Analyse psychométrique des outils d'évaluation mathématique utilisés auprès des enfants francophones

Psychometric Analysis of Mathematics Assessment Tools Used with French-speaking Children

Anne Lafay
Julie Cattini

MOTS-CLÉS
ÉVALUATION
MATHÉMATIQUE
TROUBLE DES APPRENTISSAGES EN MATHÉMATIQUES
DYSCALCULIE
ENFANT
FRANCOPHONE
PSYCHOMÉTRIE
VALIDITÉ
FIDÉLITÉ

Abrégé

Si nous nous référons au *Manuel diagnostique et statistique des troubles mentaux* (5e éd.; DSM-5; American Psychiatric Association, 2013 [version anglaise], 2016 [version française]), l'évaluation d'un enfant en difficulté mathématique doit comporter une évaluation objective. Cette évaluation vient aider le clinicien à déterminer si les compétences scolaires de l'enfant sont nettement en-dessous du niveau escompté pour l'âge chronologique. Jusqu'à présent, aucune étude ne s'est intéressée à évaluer les qualités psychométriques des tests disponibles en français pour évaluer les capacités mathématiques des enfants francophones. Pourtant, les professionnels sont amenés à faire un choix éclairé sur le ou les test(s) qu'ils utiliseront avec leurs patients. La présente étude vise, d'une part, à mettre à jour la recension des outils disponibles en français pour l'évaluation mathématique qui avait été établie par Lafay, St-Pierre et Macoir (2014) et, d'autre part, à analyser leurs qualités psychométriques. Les résultats obtenus montrent que, bien que plusieurs outils soient disponibles, peu d'entre eux répondent aux standards psychométriques. Cela remet donc en question la valeur discriminante des outils disponibles. Ainsi, cette étude promeut l'utilisation d'une pratique basée sur les données probantes pour aider les cliniciens à adopter une pratique réflexive lors du choix des tests diagnostiques.

Anne Lafay
Concordia University, Montréal, QC, CANADA

Julie Cattini
Luxembourg, LUXEMBOURG

Abstract

According to the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; *DSM-5*; American Psychiatric Association, 2013 [English version], 2016 [French version]), objective assessment is an important step in evaluating children with mathematical difficulties. It can help professionals determine whether children's academic competence is below that expected for their chronological age. Until now, no study has investigated the psychometric characteristics of French clinical tests to assess the mathematical abilities of French-speaking children. However, clinicians must carefully select the test(s) they would use to assess their patients. The present study has two goals: (a) to update the recension of French clinical tools for the assessment of mathematical abilities realized by Lafay, St-Pierre, and Macoir (2014) and (b) to analyze the psychometric properties of these tools. The results showed that although many assessment tools are available, few satisfy psychometric standards. Thus, the discriminating value of these tools is called into question. This study promotes using an evidence-based approach to help clinicians adopt a reflexive practice when selecting diagnostic tests.

L'évaluation d'un enfant en difficulté mathématique doit comporter une évaluation objective des compétences scolaires. Jusqu'à présent, aucune étude ne s'est intéressée à évaluer les qualités psychométriques des tests disponibles en français pour évaluer les capacités mathématiques des enfants francophones. La présente étude a pour objectif d'aider les orthophonistes à faire des choix éclairés dans la sélection des outils d'évaluation des habiletés mathématiques. Ces renseignements pourront également leur permettre d'être en mesure de mieux comprendre ou d'interpréter les résultats d'évaluations complétées par d'autres professionnels.

Diagnostic de trouble des apprentissages en mathématiques

En fonction du milieu professionnel dans lequel il travaille, l'orthophoniste peut être amené à travailler avec des enfants ayant un trouble de la communication ou qui sont aux prises avec un problème de langage. Parmi ces enfants, plusieurs manifestent des difficultés concomitantes en mathématiques. Par exemple, des difficultés en mathématiques ont été observées chez les enfants sourds (pour une revue, voir Roux, 2014) et chez les enfants ayant un trouble développemental du langage oral (Donlan, Cowan, Newton et Lloyd, 2007; Durkin, Mok et Conti-Ramsden, 2013). Ajoutons que le trouble des apprentissages en mathématiques – autrement appelé dyscalculie – est très fréquemment associé à la dyslexie. En effet, entre 17% (Gross-Tsur, Manor et Shalev, 1996) et 43,3-65% (Barbaresi, Katusic, Colligan, Weaver et Jacobsen, 2005) des enfants présentant un trouble des apprentissages en mathématiques sont aussi dyslexiques selon les critères diagnostiques utilisés dans chacune des études. Certains auteurs proposent que la capacité à traiter les nombres, qui sous-tend le développement des habiletés mathématiques, est innée et disponible à tous, incluant aux adultes sans culture et langage mathématique (Butterworth, Reeve, Reynolds et Lloyd, 2008; Frank, Everett, Fedorenko et Gibson, 2008; Gordon, 2004), aux bébés (Antell et Keating, 1983; Starkey et Cooper, 1980; Wynn, 1992) et même aux animaux (Brannon, 2005). D'autres proposent plutôt que les habiletés mathématiques se développent grâce à un système numérique exact et lié spécifiquement au langage humain. À titre d'exemple, Carey (2001, 2004) attribue une importance primordiale au langage dans le développement du concept de nombre. Selon cet auteur, même si les enfants perçoivent presque instantanément et de manière quasi innée les très petites quantités (c.-à-d. *subitizing*), ce n'est que grâce à l'acquisition des mots-nombres qu'ils deviennent capables d'associer une quantité précise à un mot-nombre précis. En résumé, un fait demeure : le langage a une place prépondérante dans le développement de la compréhension et dans l'application des concepts mathématiques.

D'après la définition du *Manuel diagnostique et statistique des troubles mentaux* (5 éd.; DSM-5; American Psychiatric Association, 2013 [version anglaise], 2016 [version française]), le trouble des apprentissages en mathématiques, observé chez 1% à 10% des enfants d'âge scolaire, est défini comme étant un déficit des apprentissages dans les sphères du sens du nombre, du calcul et de la résolution de problèmes qui ne peut être expliqué par des troubles d'ordre sensoriel, neurologique, psychiatrique ou environnemental. Le trouble des apprentissages en mathématiques interfère fortement avec la scolarité des enfants en étant atteints et avec les activités de la vie quotidienne impliquant des compétences numériques, et ce, de manière durable et en dépit des interventions offertes. Les symptômes décrits dans le DSM-5 sont, d'une part, des difficultés à maîtriser le sens des nombres, les données chiffrées ou le calcul et/ou, d'autre part, des difficultés avec le raisonnement mathématique (p. ex. des difficultés à appliquer des concepts ou des méthodes mathématiques pour résoudre les problèmes). En s'appuyant sur cette définition, les cliniciens doivent alors inclure, dans l'évaluation, une investigation minimale des habiletés mathématiques suivantes : dénombrement (c.-à-d. action d'indiquer le nombre d'éléments d'une collection), lecture et dictée de nombres (c.-à-d. action de passer d'un nombre écrit en code arabe [p. ex. 15] à un nombre en code oral [p. ex. quinze] et inversement), calcul (p. ex. 2 + 4, 3 x 12, 150 x 3, 1259 - 856, etc.) et résolution de problèmes à énoncé verbal (p. ex. Marie-Ève a neuf jujubes dans son sac. Elle en a trois de plus que son ami Marc-Antoine. Combien de jujubes Marc-Antoine a-t-il?).

Le cadre théorique de l'approche cognitive (Butterworth, 2005; Noël et Rousselle, 2011; Von Aster et Shalev, 2007; Wilson et Dehaene, 2007) soutient par ailleurs que les difficultés mathématiques mentionnées précédemment découlent d'une faiblesse au niveau du *sens du nombre* (c.-à-d. du traitement des nombres présentés non symboliquement) et de l'accès au *sens du nombre* via les nombres symboliques (c.-à-d. du traitement des nombres présentés symboliquement). Cela suggère, par exemple, des difficultés à comparer des nombres en format analogique (p. ex. ***) ou en format arabe (p. ex. 3), à identifier et à estimer des quantités en format analogique, ou encore, à placer des nombres sur une ligne numérique. En s'appuyant sur ces hypothèses théoriques, les cliniciens doivent alors inclure, dans l'évaluation, une investigation minimale

des habiletés du traitement des nombres symboliques et non symboliques.

Selon le DSM-5, un enfant ayant un trouble des apprentissages en mathématiques présente un niveau mathématique significativement inférieur à ce qui est attendu pour l'âge, tel qu'évalué par des tests standardisés de calcul et de raisonnement. Dans une démarche de pratique basée sur les données probantes, on mesure ici toute l'importance d'utiliser les tests les plus valides et de connaître, au préalable, leurs propriétés psychométriques afin de faire un choix éclairé (Betz, Eickhoff et Sullivan, 2013; Gaul Bouchard, Fitzpatrick et Olds, 2009; Leclercq et Veys, 2014; McCauley, 1989). Gaul Bouchard et al. (2009) ont par ailleurs expliqué que « l'Ordre des orthophonistes et audiologistes du Québec prévient leurs membres que le fait de tirer des conclusions basées sur des tests non standardisés ou qui ne possèdent pas des niveaux de qualités psychométriques appropriés va à l'encontre du code déontologique de leur ordre professionnel ». Betz et al. (2013) ont récemment investigué les habitudes d'utilisation des tests de 364 cliniciens américains et ont rapporté la présence d'un biais concernant la popularité d'un test. Ils ont en effet montré que la fréquence d'utilisation des tests standardisés pour le diagnostic du trouble développemental du langage est uniquement corrélée à la date de publication mais aucunement, malheureusement, à leurs qualités psychométriques (p. ex. fidélité, validité, pouvoir discriminant). Ce résultat est tout à fait surprenant et va à l'encontre des recommandations d'une pratique basée sur les données probantes. De plus, il convient de reconnaître que les manuels de tests ne sont pas toujours clairs et il est parfois difficile de s'y retrouver. Outre la lecture du manuel, il est donc important de savoir ce que l'on cherche.

Les qualités psychométriques d'un outil d'évaluation

Un outil d'évaluation doit respecter plusieurs propriétés psychométriques pour être considéré de bonne qualité : il doit être standardisé, valide, fidèle et posséder des données normatives.

Standardisation. Un test est standardisé lorsque les conditions de passation et de cotation ont été systématisées et uniformisées lors de l'étalonnage. De plus, le manuel est suffisamment clair et précis pour permettre à un utilisateur ultérieur de reproduire de façon identique les procédures et conditions dans sa pratique clinique. Ceci permet de limiter la subjectivité, ainsi que les erreurs de mesure ou les biais d'interprétation. Vérifier les caractéristiques de standardisation d'un test est primordial pour une utilisation optimale.

Validité. La validité d'un test réfère au degré avec lequel un test mesure vraiment ce qu'il prétend mesurer. Plusieurs types de validité peuvent être investigués. Tout d'abord, la validité de surface (ou validité d'apparence; Ivanova et Hallowell, 2013) est une mesure subjective qui concerne la compréhension et l'acceptation du test par les utilisateurs (patients et évaluateurs). Il s'agit de mesurer si l'évaluateur peut décrire l'objectif du test, s'il comprend les consignes, s'il est capable d'utiliser le test et s'il juge que la présentation du test est adéquate pour la tranche d'âge visée.

