiques ? Quel est son pouvoir de diagnostic ?

4. Partie « fidélité »

La partie suivante concerne la fidélité, à savoir le degré de précision des mesures qu'il permet d'évaluer, il s'agit de vérifier si la fidélité inter-observations et inter-observateurs a été contrôlée et/ou vérifiée.

5. Partie « validité »

La partie suivante s'intéresse surtout à la validité théorique, il s'agit de chercher si le test ne présente pas un biais lié à la procédure de passation, aux choix des sujets, aux choix des items et de l'outil proposé. L'UNADREO étudie également la validité de contenu, y'a-t-il eu des études de corrélation interne, de consistance interne, de cohérence ?

L'examen du choix et de la taille de la population est également analysé. Est-elle bien représentative, la taille de l'échantillon est-elle assez importante ? D'où provient l'échantillon censé représenté une population ? S'agit-il d'une région, de différentes régions, France métropole ou autre ? Tient-on compte de la variable sexe ? La population est-elle ou non répartie par milieu socio-culturel ? Des études statistiques ont-elles été réalisées pour vérifier l'influence de ces facteurs ou non (sexe, milieu etc.) ?

Quelles est la validité écologique, la validité concourante, la validité théorique ?

Tableau 2. Grille « Evaluation Test en Orthophonie ».

Grille d'évaluation élaborée par l'UNADREO
EVALUATION TEST en ORTHOPHONIE
Nom de l'outil :
Auteur(s) :

Qualités du test	Avantages	Limites
Standardisation		
Présence des étalonnages, de la stratification		
Présence des variables linguistiques		
Représentativité de la population testée		
Présence de groupes témoins		
Sensibilité		
Fidélité		
Fidélité inter-observations		
Fidélité inter-observateurs		

Validité		
Validité de contenu Coefficient de l'alpha de Cronbach		
Validité empirique / critérielle Analyse de corrélation		
Validité écologique		
Validité théorique		

Conclusion :

Intérêt clinique :

Cette grille est en cours de passation. Chaque outil sera évalué de façon aveugle par deux groupes distincts et la grille présentée ci-dessus remplie. Ensuite, une réunion de concertation permettra d'affiner et de compléter l'analyse avant de la rendre publique auprès des professionnels qui pourront s'y référer.

La partie suivante présente l'inventaire des tests qui ont été sélectionnés.

III – Inventaire des tests

A – Recensement des tests

L'équipe de l'UNADREO a recensé tous les outils publiés depuis les quinze dernières années mais également ceux plus anciens mais toujours considérés en clinique comme des outils nécessaires. Sont pris en compte aussi bien les batteries complètes de tests que des tests plus restrictifs.

85 outils d'évaluation ont été recensés et sont en cours de labellisation. Ces batteries/tests concernent l'évaluation lors de différents bilans correspondant à notre nomenclature :

- Bilan de la déglutition et des fonctions oro-myo-faciales
- Bilan de la phonation
- Bilan de la communication et du langage oral et/ou bilan d'aptitudes à l'acquisition de la communication et du langage écrit

- Bilan de la communication et du langage écrit
- Bilan de la dyscalculie et des troubles du raisonnement logico-mathématique
- Bilan des troubles d'origine neurologique
- Bilan des bégaiements et des autres troubles de la fluence
- Bilan de la communication et du langage dans le cadre des handicaps moteurs, sensoriels ou mentaux (inclus surdité, paralysies cérébrales, troubles envahissants du développement, maladies génétiques).

Sur ces 85 tests, quatre ont été publiés avant 2001 mais restent encore très utilisés par les cliniciens, les 81 restants ont été publiés entre 2001 et 2016.

Cette première partie présente l'état des lieux des tests publiés au cours des quinze dernières années.

B – Tendance développement de nouveaux tests au cours des quinze dernières années

La figure 1 présente le nombre de nouveaux tests d'évaluation en orthophonie publiés au cours des quinze dernières années. On peut tout d'abord constater que le nombre de tests s'est surtout accru au cours des cinq dernières années, soit entre 2011 et 2016.