La validité de contenu (Gaul Bouchard et al., 2009), parfois appelée validité théorique (Leclercq et Veys, 2014), réfère à la pertinence du contenu du test. On ne peut pas affirmer qu'un test est valide pour toujours. La conception de l'outil et le choix des items qui le composent doivent reposer sur les modèles théoriques récents de la fonction cognitive évaluée.

La validité de critère (ou validité empirique; Ivanova et Hallowell, 2013) est la capacité d'un test à évaluer adéquatement la performance par rapport à un critère de référence (critère externe, indépendant). On distingue deux types de validation critériée : la validité concomitante (autrement appelée validité concourante par Leclercq et Veys, 2014, ou encore, validité concordante par Gaul Bouchard et al., 2009) et la validité prédictive. La validité concomitante implique une comparaison, au même moment de mesure, entre le test et un critère de référence externe (p. ex. un autre test standardisé mesurant le même construit théorique). La validité prédictive implique une comparaison, en temps différé, entre le test et un critère qui sert d'indicateur d'une performance future pour une tâche de nature similaire que l'on cherche à prédire. La pertinence fonctionnelle de l'outil doit être attestée via une concordance entre les scores observés à l'outil et le fonctionnement dans les activités de vie quotidienne mettant en œuvre la fonction évaluée (p. ex. la note obtenue en mathématiques lors des examens scolaires).

Enfin, la validité de construit réfère à la capacité d'un test ou d'une batterie de tests à mesurer un construit théorique. Plusieurs analyses empiriques peuvent être rapportées dans les manuels de test. Ces analyses devraient pouvoir s'expliquer en lien avec la théorie avancée par le test. Ce n'est donc pas uniquement la valeur des analyses effectuées qu'il importe de regarder, mais également son lien avec la définition du construit qu'il sous-tend. De ce fait, le manuel d'un test doit, au préalable, définir les objectifs du test et des sous-tests de manière claire et simple, en plus de définir le construit (c.-à-d. la définition conceptuelle, théorique ou opérationnelle de ce qui est mesuré dans le test). La validité de construit est vérifiée

par des analyses portant sur la validité en lien avec les caractéristiques de l'individu, la validité factorielle et la précision (Ivanova et Hallowell, 2013). La validité en lien avec les caractéristiques de l'individu concerne le fait que lorsque le construit mesuré est intrinsèquement relié à une ou plusieurs caractéristiques « évidentes » de l'individu, la mesure de ce construit doit être sensible à cette relation (p. ex. sexe, âge, niveau socioéconomique, pathologie, etc.). La validité factorielle est une mesure dans laquelle la structure théorique du test correspond à la structure statistique observée. En d'autres mots, différents items ou sous-tests, malgré des différences de contenu, de format ou de tâches, mesurent une dimension commune qui influence la performance des individus à tous les items ou sous-tests. Enfin, la précision ou le pouvoir discriminant (également appelé pouvoir classificatoire; Ivanova et Hallowell, 2013) d'un outil correspond à sa sensibilité et sa spécificité et doit garantir son pouvoir diagnostique. La sensibilité est le pouvoir qu'un test possède pour repérer un enfant en difficulté comme étant effectivement en difficulté (c.-à-d. un vrai positif). Selon Plante et Vance (1994), un outil est reconnu comme étant sensible s'il permet de classifier correctement une forte proportion des personnes présentant des difficultés (80% à 95%). La spécificité est le pouvoir qu'un test possède pour repérer une personne saine comme étant effectivement saine (c.-à-d. un vrai négatif). Un outil est reconnu comme étant spécifique s'il permet de classifier correctement une forte proportion des personnes ne présentant pas de difficulté (80% à 95%). En conclusion, la validité d'un instrument se détermine entre autres en évaluant dans quelle mesure le test mesure réellement ce qu'il dit vouloir mesurer. Prendre connaissance des différents éléments de validation d'un outil est une avenue indispensable pour juger de sa pertinence dans un contexte particulier.

Fidélité. La fidélité d'un test porte sur son degré de cohérence, de précision et de reproductibilité. Plusieurs types de fidélité peuvent également être investigués. Tout d'abord, la cohérence interne (Ivanova et Hallowell, 2013) concerne le fait qu'un test psychologique soit cohérent avec lui-même et que chacune de ses composantes réagisse de manière cohérente à une même réponse. Il existe plusieurs analyses empiriques qui permettent d'évaluer la cohérence interne. Tout d'abord, la cohérence interne peut être estimée par le calcul du coefficient alpha de Cronbach. Il s'agit d'une valeur calculée qui s'étend entre 0 et 1. Plus la valeur alpha s'approche de 1, plus l'ensemble des éléments est homogène. Le seuil minimal d'acceptabilité pour l'alpha de Cronbach est estimé à 0,70 (Nunnally, 1978). Il faut toutefois noter qu'un alpha de Cronbach trop élevé peut être une indication de redondance. La cohérence interne d'un test ou d'une batterie de tests peut également être évaluée à partir d'une analyse de corrélations inter-items à l'intérieur d'une même épreuve, ou encore, d'une analyse de corrélations inter-épreuves à l'intérieur d'une batterie de tests. Enfin, la cohérence interne peut être estimée par une bissection des items (ou *split-half*), ce qui consiste à partager aléatoirement un test en deux groupes d'items et à vérifier leur corrélation. Les balises utilisées sont celles indiquées par Cohen (1988), à savoir qu'une corrélation autour de 0,10 est faible, qu'une corrélation autour de 0,30 est moyenne et qu'une corrélation autour de 0,50 est forte.

Ensuite, la fidélité peut être investiguée par une évaluation de la stabilité, qui consiste à vérifier si le test donne des résultats relativement similaires (reproductibles) dans des situations différentes et comparables. La fidélité temporelle (ou test-retest; Ivanova et Hallowell, 2013) stipule que l'outil est en mesure de fournir des résultats comparables entre deux passations à des temps différents, ce qui assure que les résultats obtenus ne sont pas l'effet du hasard. La fidélité inter-juges (Ivanova et Hallowell, 2013) assure généralement que les résultats obtenus par une personne sont le reflet de sa performance, indépendamment du professionnel qui a administré et corrigé le test. Il importe donc que différents juges soient en mesure d'évaluer les performances de la même manière. Ajoutons que lorsque deux versions parallèles d'un même test existent, l'outil doit montrer que l'application de ces deux versions aux mêmes personnes résultent en des scores équivalents (fidélité par versions parallèles; Ivanova et Hallowell, 2013). En conclusion, la fidélité d'un instrument se détermine non seulement en évaluant dans quelle mesure les items censés mesurer un même construit mènent à des résultats similaires, mais également dans quelle mesure ces résultats concordent.

Normes. Une norme correspond à la distribution des scores obtenus par un échantillon de personnes, représentatif d'une population définie, à un instrument qui a été administré dans des conditions standardisées. Tout d'abord, le manuel doit faire état de la population d'étalonnage pour que l'utilisateur puisse savoir si celle-ci est représentative de la situation de son patient. Plusieurs informations sont nécessaires (Ivanova et Hallowell, 2013), ce qui inclut les caractéristiques des enfants formant l'échantillon (p. ex. l'âge et le niveau de scolarité des enfants, la répartition géographique/l'origine, la répartition des statuts socioéconomiques des parents, la proportion de filles et garçons, le nombre d'enfants présentant des difficultés intégrés dans l'échantillon des enfants sans difficulté). En bref, l'échantillon doit être décrit en précision. La taille de l'échantillon est également une variable

importante à considérer. Selon le consensus généralement établi et rapporté dans Gaul Bouchard et al. (2009) et dans Leclercq et Veys (2014), la loi de la limite inférieure exige un minimum de 100 personnes dans chaque sous-groupe. Ajoutons qu'il est important que le manuel du test précise le moment de l'étalonnage et les qualifications de l'évaluateur.

Enfin, les tests doivent faire état des mesures de tendance centrale (Ivanova et Hallowell, 2013), c'est-à-dire de la moyenne et de l'écart-type (ou des rangs centiles de performances) de l'échantillon d'étalonnage, afin d'avoir un repère quantitatif clair auquel comparer les performances des enfants, et ainsi, être en mesure de les situer par rapport à la moyenne, ou encore, de mettre en évidence leur faiblesse ou leur déficit. Le DSM-5 préconise l'utilisation de tests formels ciblés (c.-à-d. standardisés et normés) avec un seuil de performance correspondant à 1,5 écarts-types sous la moyenne, ou encore, au 7e percentile pour conclure à la présence d'un trouble des apprentissages en mathématiques. Le DSM-5 précise également qu'un seuil plus indulgent (p. ex. -1 écart-type) peut être utilisé pour identifier la présence de difficultés en mathématiques. Abondant dans une direction similaire, Green et Gallagher (2014) rapportent dans leur synthèse de la littérature que la recherche scientifique considère que des scores à des tests évaluant les habiletés mathématiques se situant sous le 10e percentile indiqueraient la présence d'un trouble des apprentissages en mathématiques (*Mathematic Learning Disabilities*), alors que des scores se situant sous le 35e percentile indiqueraient simplement la présence de difficultés mathématiques (*Mathematic Difficulties*). Ce seuil plus large a notamment sa place dans une démarche de dépistage des enfants à risque. Dans le cas de la prévention, il est effectivement préférable et plus prudent d'obtenir plus de faux positifs que de faux négatifs. Par ailleurs, un score brut d'un test normé n'est qu'une mesure approximative du score véritable de l'individu. Afin de minimiser l'impact de cette estimation, McCauley et Swisher (1984) préconisent de fixer un intervalle de confiance (IC) à 95%. Il s'agit d'un intervalle de valeurs (dépendant de l'écart-type de la distribution des scores et du degré de fidélité des tests) qui détermine 95% de chance de contenir la vraie valeur du paramètre estimé. Autrement dit, l'intervalle de confiance représente la fourchette de valeurs à l'intérieur de laquelle nous sommes certains à 95% de trouver la vraie valeur recherchée.

Outils d'évaluation des capacités mathématiques

Lafay, St-Pierre et Macoir (2014) ont réalisé une recension des outils disponibles en français pour l'évaluation mathématique et ont conclu que les professionnels ont quelques outils à disposition pour évaluer les habiletés mathématiques des enfants. Ces auteurs mettent toutefois en évidence des limitations, telles que le manque de standardisation ou de normes pour plusieurs outils, ou encore, le fait que certains outils ne s'appuient pas sur les modèles théoriques actuels du traitement numérique et ne permettent donc pas de documenter les processus déficitaires nécessaires pour diagnostiquer un trouble des apprentissages en mathématiques. Jusqu'à présent, aucune étude ne s'est intéressée à évaluer les qualités psychométriques des tests disponibles en français pour évaluer les capacités mathématiques des enfants francophones.

Objectifs

L'objectif général du présent article est d'aider l'orthophoniste à faire un choix éclairé dans la sélection des outils d'évaluation mathématique dont il a besoin, en plus de lui permettre d'être en mesure de mieux comprendre ou d'interpréter les résultats des évaluations complétées par d'autres professionnels. Pour cela, les objectifs spécifiques sont : 1) mettre à jour la recension des outils disponibles en français pour l'évaluation mathématique établie par Lafay et al. (2014) et 2) faire une analyse des qualités psychométriques des outils standardisés faisant partie de la recension.

Recension des outils

Méthodologie

La recension a porté sur les outils permettant l'évaluation des capacités mathématiques auprès de la population pédiatrique francophone. Celle-ci a d'abord été effectuée à partir des résultats de la recension de Lafay et al. (2014). Une mise à jour a ensuite été effectuée en utilisant plusieurs moyens. Une première recherche a été effectuée dans les bases de données PubMed et PsycInfo à l'aide des mots-clés « évaluation » et « mathématiques ». Une recherche identique a également été réalisée sur le site de la revue orthophonique Glossa, puisque celle-ci n'est pas référencée dans les bases de données mentionnées précédemment. Toutefois, la plupart des outils d'évaluation ne sont pas référencés dans les bases de données scientifiques. De ce fait, les catalogues des grandes maisons d'édition de tests (c.-à-d. Édition du centre de psychologie appliquée et Pearson) et des maisons d'édition spécialisées dans le matériel orthophonique (c.-à-d. Ortho Édition, HappyNeuron, Orthopratic et Cogilud) ont été consultés. Les catalogues ont été parcourus page à page dans les rubriques concernant l'évaluation et les mathématiques. Finalement, des chercheurs dans le domaine de la psychologie ou de l'éducation en mathématiques, ainsi que des cliniciens (orthophonistes, neuropsychologues et

orthopédagogues), ont été consultés dans le but d'identifier d'autres outils utilisés. La recherche a été menée afin de repérer les outils permettant le dépistage et l'évaluation des difficultés mathématiques auprès la population pédiatrique francophone édités entre l'année 1990 et février 2017.

Résultats

La recension de Lafay et al. (2014) avait mené à l'identification de 25 outils : trois échelles d'intelligence, six outils d'évaluation du rendement scolaire et 15 outils spécialisés dans l'évaluation mathématique cognitive. La présente recherche a mené à l'identification de six outils supplémentaires, pour un total de 31 outils. En effet, les quatre outils suivants ont été repérés dans les catalogues des maisons d'édition : *Evaluation Des fonctions cognitives et des Apprentissages de 4 à 11 ans* (Billard et Touzin, 2012), *Épreuve verbale d'aptitudes cognitives* (Flessas et Lussier, 2003), *Test diagnostique des compétences de base en mathématiques pour les enfants du CE2 à la 5ème* (Tedi-MATH Grands; Noël et Grégoire, 2015) et *Examath 8-15 : batterie informatisée d'examen des habiletés mathématiques* (Examath 8-15; Lafay et Helloin, 2016). De plus, deux articles présentant les données de normalisation en franco-québécois d'outils existants, soit la *Batterie pour l'évaluation du traitement des nombres et du calcul chez l'enfant* (Von Aster, 2006; Lafay, St-Pierre et Macoir, 2016) et le *Tempo Test Rekenen* (De Vos, 1992; Lafay, St-Pierre et Macoir, 2015), ont été repérés.

Parmi les 31 outils répertoriés, neuf outils ont été retirés à la suite d'application de critères d'exclusion. D'abord, les trois échelles d'intelligence suivantes ont été retirées, car leur sous-test mathématique ne pouvait donner sens sans le contexte complet de l'échelle d'intelligence : *Batterie pour l'examen psychologique de l'enfant* (2e éd.; Kaufmann et Kaufmann, 2008), *Échelle d'intelligence de Wechsler pour enfants et adolescents* (4e éd.; Weschler, 2005b) et *Nouvelle échelle métrique de l'intelligence* (Cognet, 2006). Ajoutons que le public lecteur visé par cet article est principalement l'orthophoniste et celui-ci ne fait pas passer d'échelles d'intelligence. Ensuite, deux outils ont été retirés parce que le manuel n'était pas disponible pour consultation, soit la *Batterie d'épreuves pour l'école élémentaire* (Savigny, 2001) et les *Tests d'acquisitions scolaires mathématiques* (Riquier, 1997). Deux autres ont été retirés, car il s'agissait d'épreuves totalement descriptives : *Épreuve de décision logique* (publiée dans Ménissier, 2011) et *Difficultés en mathématiques, évaluation et rééducation* (Koppel, 1998). Enfin, deux outils ont été retirés parce que le manuel était rédigé dans une langue autre que le français, soit le *Keymath™3 diagnostic assessment: Canadian edition* (Connoly, 2008) et le *Test de Calcul de Courtrai Révision 2006* (Baudonck, Debusschere, Dewulf, Samyn et Vercaemst, 2006). Cela résulte en un total de 22 outils ayant été inclus dans la présente étude. Parmi ceux-ci, 14 sont des outils évaluant uniquement les habiletés mathématiques, alors que les huit autres sont des batteries de langage ou de rendement scolaire comportant un ou quelques sous-tests d'évaluation mathématique. Le tableau 1 présente les caractéristiques générales des tests : le titre, le(s) auteur(s), la date de publication, les informations sur la modalité de présentation (informatisée ou papier), ainsi que les caractéristiques de la population et du moment d'étalonnage. Les batteries évaluant uniquement les habiletés mathématiques permettent d'évaluer les enfants âgés de 4 ans 0 mois à 17 ans 11 mois. Huit sont normées pour la population française, trois pour la population belge francophone, trois pour la population franco-québécoise et un pour la population suisse francophone.

Les outils mesurent les habiletés mathématiques (c.-à-d. dénombrement, numération, transcodage, calcul, vocabulaire mathématique, résolution de problèmes, raisonnement) et les habiletés cognitives de traitement du nombre. Les domaines mathématiques couverts varient toutefois d'un outil à l'autre. Nous renvoyons le lecteur au tableau 2 pour le détail des domaines mathématiques couverts par chaque outil.

Analyse des qualités psychométriques des outils

Méthodologie

Un total de vingt-deux outils ont été soumis à l'analyse des qualités psychométriques. Une grille d'analyse a été construite pour les besoins de l'étude à partir d'une synthèse de plusieurs références traitant du sujet des qualités psychométriques d'outils d'évaluation (c.-à-d. Gaul Bouchard et al., 2009; Ivanova et Hallowell, 2013; Leclercq et Veys, 2014). Ces références ont été choisies pour deux raisons principales : elles portaient sur un domaine proche de celui de la présente étude, à savoir un domaine de l'orthophonie (c.-à-d. le langage oral), et elles apportaient une analyse rigoureuse d'autres outils d'évaluation. Gaul Bouchard et al. (2009) ont utilisé une grille composée de 16 critères tirés des recommandations de McCauley et Swisher (1984). Leclercq et Veys (2014) ont quant à eux employé une grille composée de 13 critères, ceux-ci également tirés des recommandations de McCauley et Swisher (1984). Si certains critères sont communs dans les deux grilles, certains ne sont présents que dans l'une ou l'autre. Finalement, Ivanova et Hallowell (2013) ont décrit certains autres critères supplémentaires, tels que la nécessité d'évaluer la validité de surface, la structure du test

Tableau 1. Relevé des caractéristiques concernant la population d'étalonnage pour les différentes batteries d'évaluation analysées[1]

Nom du test ou de la batterie	Auteur(s)	Date	Présentation informatisée	Âge (années)	Classe[2]	Échantillon (nombre total)	Nombre de groupe(s)	Nombre par groupe	Pays	Comparaison avec autre(s) pays	Niveau socioéconomique	Moment de l'étalonnage
Batteries ou tests évaluant spécifiquement les habiletés mathématiques												
B-LM	Métral	2008	Non	5-8	GSM-CE1	299	14	11 à 31	France	/	/	Janvier-avril
ECPN	Duquesne	2003	Non	4-9	/	132	5	Non indiqué (26 en moyenne)	France	/	/	/
ERLA	Legeay, Morel et Voye	2009	Non	/	/	/	/	/	/	/	/	/
Examath 8-15	Lafay et Helloin	2016	Oui	8-15	CE2-3e collège	443	5	74 à 127	France et Belgique	Québec	Répartition proche des indices INSEE	Mars-mai
MathEval	Heremans	2011	Oui	/	3e maternelle-2e primaire	65	3	65 par groupe environ	Belgique	/	Selon l'auteur, échantillon plutôt favorisé, non représentatif des normes INSEE	Avril à juin
Numerical	Gaillard	2000	Non	7-10	2e-4e primaire	280	2	126 à 154	Suisse	Comparaison entre France, Finlande et Argentine	/	Février
Protocole du calcul élémentaire	Ménissier	2003	Non	7-11	CE1-CM2	406	4	91 à 109	France	/	/	Octobre-novembre
Tedi-MATH	Van Nieuwenhoven, Grégoire et Noël	2001	Non	/	MSM-CE2 / 2e maternelle-3e primaire	583	8	67 à 76	France et Belgique	Le manuel indique qu'il n'y a pas de différence entre France et Belgique mais données absentes	/	Novembre et mai
Tedi-MATH Grands	Noël et Grégoire	2015	Les deux	/	CE2-5e collège	254	5	46 à 56	France	/	INSEE	Mars-juin
TTR	Lafay, St-Pierre et Macoir	2015	Non	8-9	3e primaire	77	1	77	Canada francophone (Québec)	Comparaison qualitative avec échantillon du manuel Pays-Bas	Répartition selon Indices de défavorisation du MELS	Février-juin
UDN-II	Meljac et Lemmel	1999	Non	4-11	/	420	8	49 à 57	France	/	/	/
WIAT-II	Wechsler	2005a	Non	6-17 ans 11 mois	1ère primaire-5e pecondaire	294 si âge, 304 si classe	12	44 à 56 si classe, 18 à 32 si âge	Canada francophone (Québec)	Le manuel indique une comparaison Franco-Québécois et Franco-Ontariens mais données absentes	Information sur le niveau d'études des parents	Février à décembre
ZAREKI-R	Von Aster	2006	Non	6-11 ans et demi	CP-CM2	249	5	43 à 59	France	/	50% ZEP, 28% ne parlent pas français à la maison	Janvier à mars
ZAREKI-R	Lafay, St-Pierre et Macoir	2016	Non	8-9	3e Primaire	81	1	81	Canada francophone (Québec)	Comparaison qualitative avec échantillon du manuel France	Répartition selon Indices de défavorisation du MELS	Février-juin