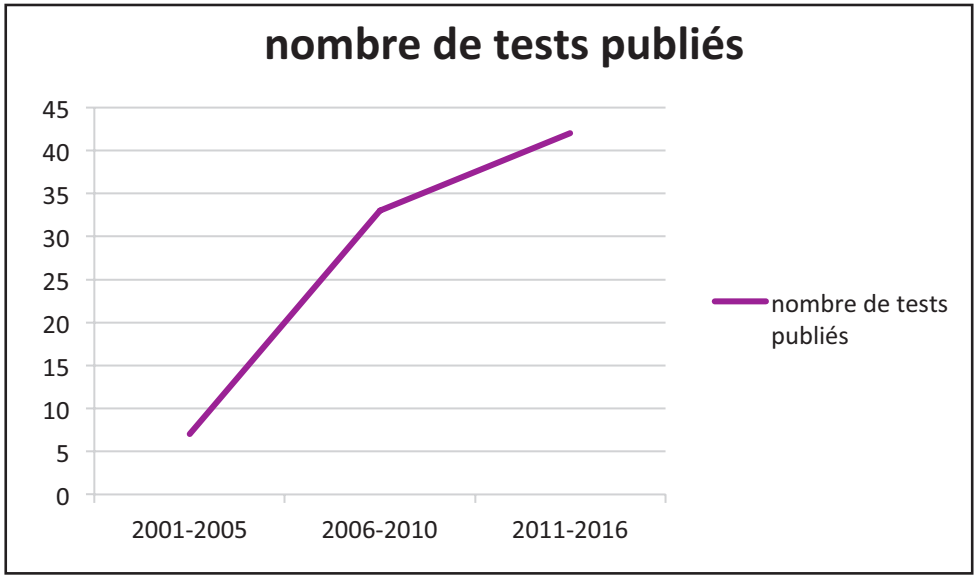

Figure 1. Nombre de tests publiés entre 2001 et 2016.

Entre 2001 et 2005, seulement sept nouveaux tests d'évaluation ont été recensés, mais la tendance est à l'accroissement puisqu'entre 2006 et 2010, 33 nouveaux tests sont apparus sur le marché et entre 2011 et 2016, 42 nouveaux outils ont été publiés.

L'élaboration des nouveaux outils d'évaluation est une réponse au manque d'outils utilisés lors du diagnostic. La figure 2 présente la répartition de ces outils d'évaluation en fonction des thèmes qu'ils abordent.

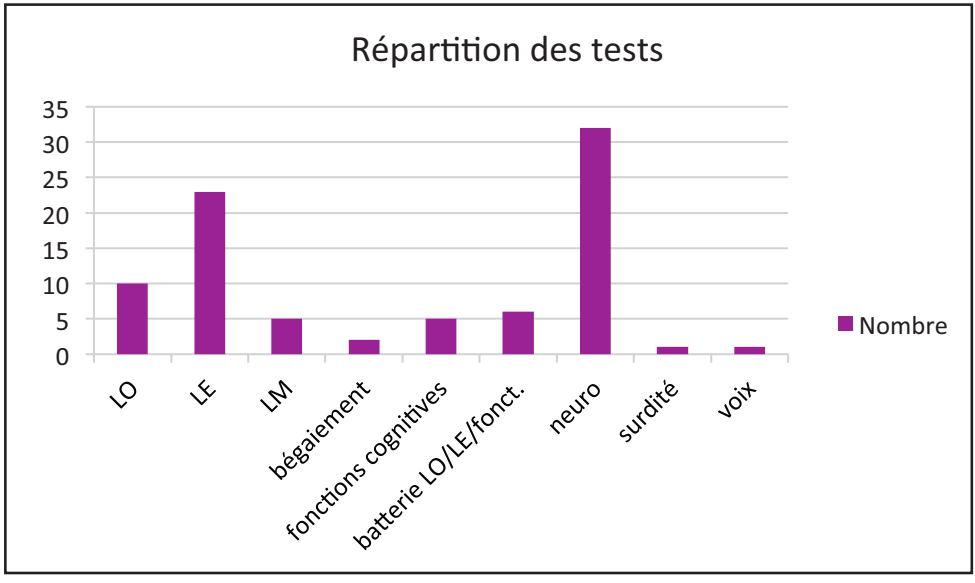

Figure 2. Répartition des tests d'évaluation en fonction des domaines.

On constate qu'au cours des quinze dernières années, de nombreux tests ont été développés en langage écrit avec le souci d'élaborer des outils permettant d'évaluation la lecture/l'orthographe pour toutes les tranches d'âge. L'augmentation concerne l'apparition d'outils pour les collégiens et lycéens dont quelques-uns pour les étudiants post-bac. Cette augmentation pour ces tranches d'âge permet de répondre aux demandes de bilan de plus en plus fréquentes pour les plus grands en raison de la mise en place des aménagements scolaires et les tiers-temps pour les examens proposés aux enfants présentant un trouble de type dys.