Batteries générales comportant un ou plusieurs sous-tests mathématiques												
ECHAS	Simonart	1998a	Non	/	3e-6e primaire	1013	4	201 à 327	Belgique	/	/	Mai-juin
EDA	Billard et Touzin	2012	Non	4-11 (6-11 pour math)	MSM-CM2 (CP-CM2 pour math)	626	6	94 à 111	France	/	Répartition homogène selon indices INSEE	Septembre-juin
EVAC	Flessas et Lussier	2003	Non	8-14	CE2-3e collège	886 à 919	7	109 à 154 si âge, 113 à 143 si classe	France	Québec	/	Premier trimestre de l'année
Exalang 3-6	Thibault et Helloin	2006	Oui	2 ans 8 mois-5 ans 10 mois	MSM-GSM / 2e-3e maternelle	468	6	59 à 96	France et Belgique	/	Répartition proche des indices INSEE	/
Exalang 8-11	Thibault, Lenfant et Helloin	2012	Oui	8-11	CE2-CM2	461	3	93 à 150	France	/	Répartition proche des indices INSEE	Février-avril
Exalang 11-15	Thibault, Helloin et Lenfant	2009	Oui	11-15	6e collège-3e collège	322	4	85 à 97	France	/	Répartition proche des indices INSEE	Janvier-mai
N-EEL	Chevrie-Muller et Plaza	2001	Non	3 ans 7 mois-8 ans 7 mois	PSM-CE2	541	5	108 à 109	France	/	Information sur la catégorie socioprofessionnelle des parents	Septembre-juin
PEDA 1C	Simonart	1998b	Non	/	1ère-2e primaire	232 à 290	3	232 à 290	Belgique	/	/	Mai-juin

Note. [1]Les tests sont classés en ordre alphabétique selon leur titre. [2]Équivalent des classes entre la France, la Belgique, la Suisse et le Québec : MSM en France = 2e année de maternelle en Belgique, 1ère année de maternelle en Suisse et prématernelle au Québec; GSM en France = 3e année de maternelle en Belgique, 2e année de maternelle en Suisse et maternelle au Québec; CP en France = 1ère année du primaire en Belgique, en Suisse et au Québec; CE1 en France = 2e année du primaire en Belgique, en Suisse et au Québec; CE2 en France = 3e année du primaire en Belgique, en Suisse et au Québec; CM1 en France = 4e année du primaire en Belgique, en Suisse et au Québec; CM2 en France = 5e année du primaire en Belgique et au Québec, 5e année de transition en Suisse; 6e collège en France = 6e année du primaire en Belgique et au Québec et 6e année de transition en Suisse; 5e collège en France = 1ère secondaire en Belgique, en Suisse et au Québec; 4e collège en France = 2e secondaire en Belgique, en Suisse et au Québec; 3e collège en France = 3e secondaire en Belgique, en Suisse et au Québec; Seconde Lycée en France = 4e secondaire en Belgique et au Québec et 1ère secondaire degré 2 en Suisse; Première Lycée en France : 5e secondaire en Belgique et au Québec et 2e secondaire degré 2 en Suisse. B-LM = Mallette B-LM cycle II; CE1 = Cours élémentaire 1; CE2 = Cours élémentaire 2; CM1 = Cours moyen 1; CM2 = Cours moyen 2; CP = Cours préparatoire; ECHAS = Échelle d'apprentissages scolaires primaires; ECPN = Epreuves Conceptuelles de résolution des Problèmes Numériques; EDA = Evaluation Des fonctions cognitives et Apprentissages de 4 à 11 ans; ERLA = Mallette Bilan; EVAC = Épreuves verbale d'aptitudes cognitives; Exalang 3-6 = Exalang 3-6 : batterie d'examen des fonctions langagières chez l'enfant de 3 à 6 ans; Exalang 8-11 = Exalang 8-11 : bilan informatisé pour l'examen du langage et des compétences transversales chez l'enfant de 8 à 11 ans; Exalang 11-15 = Exalang 11-15 : batterie informatisée pour l'examen du langage oral, du langage écrit et des compétences transversales chez le collégien; Examath 8-15 = Examath 8-15 : batterie informatisée d'examen des habiletés mathématiques; GSM = Grande section de maternelle; INSEE = Institut national de la statistique et des études économiques; MELS = Ministère de l'éducation et de l'enseignement supérieur; MSM = Moyenne section de maternelle; N-EEL = Nouvelles Épreuves pour l'Examen du Langage; Numerical = Numerical : test neurocognitif pour l'apprentissage du nombre et du calcul; PEDA 1C = Tests pédagogiques de premier cycle primaire; PSM = Petite Section de Maternelle; Tedi-MATH = Test diagnostique des compétences de base en mathématiques; Tedi-MATH Grands = Test diagnostique des compétences de base en mathématiques pour les enfants du CE2 à la 5ème; TTR = Tempo Test Rekenen; UDN-II = Construction et Utilisation du Nombre (2e éd.); WIAT-II = Wechsler Individual Achievement Test (2e éd.); ZAREKI-R = Batterie pour l'évaluation du traitement des nombres et du calcul chez l'enfant; ZEP = Zone d'éducation prioritaire.

Tableau 2. Domaines évalués par les différentes batteries d'évaluation analysées[1]

Nom du test ou de la batterie	Traitement cognitif du nombre	Dénombrement	Numération et transcodage	Calcul	Résolution de problèmes	Langage et raisonnement
Batteries ou tests évaluant spécifiquement les habiletés mathématiques						
B-LM	+/-	+	+	+	+	+
ECPN		+				
ERLA		+	+ (transcodage)		+	+
Examath 8-15	+	+	+	+	+	+
MathEval	+	+	+ (transcodage)	+		
Numerical	+	+	+ (transcodage)	+		+ (vocabulaire math)
Protocole du calcul élémentaire				+		
Tedi-MATH	+	+	+	+	+	
Tedi-MATH Grands	+		+	+	+	+
TTR				+		
UDN 2			+ (transcodage)	+		+ (vocabulaire math)
WIAT-II				+	+	+
ZAREKI-R	+	+	+ (transcodage)	+	+	
Batteries générales comportant un ou plusieurs sous-tests mathématiques						
ECHAS				+	+	+ (vocabulaire math)
EDA	+	+	+	+	+	+
EVAC						+
Exalang 3-6		+				+ (vocabulaire math)
Exalang 8-11					+	+ (vocabulaire math)
Exalang 11-15						+
N-EEL						+ (vocabulaire math)
PEDA 1C				+		

Note. [1]Les tests sont classés en ordre alphabétique selon leur titre. B-LM = Mallette B-LM cycle II; ECHAS = Échelle d'apprentissages scolaires primaires; ECPN = Epreuves Conceptuelles de résolution des Problèmes Numériques; EDA = Evaluation Des fonctions cognitives et Apprentissages de 4 à 11 ans; ERLA = Mallette Bilan; EVAC = Épreuves verbale d'aptitudes cognitives; Exalang 3-6 = Exalang 3-6 : batterie d'examen des fonctions langagières chez l'enfant de 3 à 6 ans; Exalang 8-11 = Exalang 8-11 : bilan informatisé pour l'examen du langage et des compétences transversales chez l'enfant de 8 à 11 ans; Exalang 11-15 = Exalang 11-15 : batterie informatisée pour l'examen du langage oral, du langage écrit et des compétences transversales chez le collégien; Examath 8-15 = Examath 8-15 : batterie informatisée d'examen des habiletés mathématiques; N-EEL = Nouvelles Épreuves pour l'Examen du Langage; Numerical = Numerical : test neurocognitif pour l'apprentissage du nombre et du calcul; PEDA 1C = Tests pédagogiques de premier cycle primaire; Tedi-MATH = Test diagnostique des compétences de base en mathématiques; Tedi-MATH Grands = Test diagnostique des compétences de base en mathématiques pour les enfants du CE2 à la 5ème;TTR = Tempo Test Rekenen; UDN-II = Construction et Utilisation du Nombre (2e éd.); WIAT-II = Wechsler Individual Achievement Test (2e éd.); ZAREKI-R = Batterie pour l'évaluation du traitement des nombres et du calcul chez l'enfant.

par une analyse factorielle, etc. La mise en commun de ces travaux a ainsi mené à l'élaboration de la présente grille. Au total, 21 critères ont été établis comme étant des caractéristiques de base devant être considérées par le clinicien avant d'utiliser un test dans le but de poser un diagnostic ou d'émettre une décision clinique à propos d'une performance d'un enfant à un test. Ces critères sont présentés et expliqués dans le tableau 3.

L'un des critères présentés dans le tableau 3 est la validité de contenu. Celle-ci a été établie à partir de la lecture d'ouvrages de référence en cognition mathématique qui font état des lieux des modèles théoriques actuels du développement mathématique et du trouble des apprentissages en mathématiques chez l'enfant (Butterworth, 1999, 2005; Cappelletti et Fias, 2016; Dehaene, 2010; Habib, Noël, George-Poracchia et Brun, 2011; Habib, 2014; Kadosh et Dowker, 2015), afin d'identifier si les outils évaluant spécifiquement les habiletés mathématiques ou les sous-tests de batteries plus générales s'appuyaient sur les modèles du trouble des apprentissages en mathématiques définis dans ces ouvrages. Dans le cas d'une batterie de langage comportant quelques sous-tests mathématiques, la validité de contenu a été établie non pas pour la batterie au complet, mais pour les sous-tests en question. En particulier, nous avons accordé 1 point à un outil faisant référence à un ouvrage précis et à un modèle actuel basé sur les données probantes, 0,5 point à un outil manquant de précision (c.-à-d. indiquant une référence basée sur des données probantes mais n'expliquant pas le modèle précis) et 0 point à un outil s'appuyant sur un modèle théorique non reconnu par la littérature actuelle ou à un outil ne précisant aucune référence.

Un test a été considéré comme satisfaisant un critère si le manuel présentait, dans son entier, suffisamment d'informations en lien avec le critère en question pour en permettre l'évaluation. Dans ce cas, 1 point a été attribué. Au contraire, si aucune information n'était donnée, aucun point n'a été attribué. Dans certains cas, nous avons décidé d'accorder seulement 0,5. Par exemple, si le manuel stipulait qu'une des validités considérées dans l'analyse psychométrique avait été vérifiée mais qu'aucune donnée chiffrée ne permettait de réellement approuver la présence du critère, 0,5 point était accordé en guise de confiance aux auteurs. Ce même score (0,5 point) était également attribué si le manuel donnait les informations relatives à un critère, mais les données statistiques révélaient des résultats non significatifs ou faibles (p. ex. une corrélation faible). Enfin, dans trois cas, nous avons décidé d'attribuer 0,75 point, car les manuels indiquaient que les tests remplissaient presque totalement le critère. Par exemple, un score de 0,75 point a été attribué à l'outil Examath 8-15 pour le critère « taille de l'échantillon », car certains groupes dépassaient le seuil de 100 enfants alors que d'autres en était proche (p. ex. 87).

Deux types de totaux ont ainsi été calculés. Premièrement, un score de qualité a été attribué à chaque test en utilisant la formule suivante : nombre de critères que le test satisfait / 21 critères au total * 100. Pour ce faire, le nombre de critères remplis pour chaque test a été additionné, chaque critère possédant une importance relative équivalente dans la présente grille constituée (c.-à-d. 1 critère = 1 point). Toutefois, dans les faits, certains critères nous semblent avoir une plus grande importance (p. ex., la sensibilité/spécificité versus la validité d'apparence). Une note totale sur 21 a été attribuée et le pourcentage correspondant calculé. Deuxièmement, un score de validation a été calculé pour chaque critère en utilisant la formule suivante : nombre de tests satisfaisant le critère / 22 tests au total * 100. Pour ce faire, le nombre de tests remplissant chaque critère a été additionné, chaque critère possédant une importance relative équivalente (c.-à-d. 1 test = 1 point). Une note totale sur 22 a été attribuée et le pourcentage correspondant calculé.

L'analyse psychométrique a été effectuée par deux juges (auteures de l'article) qui sont toutes deux orthophonistes et impliquées dans la recherche et/ou la pratique orthophonique basée sur les données probantes. Chacune a suivi une formation de base sur les qualités psychométriques des outils d'évaluation dans son cursus de formation continue professionnelle. Tout d'abord, les deux juges ont évalué chaque outil, de manière séparée et à l'aveugle, à partir de la consultation des manuels, de la consultation des sites commerciaux et des échanges avec les auteurs (quand ceux-ci ont bien accepté de répondre aux interrogations suscitées par la lecture des manuels). Ensuite, le premier juge a vérifié l'adéquation des points attribués par elle-même et son co-juge (et inversement). L'analyse a d'abord montré une adéquation globale de 87% entre les deux juges. Par la suite, les deux juges ont discuté et parcouru à nouveau les manuels ensemble pour parvenir à un consensus complet. C'est d'ailleurs à ce moment qu'ont été définies les cotations intermédiaires de 0,5 ou 0,75 précédemment explicitées.

Tableau 3. Synthèse des critères et recommandations concernant les outils d'évaluation

Critères			Explications
QUALIFICATIONS DE L'ÉVALUATEUR			Les qualifications de la personne qui va administrer le test, le corriger et l'interpréter sont clairement explicitées afin de garantir la validité des résultats.
STANDARDISATION	Consignes de passation et de cotation		Les consignes d'administration et de cotation sont clairement spécifiées dans le manuel afin de minimiser la subjectivité lors de l'administration et de la cotation.
VALIDITÉ	De surface		L'outil est recevable par les utilisateurs. Il s'agit de la compréhension et de l'acceptation du test par les utilisateurs (patients et évaluateurs).
	De contenu	Validité théorique	La conception de l'outil et le choix des items qui le composent reposent sur les modèles théoriques récents de la fonction cognitive évaluée. Dans le cas d'une batterie de langage comportant quelques sous-tests mathématiques, la validité de contenu a été établie, non pas pour la batterie au complet, mais pour les sous-tests en question : 1 point si référence précise et modèle actuel sur la base des données probantes, 0,5 point si référence manquant de précision mais basée sur des données probantes et 0 point si outil s'appuyant sur un modèle théorique non appuyé par la littérature actuelle ou à un outil ne précisant aucune référence.
		Objectif des tests précisé	Les concepteurs posent un choix clair concernant l'objectif de leur outil (diagnostic, détermination d'un niveau de sévérité, orientation thérapeutique) et le précisent.
	De critère	Concomitante	L'outil montre une bonne corrélation entre ses résultats et ceux d'autres épreuves mesurant les mêmes fonctions cognitives et ayant prouvé leur pertinence diagnostique : 1 point est attribué si le test indique les analyses effectuées et si les corrélations indiquées sont moyennes (autour de 0,3) ou bonnes (autour de 0,5); 0,5 point est attribué si les corrélations indiquées sont faibles (c.-à-d. autour de 0,1).
		Prédictive	La pertinence fonctionnelle de l'outil est attestée via une concordance entre les scores observés à l'outil et le fonctionnement dans les activités de vie quotidienne mettant en œuvre la fonction évaluée (p. ex., la note scolaire en mathématique) : 1 point est attribué si le test indique les analyses effectuées et si les corrélations indiquées sont moyennes (autour de 0,3) ou bonnes (autour de 0,5); 0,5 point est attribué si les corrélations indiquées sont faibles (c.-à-d. autour de 0,1).
	De construit	Relations avec les caractéristiques individuelles	Lorsque le construit mesuré est intrinsèquement relié à une ou plusieurs caractéristiques « évidentes » de l'individu, la mesure du construit est sensible à cette relation (sexe, âge, intelligence, etc.) : 1 point est attribué si le test a été évalué selon au moins deux caractéristiques (p. ex., genre, âge, classe, niveau socio-économique, latéralité, pays, présence de trouble ou non); 0,5 point est attribué si le test a été évalué selon une seule caractéristique.
		Validité factorielle	Différents items ou sous-tests (malgré des différences de contenu, de format ou de tâches) mesurent une dimension commune qui les influence tous. Le test a la capacité d'établir des associations statistiques entre ses items (ou sous-tests) en conformité avec les dimensions (facteurs) supposément mesurées : 1 point est attribué si une analyse factorielle a été analysé et met en évidence des facteurs de regroupement correspondant aux modèles théoriques apportés par les auteurs et aux regroupements en modules effectués par les auteurs.
		Sensibilité/ spécificité	Le pouvoir discriminant de l'outil, c'est-à-dire sa sensibilité et sa spécificité, a fait l'objet d'analyses spécifiques (incluant notamment une population en difficulté), afin de garantir son pouvoir diagnostique. Il s'agit du calcul des pourcentages de vrais positifs, vrais négatifs, faux positifs et faux négatifs : 1 point est attribué si le test indique une sensibilité et une spécificité supérieure à 0,80; 0,5 point si le test indique une sensibilité et une spécificité inférieure à 0,80 est attribué; aucun point n'est attribué si cela n'est pas été testé.
FIDÉLITÉ	Stabilité	Temporelle	L'outil fait preuve d'une fidélité test-retest suffisante afin de garantir la stabilité des résultats dans le temps. Friberg (2010) recommande un coefficient de corrélation de 0,90. Un coefficient de 0,80 est acceptable.
		Versions parallèles	L'outil montre que deux versions du même test aux mêmes personnes sont équivalentes. L'équivalence d'un test indique à quel point les scores fournis sont indépendants du contenu spécifique des items qui composent le test : 1 point est attribué si le test indique ses analyses et si les corrélations indiquées sont moyennes ou bonnes; 0,5 point est attribué si les corrélations indiquées sont faibles.
		Inter-juges	L'outil fait preuve d'une fidélité inter-juges suffisante afin de garantir que les résultats obtenus sont les plus objectifs possibles et indépendamment de la personne qui a administré et corrigé le test : 1 point a été attribué si les corrélations indiquées étaient égales ou supérieures à 0.90 (recommandation de Friberg, 2010) ou si le Kappa de Cohen indiqué était égal ou supérieur à 0,60 (recommandation de Fleiss, 1981).
	Cohérence interne	Corrélations	Le test montre les corrélations obtenues entre chacun des items du test et le score total au test, ainsi qu'entre les scores totaux de chaque module du test : 1 point est attribué si le test indique ses analyses et si les corrélations indiquées sont moyennes ou bonnes; 0,5 point est attribué si les corrélations indiquées sont faibles.
		Bissection	L'outil fait preuve de fidélité par bissection, technique qui consiste à diviser un test (une seule version) en deux parties « équivalentes » afin de calculer un « sous-score » pour chacune de ces parties. Les deux parties doivent être corrélées : 1 point est attribué si le test indique ses analyses et si les corrélations indiquées sont moyennes ou bonnes; 0,5 point est attribué si les corrélations indiquées sont faibles.
		Cohérence inter-items	Le manuel fait état d'une analyse statistique de la pertinence des items inclus dans les épreuves, notamment en démontrant la cohérence interne (alpha de Cronbach). Le seuil minimal d'acceptabilité étant estimé à 0,70.
NORMES	Taille de l'échantillon (nombre par groupe)		La taille de l'échantillon d'étalonnage est suffisamment importante : 1 point est attribué si l'échantillon comporte au minimum 100 participants par tranche d'âge/sous-groupe; 0,5 point est attribué si l'échantillon comporte au minimum 80 participants par tranche d'âge/sous-groupe; 0,75 point est attribué si certains groupes sont au-dessus de 100 et certains autres à 80.
	Description de l'échantillon		Les caractéristiques géographiques, socioéconomiques, linguistiques, l'âge et le genre de la population de l'échantillon d'étalonnage sont clairement explicités : 1 point est attribué si le manuel précise au moins deux caractéristiques (p. ex., genre, âge, classe, niveau socioéconomique, latéralité, pays, présence de trouble ou non); 0,5 point est attribué si le manuel indique une seule caractéristique.
	Représentativité de l'échantillon		Les caractéristiques géographiques, socioéconomiques, linguistiques, l'âge et le genre de la population de l'échantillon d'étalonnage sont représentatives de la population tout venant : 1 point est attribué si l'échantillon est représentatif sur au moins deux caractéristiques; 0,5 point est attribué si l'échantillon est représentatif sur une seule caractéristique.
	Mesures de tendance centrale		Les moyennes et écarts-types de l'échantillon d'étalonnage sont mentionnés pour chaque tranche d'âge (et/ou les percentiles si la distribution des scores n'est pas gaussienne).
	Intervalle de confiance		L'intervalle de confiance (IC) à 95% est rapporté pour chaque norme calculée. C'est un intervalle de valeurs qui a 95% de chance de contenir la vraie valeur du paramètre estimé.

Résultats

Les tableaux 4a et 4b présentent une synthèse des caractéristiques psychométriques des 22 outils analysés : la note de 1 point, 0,75 point, 0,50 point ou 0 point est indiquée dans chaque case.

Aucun critère n'a été rempli par l'ensemble des tests. La moyenne des scores de validation des critères est plutôt faible (35%; écart-type = 27%). Les scores attribués s'étendent de 0% (fidélité de type stabilité : versions parallèles) à 93% (description de l'échantillon). Cinq critères obtiennent un score de validation au-dessus de 75%, deux obtiennent un score de validation entre 75% et 50%, cinq obtiennent un score de validation entre 50% et 25% et neuf obtiennent un score de validation en-dessous de 25%. Les critères les plus respectés sont la qualification de l'évaluateur, la standardisation de l'outil (consignes de passation et cotation), la précision de l'objectif des tests, la description de l'échantillon d'étalonnage et la présence de mesures de tendance centrale. En revanche, les critères les moins respectés sont la validité de surface, la validité de critère concomitante, la validité de construit évaluant la conception factorielle de l'outil, la sensibilité et la spécificité de l'outil, la fidélité de type stabilité (temporelle, versions parallèles, inter-juges) et la cohérence interne (bissection et cohérence inter-items).

Aucun test n'obtient un score de qualité de 100%. La moyenne des scores de qualité est plutôt faible (39%; écart-type = 14%). Les scores attribués pour les qualités psychométriques des outils s'étendent de 12% à 67%. Aucun n'obtient un score de qualité au-dessus de 75%, cinq obtiennent un score de qualité entre 75% et 50%, onze obtiennent un score de qualité entre 50% et 25% et six obtiennent un score de qualité en-dessous de 25%.