Concernant la catégorie Langage oral (LO), on note l'apparition d'outils élaborés pour les adolescents et notamment l'apparition d'outils pour évaluer plus finement le langage élaboré et ses subtilités.

Apparaissent également depuis peu des outils pour explorer les fonctions cognitives (mémoire, attention, praxie). Ces outils sont utilisés pour comprendre et expliquer de façon plus précise les difficultés du patient, ils permettent de repérer des profils et de révéler l'existence d'un trouble cognitif qui permettra plus facilement de poser un diagnostic (par exemple de dyslexie et/ou dysorthographie). Ne sont comptabilisés ici que les tests utilisés pour les enfants et adolescents. D'autres tests ont également été développés dans le cadre de l'aphasie, ces tests ont été classés dans la catégorie « Neurologie ».

On note également l'apparition de batteries plus globales ayant comme objet l'évaluation du langage oral, le langage écrit et les fonctions cognitives associées.

La catégorie « Neuro » comprend les tests développés pour l'adulte dans le cadre de l'aphasie, des traumatisés crâniens et des troubles neurodégénératifs. On constate que cette catégorie s'est beaucoup développée au cours des quinze dernières années avec un total de 32 outils d'évaluation publiés, ce qui est assez considérable.

On constate toujours que les parents pauvres dans le domaine de l'évaluation restent le bégaiement avec seulement deux outils développés, la surdité (1) et la voix (1).

Enfin, depuis peu, de nouveaux outils en logico-mathématique (LM) sont nés en lien avec les nouveaux courants de pensées (cognition mathématique).

C – Va-t-on vers une informatisation des outils ?

Comme le rappelle Marin-Curtoud *et al.* (2010), l'informatisation des tests a l'avantage de permettre de maîtriser la variable « temps de traitement d'une tâche », l'outil informatique permet à la fois de traiter la précision des réponses et les temps de réponse (French, 1994 ; Le Gall & Allain, 2001). Avec l'utilisation de l'outil informatique, la vitesse de traitement devient un indice de degré d'automatisation, et donne des informations sur le coût cognitif que peuvent entraîner certains traitements cognitifs. (Bonin, 2003,

Martin, 1999). Ces outils vont donc se développer de plus en plus notamment dans le cadre d'outils évaluant les fonctions cognitives. Sont présentés dans la figure 3 les pourcentages de tests « informatisés ». Nous entendons par « informatisés », les tests ayant une partie du support présenté sur écran à la totalité de la passation. Sont exclus les outils ne présentant qu'un accès à la feuille de passation ou cotation des résultats via un DVD annexe ou un site.

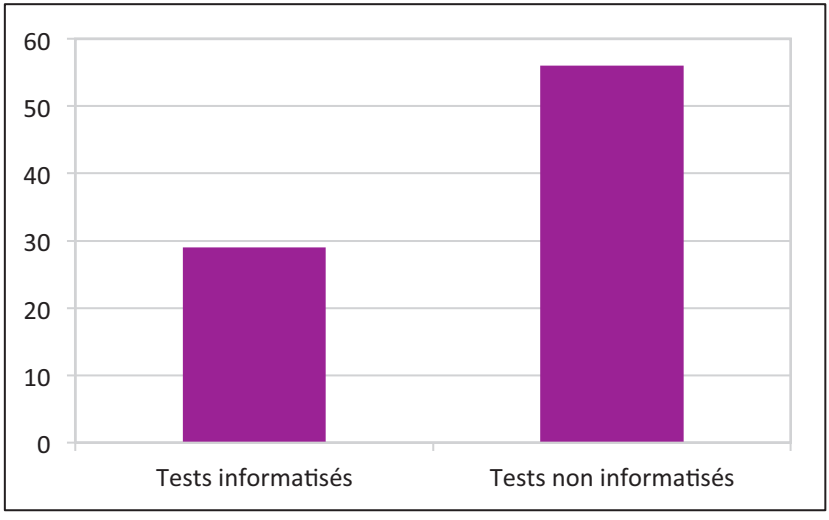

Figure 3. Répartition des tests informatisés vs non informatisés sur le nombre total de tests recensés.

On note que seulement 29 tests publiés au cours des quinze dernières années sont informatisés, soit 34 %. Les plus anciens ont une partie du support de passation réalisée sur écran alors que les plus récents, les tests sont dits totalement informatisés, du support au recueil des réponses avec accès aux scores du patient, voire au profil.

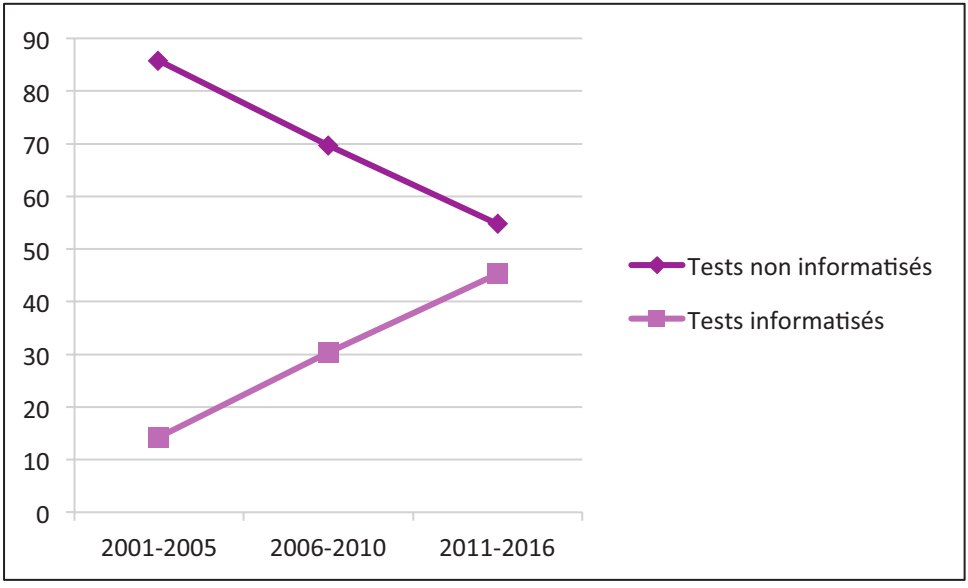

Figure 4. Pourcentage de tests informatisés vs non informatisé (évolution).

La figure 4 montre l'existence d'une tendance à informatiser les outils d'évaluation. En effet, seulement 14 % des outils étaient informatisés entre 2001 et 2005, ensuite cette tendance est passée à 30 % entre 2006 et 2010. Elle est actuellement de 45 %.

Le support actuel privilégié est l'ordinateur, peu d'outils sont actuellement disponibles sur tablette, en effet le pourcentage actuel d'outils sur tablette n'est que de 3,3 % (sur 30 tests informatisés recensés), cette tendance devrait s'accroître au cours de la prochaine décennie en raison de la praticabilité de cet outil numérique.

La question que l'on peut se poser est la suivante : existe-t-il des domaines en orthophonie pour lesquels on voit s'accroître le nombre d'outils informatisés versus non informatisés ?

La figure 5 présente la répartition des tests en fonction des domaines.

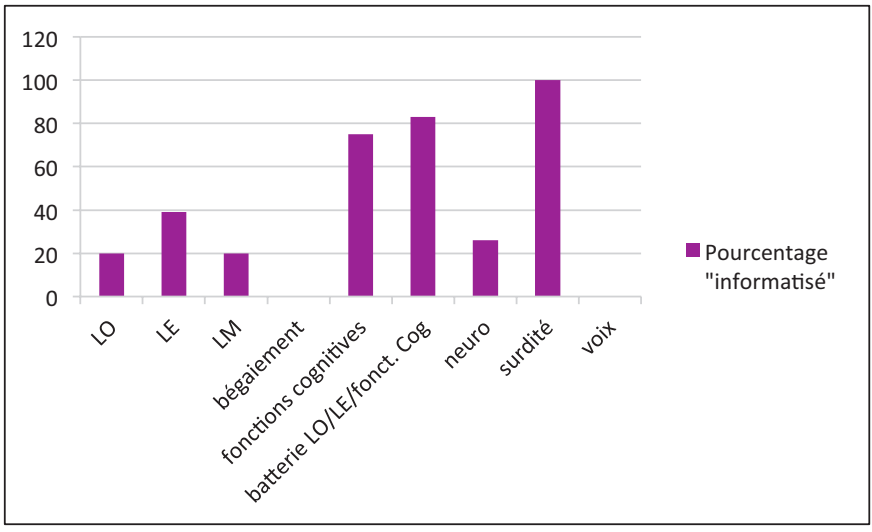

Figure 5. Répartition en pourcentage des tests informatisés en fonction des domaines.