Discussion

La présente étude avait pour objectif d'aider les orthophonistes à faire un choix éclairé dans la sélection des outils d'évaluation mathématique dont ils ont besoin, en plus de leur permettre d'être en mesure de mieux comprendre ou d'interpréter les résultats d'évaluations complétées par d'autres professionnels. Pour cela, les objectifs spécifiques étaient de : 1) mettre à jour la recension des outils disponibles en français pour l'évaluation mathématique établie par Lafay et al. (2014) et 2) faire une analyse des qualités psychométriques des outils standardisés faisant partie de la recension.

Premier objectif : recension des outils. Relativement au premier objectif, la présente étude a permis de mettre en évidence l'existence de 22 outils disponibles en français pour l'évaluation mathématique, dont 14 outils évaluant spécifiquement les habiletés mathématiques et huit batteries de langage ou de rendement scolaire comportant un ou quelques sous-tests d'évaluation mathématique. Les outils évaluant spécifiquement les habiletés mathématiques permettent d'évaluer les enfants âgés de 4 ans 0 mois à 17 ans et 11 mois. Huit sont normés pour la population française, trois pour la population belge francophone, trois pour la population franco-québécoise et un pour la population suisse francophone. Les domaines mathématiques couverts sont les habiletés mathématiques (dénombrement, numération, transcodage, calcul, vocabulaire mathématique, résolution de problèmes, raisonnement) et les habiletés cognitives de traitement du nombre. Ils varient toutefois d'un outil à un autre.

Deuxième objectif : analyse des qualités psychométriques des outils. Concernant les qualités psychométriques des outils d'évaluation mathématique recensés, la présente étude a permis de mettre en évidence que certains critères sont très bien considérés alors que d'autres ne le sont peu ou pas. De plus, l'analyse a montré que les outils n'ont pas tous un score de qualité psychométrique global équivalent.

Tout d'abord, l'analyse montre qu'aucun critère n'est pris en compte par l'ensemble des tests. De manière générale, les outils que les professionnels ont à leur disposition respectent plutôt bien les critères suivants : la qualification de l'évaluateur, la standardisation de l'outil (consigne de passation et de cotation), la précision de l'objectif de l'épreuve, la description de l'échantillon d'étalonnage et la présence de mesures de tendance centrale. Ces critères sont les plus simples à considérer et à mettre en place d'après Gaul Bouchard et al. (2009). En revanche, neuf des 21 critères ont été presque totalement négligés dans les tests disponibles en français, soit la validité de surface (ou d'apparence), la validité de critère concomitante, la conception factorielle de l'outil, la sensibilité et la spécificité de l'outil, la fidélité de type stabilité (temporelle, versions parallèles, inter-juges) et la cohérence interne (bissection et consistance inter-items). En particulier, les outils actuels ont généralement une bonne spécificité mais la sensibilité est limitée. Une explication potentielle est le fait qu'aucune mesure n'est réalisée avec un groupe comportant suffisamment d'enfants en difficulté. Or, des scores-seuils devraient être calculés avec la distribution d'une population de personnes saines et de personnes en difficulté. L'outil Examath 8-15 est le seul outil, dans cette recension, à fournir des données concernant la sensibilité et la spécificité de la batterie. En effet, le manuel présente l'étude des performances de 126 enfants, dont 63 présentaient des difficultés mathématiques et 63 avaient un développement

Tableau 4a. Relevé des caractéristiques psychométriques des différentes batteries d'évaluation analysées : qualification, standardisation et validité[1]

Nom du test ou de la batterie	Qualification de l'évaluateur	Standardisation : consignes de passation et de cotation	Validité							
			De surface	De contenu		De critère		De construit		
				Validité théorique	Objectif des tests précisé	Concomitante	Prédictive	Relations avec les caractéristiques individuelles	Validité factorielle	Sensibilité/ spécificité
Batteries ou tests évaluant spécifiquement les habiletés mathématiques										
B-LM	1	1	0	0	1	0	0	0	0	0
ECPN	0	0,5	0	1	1	0	0	0,5	0	0
ERLA	1	0,5	0	0	1	0	0	0	0	0
Examath 8-15	1	1	0,5	1	1	0	1	1	0	0,5
MathEval	0	1	1	1	1	0	1	0,5	0	0
Numerical	0	1	0	0,5	1	0,5	0	1	1	0
Protocole du calcul élémentaire	0	0	0	1	1	0	0	0	0	0
Tedi-MATH	1	1	0	0,5	1	0	0,5	0	0	0
Tedi-MATH Grands	1	1	0	1	1	0	0,5	0	1	0
TTR	1	1	0	1	1	1	1	1	0	0
UDN-II	1	1	0	0	1	0	0	0	0	0
WIAT-II	1	1	0	0	1	0	0	1	0	0
ZAREKI-R	1	1	0	1	1	0,5	1	0,5	1	0
ZAREKI-R (article)	1	1	0	1	1	0	0	1	0	0
Nombre de tests remplissant le critère	10	12	1,5	9	14	2	5	6,5	3	0,5
% de tests remplissant le critère	71%	86%	11%	64%	100%	14%	36%	46%	21%	4%
Batteries générales comportant un ou plusieurs sous-tests mathématiques										
ECHAS	1	1	0	0	0	0	0	1	0	0
EDA	1	1	0	1	1	0	0	1	0	0
EVAC	1	1	0	0	1	0	1	1	0	0
Exalang 3-6	1	1	0,5	0	1	0,5	0,5	0,5	0	0
Exalang 8-11	1	1	0,5	0	1	0,5	0,5	1	0	0
Exalang 11-15	1	1	0,5	0	1	0,5	0,5	1	0	0
N-EEL	1	1	0	0	1	0	0	0,5	0	0
PEDA 1C	1	1	0	0	0	0	0	0	0	0
Nombre de tests remplissant le critère	7	7	1,5	1	6	1,5	2,5	5	0	0
% de tests remplissant le critère	100%	100%	19%	13%	75%	19%	31%	75%	0%	0%
Tout test										
Nombre de tests remplissant le critère	18	20	3	10	20	3,5	7,5	12,5	3	0,5
% de tests remplissant le critère	82%	91%	14%	45%	91%	16%	34%	57%	14%	2%

Note. [1]Les tests sont classés en ordre alphabétique selon leur titre. B-LM = Mallette B-LM cycle II; ECHAS = Échelle d'apprentissages scolaires primaires; ECPN = Epreuves Conceptuelles de résolution des Problèmes Numériques; EDA = Evaluation Des fonctions cognitives et Apprentissages de 4 à 11 ans; ERLA = Mallette Bilan; EVAC = Épreuves verbale d'aptitudes cognitives; Exalang 3-6 = Exalang 3-6 : batterie d'examen des fonctions langagières chez l'enfant de 3 à 6 ans; Exalang 8-11 = Exalang 8-11 : bilan informatisé pour l'examen du langage et des compétences transversales chez l'enfant de 8 à 11 ans; Exalang 11-15 = Exalang 11-15 : batterie informatisée pour l'examen du langage oral, du langage écrit et des compétences transversales chez le collégien; Examath 8-15 = Examath 8-15 : batterie informatisée d'examen des habiletés mathématiques; N-EEL = Nouvelles Épreuves pour l'Examen du Langage; Numerical = Numerical : test neurocognitif pour l'apprentissage du nombre et du calcul; PEDA 1C = Tests pédagogiques de premier cycle primaire; Tedi-MATH = Test diagnostique des compétences de base en mathématiques; Tedi-MATH Grands = Test diagnostique des compétences de base en mathématiques pour les enfants du CE2 à la 5ème; TTR = Tempo Test Rekenen; UDN-II = Construction et Utilisation du Nombre (2e éd.); WIAT-II = Wechsler Individual Achievement Test (2e éd.); ZAREKI-R = Batterie pour l'évaluation du traitement des nombres et du calcul chez l'enfant.

Tableau 4b. Relevé des caractéristiques psychométriques des différentes batteries d'évaluation analysées : fidélité et normes[1]

Nom du test ou de la batterie	Fidélité						Normes					Score de qualité psychométrique	
	Stabilité			Cohérence interne									
	Temporelle	Versions parallèles	Inter-juges	Corrélations	Bissection	Cohérence inter-items (alpha de Cronbach)	Taille de l'échantillon	Description de l'échantillon	Représentativité de l'échantillon	Mesures de tendance centrale	Intervalle de confiance	Nombre de critères remplis par chaque test	% de critères remplis par chaque test
Batteries ou tests évaluant spécifiquement les habiletés mathématiques													
B-LM	0	0	0	0	0	0	0	1	0,5	0	0	4,5	21%
ECPN	0	0	0	0	0	0	0	1	0	0	0	4	19%
ERLA	0	0	0	0	0	0	0	0	0	0	0	2,5	12%
Examath 8-15	0,5	0	1	0,75	0	0	0,75	1	1	1	1	14	67%
MathEval	0	0	0	1	0	1	0	0,5	0	1	0	9	43%
Numerical	0	0	0	1	0	1	1	1	0	1	0	10	48%
Protocole du calcul élémentaire	0	0	0	0	0	0	1	1	0	0	0	4	19%
Tedi-MATH	0	0	0	0	0	1	0	1	0,5	1	1	8,5	40%
Tedi-MATH Grands	0	0	0	0,75	0	0,5	0	1	1	1	1	10,75	51%
TTR	0	0	0	0	0	0	0	1	1	1	0	10	48%
UDN-II	0	0	0	0	0	0	0	1	0	0	0	4	19%
WIAT-II	0	0	0,5	1	1	0	0	1	1	1	1	10,5	50%
ZAREKI-R	0	0	0	0,5	0	0	0	1	0,5	1	0	10	48%
ZAREKI-R (article)	0	0	0	0	0	0	0,5	1	1	1	0	8,5	40%
Nombre de tests remplissant le critère	0,5	0	1,5	5	1	3,5	3,25	12,5	6,5	9	4	/	/
% de tests remplissant le critère	4%	0%	11%	36%	7%	25%	23%	89%	46%	64%	29%	/	/
Batteries générales comportant un ou plusieurs sous-tests mathématiques													
ECHAS	0	0	0	0	0	0	1	1	0	1	0	6	29%
EDA	0,5	0	0	0	0	0	1	1	1	1	0	9,5	45%
EVAC	0	0	0	0	0	0	1	1	0	1	0	8	38%
Exalang 3-6	0,5	0	0	0	0	0	0	1	1	1	0	8,5	40%
Exalang 8-11	0,5	0	0,5	0,5	0	0	1	1	1	1	1	12	57%
Exalang 11-15	0,5	0	0,5	0	0	0	0,5	1	1	1	1	11	52%
N-EEL	0,5	0	0	0	0	0	1	1	1	1	0	8	38%
PEDA 1C	0	0	0	0	0	0	1	1	0	1	0	5	24 %
Nombre de tests remplissant le critère	2,5	0	1	0,5	0	0	6,5	8	5	8	2	/	/
% de tests remplissant le critère	31%	0%	13%	6%	0%	0%	81%	100%	63%	100%	25%	/	/
Tout test													
Nombre de tests remplissant le critère	3	0	2,5	5,5	1	3,5	9,75	20,5	11,5	17	6	/	/
% de tests remplissant le critère	14%	0%	11%	25%	5%	16%	44%	93%	52%	77%	27%	/	/