Les domaines les plus informatisés sont les tests explorant les fonctions cognitives et les batteries explorant différents domaines. Nous laisserons de côté les domaines de surdité et voix pour lesquels seulement un test chacun a été élaboré au cours des dernières années. Concernant les autres domaines, neurologie, langage oral, langage écrit et logico-mathématique, le pourcentage reste inférieur à 40 %, la tendance n'est donc pas à l'informatisation.

Conclusion

Cet article visait à présenter le début d'un travail et d'une réflexion que mène l'équipe de l'UNADREO sur l'évaluation des outils d'évaluation en orthophonie.

Ce travail de labellisation est en cours de réalisation et devrait présenter ces premiers résultats début 2018. Ce travail se poursuivra avec l'étude des nouveaux outils qui seront publiés par la suite. Les résultats de labellisation seront disponibles sur le site de l'UNADREO. Ils permettront de répondre aux questions des professionnels quant à la qualité d'un test et ses limites et rendront une meilleure visibilité de nos pratiques auprès des autres instances nationales et internationales.

Beech, J.R. & Harding, L. (1994). *Tests, mode d'emploi, guide de psychométrie*. Paris : ECPA.

Bonin, P. (2003). *Production verbale de mots, Approche cognitive*. Bruxelles : De Boeck Université.

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

Coltheart, M, Rastle, K, Perry, C., Langdon, R. et Ziegler, J. D. (2001). A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 104-256.

Coquet, F. (2002). Le bilan orthophonique. Rééducation *orthophonique, 212,* 13-43.

ECPA (2015). Le livre blanc « la méthode des tests ». Paris : éditions ECPA.

Ferrand, P. (2002). Des souris. *Rééducation orthophonique, 212,* 3-6.

French, C. (1994). L'évaluation assistée par ordinateur. Dans J.R. Beech et L. Harding (dir), *Tests, mode d'emploi, Guide de Psychométrie* (pp. 156-167). Paris : Centre de psychologie Appliquée.

Gulliksen, H. (1950).*Theory of mental tests*. New York : Wiley.

Huteau, M. & Lautrey, J. (1997). *Les tests d'intelligence*. Paris : La Découverte, Repères 229.

Laveault, D. & Gregoire, J. (2014). *Introduction aux théories des tests en psychologie et en sciences de l'éducation*. Bruxelles : De Boeck Supérieur.

Le Gall, A., & Allain, P. (2001). Applications des techniques de réalité virtuelle à la neuropsychologie clinique. Champ psychosomatique. *L'Esprit du temps*, *22* (2), 25-38.

Leloup, G. & Roustit, J. (2002). LABO2002 : Aide au bilan orthophonique. *Rééducation orthophonique, 212,* 7-12.

Lord, F.M. et Novick, M.R. (1968). *Statistical theories of mental test scores*. Boston : Addison-Wesley.

Magnusson, D. (1967). *Test theory*. Boston : Addison-Wesley.

Marin-Curtoud, S., Rousseau, T., & Gatignol, P. (2010). Etat des lieux sur « le test » Qu'appelle-t-on un test ? Qu'est-ce qu'évaluer ? Du test au testeur… Comment franchir le pas? *L'orthophoniste*, 296, 19-26.

Olswang, B., Rodriguez B & Timler, G. (1998). Recommanding Intervention for Toddlers with Specific, Language Learning Dificulties. *American Journal of Speech Language Pathology,* 7, 23-32.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Rousseau, T. (2016). Labellisation des outils d'évaluation utilisés en orthophonie. *L'orthophoniste, 361*, 34-35

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

http://www.sfpsy.org/IMG/pdf/recomm1.pdf.

Vrignaud, P., Chevalier, A., & Paineau, A. (2007). Recommandations internationales sur les tests informatises ou les tests distribues par Internet. Document téléchargeable sur le site de la Société Française de Psychologie.

Vrignaud, P. et Loaren, E. (2015). Qualité des pratiques d'évaluation psychologique : Garantir la validité des outils, former les praticiens, *Flash info SFP*, 2-7.