Note. [1]Les tests sont classés en ordre alphabétique selon leur titre. B-LM = Mallette B-LM cycle II; ECHAS = Échelle d'apprentissages scolaires primaires; ECPN = Epreuves Conceptuelles de résolution des Problèmes Numériques; EDA = Evaluation Des fonctions cognitives et Apprentissages de 4 à 11 ans; ERLA = Mallette Bilan; EVAC = Épreuves verbale d'aptitudes cognitives; Exalang 3-6 = Exalang 3-6 : batterie d'examen des fonctions langagières chez l'enfant de 3 à 6 ans; Exalang 8-11 = Exalang 8-11 : bilan informatisé pour l'examen du langage et des compétences transversales chez l'enfant de 8 à 11 ans; Exalang 11-15 = Exalang 11-15 : batterie informatisée pour l'examen du langage oral, du langage écrit et des compétences transversales chez le collégien; Examath 8-15 = Examath 8-15 : batterie informatisée d'examen des habiletés mathématiques; N-EEL = Nouvelles Épreuves pour l'Examen du Langage; Numerical = Numerical : test neurocognitif pour l'apprentissage du nombre et du calcul; PEDA 1C = Tests pédagogiques de premier cycle primaire; Tedi-MATH = Test diagnostique des compétences de base en mathématiques; Tedi-MATH Grands = Test diagnostique des compétences de base en mathématiques pour les enfants du CE2 à la 5ème; TTR = Tempo Test Rekenen; UDN-II = Construction et Utilisation du Nombre (2e éd.); WIAT-II = Wechsler Individual Achievement Test (2e éd.); ZAREKI-R = Batterie pour l'évaluation du traitement des nombres et du calcul chez l'enfant.

typique. Leclercq et Veys (2014) déploraient aussi l'absence d'un « critère diagnostique crucial, le pouvoir discriminant des outils » lors de l'analyse des outils d'évaluation du langage pour la population francophone. Le manque d'informations ou de moyens des concepteurs de tests sont des causes possibles à ces absences.

De plus, les qualités psychométriques ne sont pas identiques pour tous les outils. Le constat est identique à celui fait par Leclercq et Veys (2014) lors de l'analyse d'outils d'évaluation du langage pour la population francophone : les outils diagnostiques à la disposition des professionnels ne rencontrent pas l'ensemble des critères psychométriques recommandés pour une pratique de qualité. En effet, la présente analyse montre qu'aucun test n'obtient un score de qualité de 100%, ou encore, au-dessus de 75%. Si certains auteurs ont fait de gros efforts quant aux qualités psychométriques de leur outil (p. ex. les auteurs de l'Examath 8-15, du Tedi-MATH Grands, du *Wechsler Individual Achievement Test* [WIAT-II], de *l'Exalang 8-11 : bilan informatisé pour l'examen du langage et des compétences transversales chez l'enfant de 8 à 11 ans* et de l'*Exalang 11 – 15 : batterie informatisée pour l'examen du langage oral, du langage écrit et des compétences transversales chez le collégien*), peu d'outils obtiennent un score de qualité supérieur à 50%. Ce résultat mène à une recommandation de privilégier les trois outils évaluant spécifiquement les habiletés mathématiques qui obtiennent un score de qualité supérieur à 50% (c.-à-d. l'Examath 8-15, le Tedi-MATH Grands et le WIAT-II).

La présente étude met ainsi en lumière l'existence d'un écart considérable entre les qualités psychométriques des outils d'évaluation présentées et celles souhaitées. Elle permet, en cela, d'aider le clinicien à faire un choix éclairé dans la sélection des outils dont il a besoin, dans la limite des choix existants. Plusieurs critères sont primordiaux à considérer : la standardisation, la validité, la fidélité et les caractéristiques de normalisation. Selon Gaul Bouchard et al. (2009), « puisque ces tests sont utilisés par des professionnels de différents domaines, ces critères sont importants car ils assurent d'obtenir des informations plus objectives. Conséquemment, les décisions cliniques qui en découlent risquent moins d'être influencées par la manière dont les praticiens conceptualisent et interprètent les construits évalués ». La validité de contenu est particulièrement importante dans la mesure où il ne fait pas sens, dans une démarche de pratique basée sur les données probantes, d'utiliser un test employant des tâches reposant sur un modèle théorique invalide. De même, le pouvoir discriminant (autrement dit la sensibilité d'un test) parait des plus indispensables, puisque c'est la qualité qui permet d'attester qu'un outil d'évaluation permet de repérer l'ensemble des enfants présentant un trouble des apprentissages en mathématiques.

Limitations

L'étude s'est confrontée à trois limitations principales. La première concernait la difficulté à retracer les outils disponibles, car très peu ont fait l'objet de publications scientifiques et ils n'étaient donc pas recensés dans les bases de données scientifiques. La deuxième concernait la difficulté à retrouver les informations dans les manuels, car ceux-ci ne possédaient pas tous la même structure et n'utilisaient pas nécessairement le même vocabulaire. De plus, si le constat est fait que tous les outils ne remplissent pas toutes les qualités psychométriques évaluées dans la présente étude, il convient de nuancer le propos. En effet, le score de qualité est en fonction des critères retenus dans cet article. Il ne s'agit pas d'une valeur absolue, mais bien d'un résultat en fonction de la grille élaborée pour les besoins de la présente étude. Celle-ci donne une indication relative. Enfin, une autre limite de l'étude concernait l'importance relative équivalente des critères évalués (tel que discuté précédemment). En effet, deux tests pourraient obtenir un résultat équivalent en pourcentage, mais détenir des caractéristiques psychométriques bien différentes (dont certaines pourraient être plus importantes que d'autres lorsque vient le temps de choisir un test). Il semble alors indispensable de ne pas uniquement s'en tenir au résultat final pour caractériser les qualités psychométriques d'un outil et pour faire un choix d'outil, mais bien d'analyser l'ensemble des critères.

Conclusion

Jusqu'à présent, aucune étude ne s'était intéressée à évaluer les qualités psychométriques des tests disponibles en français pour l'évaluation des capacités mathématiques des enfants francophones. La présente étude est donc tout à fait originale et pertinente dans le contexte de la pratique orthophonique basée sur les données probantes. Vingt-deux outils ont été recensés et leurs propriétés psychométriques ont été analysées. L'étude a mis en évidence le fait que certains critères sont très bien considérés (p. ex. la standardisation) alors que d'autres ne le sont peu ou pas (p. ex. le pouvoir discriminant). De plus, tous les outils n'ont pas un score de qualité psychométrique global équivalent. Parmi les 22 recensés, seulement trois outils évaluant spécifiquement les habiletés mathématiques obtiennent un score de qualité supérieur à 50% (c.-à-d. l'Examath 8-15, le Tedi-MATH Grands et le WIAT-II). Il faut toutefois noter que tout critère n'a pas la même importance. Il semble alors indispensable de ne pas

s'en tenir uniquement au score global pour caractériser les qualités psychométriques d'un outil et pour faire un choix d'outil, mais bien d'analyser l'ensemble des critères.

Quelques recommandations générales peuvent être développées. À l'avenir, il est indispensable que les futurs concepteurs d'outils d'évaluation mathématique fassent l'effort de développer des outils standards, d'investiguer la validité et la fidélité des outils et de donner un maximum de précision quant à l'échantillon d'étalonnage et les normes dans les manuels d'utilisation, pour une plus grande transparence. De plus, tout comme il existe un canevas général tacitement accepté et utilisé pour la rédaction d'articles scientifiques, un canevas général concernant la rédaction de manuels de tests devrait être développé et utilisé par les concepteurs et les maisons d'édition. Chaque manuel devrait ainsi détailler les aspects liés à 1) la standardisation, 2) la validité, 3) la fidélité et 4) la normalisation de l'outil.

De même, il est indispensable que les cliniciens considèrent l'ensemble de ces critères pour juger des outils valides et pertinents à utiliser (Betz et al., 2013 ; Gaul Bouchard et al., 2009; Leclercq et Veys, 2014; McCauley, 1989). Néanmoins, l'obstacle principal à la mise en place d'une pratique basée sur les données probantes, d'après les informations recueillies auprès d'orthophonistes provenant de différents pays, reste le manque de temps (Durieux et al., 2016; O'Connort et Pettigrew, 2009; Zipoli et Kennedy, 2005). Aider les cliniciens à analyser les tests est donc essentiel pour que ces derniers soient conscientisés à la répercussion de l'absence de certaines qualités psychométriques sur leur pratique clinique.

Références

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5e éd.). Arlington, VA : Author.

American Psychiatric Association. (2016). DSM-5 : *manuel diagnostique et statistique des troubles mentaux* (5e éd.). Issy-les-Moulineaux, France : Elsevier Masson.

Antell, S. E. et Keating, D. P. (1983). Perception of numerical invariance in neonates. *Child Development, 54*, 695–701. doi:10.2307/1130057

Barbaresi, W. J., Katusic, S. K., Colligan, R. C., Weaver, A. L. et Jacobsen, S. J. (2005). Math learning disorder: Incidence in a population-based birth cohort, 1976-82, Rochester, Minn. *Ambulatory Pediatrics, 5*, 281–289. doi:10.1367/A04-209R.1

Baudonck, M., Debusschere, A., Dewulf, B., Samyn, F. Vercaemst, V. et Desoete, A. (2006). *Test de Calcul de Courtrai Révision 2006*. Courtrai, Belgique : Revalidatiecentrum Overleie.

Betz, S. K., Eickhoff, J. R. et Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools, 44*, 133–146. doi:10.1044/0161-1461(2012/12-0093)

Billard, C. et Touzin, M. (2012). *Evaluation Des fonctions cognitives et des Apprentissages de 4 à 11 ans*. Isbergues, France : Ortho Édition.

Brannon, E. M. (2005). What animals know about numbers. Dans J. I. D. Campbell (dir.), *Handbook of mathematical cognition* (p. 85–107). New-York, NY : Psychology Press.

Butterworth, B. (1999). *The mathematical brain*. London, United Kingdom : MacMillan.

Butterworth, B. (2005). The developmental dyscalculia. Dans J. I. D. Campbell (dir.), *Handbook of mathematical cognition*, (p. 455–467). New-York, NY : Psychology Press.

Butterworth, B., Reeve, R., Reynolds, F. et Lloyd, D. (2008). Numerical thought with and without words: Evidence from indigenous Australian children. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 13 179–13 184. doi:10.1073/pnas.0806045105

Cappelletti, M. et Fias, W. (2016). Progress in brain research: Vol. 227. *The mathematical brain across the lifespan*. Amsterdam, Pays-Bas : Elsevier.

Carey, S. (2001). Cognitive foundations of arithmetic: Evolution and ontogenisis. *Mind & Language, 16*, 37–55. doi:10.1111/1468-0017.00155

Carey, S. (2004). Bootstrapping & the origin of concepts. *Doedalus, 131*(1), 59–68. doi:10.1162/001152604772746701

Chevrie-Muller, C. et Plaza, M. (2001). *Nouvelles Épreuves pour l'Examen du Langage*. Paris, France : Édition du Centre de Psychologie Appliquée.

Cognet, G. (2006). *Nouvelle échelle métrique de l'intelligence* (2e éd.). Paris, France : Édition du Centre de Psychologie Appliquée.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2e éd.). Hillsdale, NJ : L. Erlbaum Associates.

Connoly, A. J. (2008). *KeyMath™ 3 diagnostic assessment: Canadian edition*. San Antonio, TX : Pearson.

Dehaene, S. (2010). *La bosse des maths*. 15 ans après. Paris, France : Odile Jacob.

De Vos, T. (1992). *Tempo Test Rekenen*. Berkhout, Pays-bas : Nijmegen.

Donlan, C., Cowan, R., Newton, E. J. et Lloyd, D. (2007). The role of language in mathematical development: Evidence from children with specific language impairments. *Cognition, 103*, 23–33. doi:10.1016/j.cognition.2006.02.007

Duquesne, F. (2003). L'ECPN : des situations problèmes pour évaluer les principales fonctions du nombre. *Glossa, 83*, 4-18.

Durieux, N., Pasleau, F., Piazza, A., Donneau, A.-F., Vandenput, S. et Maillart, C. (2016). Information behaviour of French-speaking speech-language therapists in Belgium: Results of a questionnaire survey. *Health Information and Libraries Journal, 33*, 61–76. doi:10.1111/hir.12118

Durkin, K., Mok, P. L. H. et Conti-Ramsden, G. (2013). Severity of specific language impairment predicts delayed development in number skills. *Frontiers in Psychology, 4*(581), 1–10. doi:10.3389/fpsyg.2013.00581

Flessas, J. et Lussier, F. (2003). *Épreuve verbale d'aptitudes cognitives*. Paris, France : Édition du Centre de Psychologie Appliquée.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2e éd.). New York, NY : John Wiley.

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

Frank, M. C., Everett, D. L., Fedorenko, E. et Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition, 108*, 819–824. doi:10.1016/j.cognition.2008.04.007

Gaillard, F. (2000). *Numerical : test neurocognitif pour l'apprentissage du nombre et du calcul*. Lausanne, Suisse : Institut de psychologie Université de Lausanne.

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

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science, 306*, 496–499. doi:10.1126/science.1094492

Green, K. B. et Gallagher, P. A. (2014). Mathematics for young children: A review of the literature with implications for children with disabilities. *Başkent University Journal of Education, 1*(1), 81–92.

Gross-Tsur, V., Manor, O. et Shalev, R. S. (1996). Developmental dyscalculia: Prevalence and demographic features. *Developmental Medicine & Child Neurology, 38*, 25–33. doi:10.1111/j.1469-8749.1996.tb15029.x

Habib, M. (2014). *La Constellation des dys*. Paris, France : De Boeck-Solal.

Habib, M., Noël, M.-P., George-Poracchia, F. et Brun, V. (2011). *Calcul et dyscalculie : des modèles à la rééducation*. Paris, France : Masson.

Heremans, M. (2011). *MathEval*. Repéré à https://sites.google.com/site/testmatheval/

Ivanova, M. V. et Hallowell, B. (2013). A tutorial on aphasia test development in any language: Key substantive and psychometric considerations. *Aphasiology, 27*, 891–920. doi:10.1080/02687038.2013.805728

Kadosh, R. C. et Dowker, A. (2015). *The Oxford handbook of numerical cognition*. Oxford, United Kingdom : Oxford library of psychology.

Kaufmann, A. S. et Kaufmann, N. L. (2008). *Batterie pour l'examen psychologique de l'enfant* (2e éd.). Paris, France : Édition du Centre de Psychologie Appliquée.

Koppel, H. (1998). *Difficultés en mathématiques. Évaluation et rééducation*. Neuilly-Plaisance, France : Papyrus.

Lafay, A. et Helloin, M.-C. (2016). *Examath 8-15 : batterie informatisée d'examen des habiletés mathématiques*. Grenade, France : HappyNeuron.

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

Lafay, A., St-Pierre, M.-C. et Macoir, J. (2015). Validation franco-québécoise du Tempo Test Rekenen pour l'évaluation des habiletés mathématiques auprès d'enfants de 8-9 ans. *Glossa, 118*, 27–39.

Lafay, A., St-Pierre, M.-C. et Macoir, J. (2016). Performances moyennes des enfants franco-québécois de 8-9 ans au test mathématique Zareki-R. *Glossa, 119*, 41–54.

Leclercq, L. et Veys, E. (2014). Réflexions sur le choix de tests standardisés lors du diagnostic de dysphasie. *Approche Neuropsychologique des Apprentissages chez l'Enfant, 26*, 374–382.

Legeay, M. P., Morel, L. et Voye, M. (2009). *Mallette Bilan*. Trucy sur Yonne, France : Cogilud.

McCauley, R. J. (1989). Measurement as a dangerous activity. *Journal of Speech-Language Pathology and Audiology, 13*, 29–32.

McCauley, R. J. et Swisher, L. (1984). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders, 49*, 338–348. doi:10.1044/jshd.4904.338

Meljac, C. et Lemmel, G. (1999). *Construction et Utilisation du Nombre* (2e éd.). Paris, France : Édition du Centre de Psychologie Appliquée.

Ménissier, A. (2011). Analyser, comprendre et travailler les problèmes arithmétiques. Dans M. Habib, M.-P. Noël., F. George-Poracchia et V. Brun (dir.), *Calcul et dyscalculies. Des modèles à la rééducation* (p. 79-129). Issy-les-Moulineaux, France : Elsevier-Masson.

Ménissier, A. (2003). Les variations stratégiques chez l'enfant dans le calcul d'additions et de soustractions élémentaires. *Glossa, 83*, 20–33.

Métral, E. (2008). *Mallette B-LM cycle II*. Chavanod, France : Orthopratic.

Noël, M.-P. et Grégoire, J. (2015). *Test diagnostique des compétences de base en mathématiques pour les enfants du CE2 à la 5ème*. Paris, France : Édition du Centre de Psychologie Appliquée.

Noël, M.-P. et Rousselle, L. (2011). Developmental changes in the profiles of dyscalculia: An explanation based on a double exact-and-approximate number representation model. *Frontiers in Human Neuroscience, 5*(165), 1–4. doi:10.3389/fnhum.2011.00165

Nunnally, J. C. (1978). *Psychometric theory* (2e éd.). New York, NY : McGraw-Hill.

O'Connor, S. et Pettigrew, C. M. (2009). The barriers perceived to prevent the successful implementation of evidence–based practice by speech and language therapists. *International Journal of Language & Communication Disorders, 44*, 1018–1035. doi:10.1080/13682820802585967

Plante, E. et Vance, R. (1994). Selection of preschool language test: A data-based approach. *Language, Speech, and Hearing Services in School, 25*, 15–24. doi:10.1044/0161-1461.2501.15

Riquier, M. (1997). *Tests d'acquisitions scolaires mathématiques*. Paris, France : Édition du Centre de Psychologie Appliquée.

Roux, M.-O. (2014). Surdité et difficultés d'apprentissage en mathématiques, état des lieux et problématiques actuelles. *Bulletin de Psychologie, 4*, 295–307. doi:10.3917/bupsy.532.0295

Savigny, M. (2001). *Batterie d'épreuves pour l'école élémentaire*. Paris, France : Édition du Centre de Psychologie Appliquée.

Simonart, G. (1998a). *Échelle d'apprentissages scolaires primaires*. Braine-le-château, Belgique : Eurotests édition.

Simonart, G. (1998b). *Tests pédagogiques de premier cycle primaire*. Braine-le-château, Belgique : Eurotests édition.

Starkey, P. et Cooper, R. G. (1980). Perception of numbers by human infants. *Science, 210*, 1033–1035. doi:10.1126/science.7434014

Thibault, M.-P. et Helloin, M.-C. (2006). *Exalang 3-6 : batterie d'examen des fonctions langagières chez l'enfant de 3 à 6 ans*. Mont-Saint-Aignan, France : Orthomotus

Thibault, M.-P., Helloin, M.-C. et Lenfant, M. (2009). *Exalang 11-15 : batterie informatisée pour l'examen du langage oral, du langage écrit et des compétences transversales chez le collégien*. Mont-Saint-Aignan, France : Orthomotus.

Thibault, M.-P., Lenfant, M. et Helloin, M.-C. (2012). *Exalang 8-11 : bilan informatisé pour l'examen du langage et des compétences transversales chez l'enfant de 8 à 11 ans*. Mont-Saint-Aignan, France : Orthomotus.

Van Nieuwenhoven, C., Grégoire, J. et Noël, M.-P. (2001). *Test diagnostique des compétences de base en mathématiques*. Paris, France : Édition du Centre de Psychologie Appliquée.

Von Aster, M. G. et Shalev, R. S. (2007). Number development and developmental dyscalculia. *Developmental Medicine & Child Neurology, 49*, 868–873. doi:10.1111/j.1469-8749.2007.00868.x

Von Aster, M. (2006). *Batterie pour l'évaluation du traitement des nombres et du calcul chez l'enfant* (adapté par G. Dellatolas). Paris, France : Édition du Centre de Psychologie Appliquée.

Wechsler, D. (2005a). *Wechsler Individual Achievement Test* (2e éd.). London, Angleterre : The Psychological Corporation.

Wechsler, D. (2005b). *Échelle d'intelligence de Wechsler pour enfants et adolescents* (4e éd.). Paris, France : Édition du Centre de Psychologie Appliquée.

Wilson, A. J. et Dehaene, S. (2007). Number sense and developmental dyscalculia. Dans D. Coch, G. Dawson et K. W. Fischer (dir.). *Human behavior learning, and the developing brain: Atypical development* (p. 212–238). New-York, NY : Guilford Press.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature, 358*, 749–750. doi:10.1038/358749a0

Zipoli, R. P. et Kennedy, M. (2005). Evidence-based practice among speech-language pathologists: Attitudes, utilization, and barriers. *American Journal of Speech-Language Pathology, 14*, 208–220. doi:10.1044/1058-0360(2005/021)

Notes des auteurs

Les demandes au sujet de cet article doivent être adressées à Anne Lafay, 1455 Boulevard de Maisonneuve Ouest, FG Building, Office FG 6.209, Montréal, QC, H3G 1M8, Canada. Courriel : anne.lafay@concordia.ca

Déclaration d'intérêts

La première auteure (Anne Lafay) est co-auteure de l'outil d'évaluation intitulé *Examath 8-15 : batterie informatisée d'examen des habiletés mathématiques* qui a été inclus dans la recension des outils d'évaluation mathématique et qui a fait l'objet de l'analyse psychométrique. Elle déclare avoir des liens financiers avec la maison d'édition de l'outil d'évaluation (HappyNeuron).