JSLHR

Review Article

Using Nonword Repetition to Identify Developmental Language Disorder in Monolingual and Bilingual Children: A Systematic Review and Meta-Analysis

Salomé Schwob,[a] Laurane Eddé,[a] Laure Jacquin,[a] Mégane Leboulanger,[a] Margot Picard,[a] Patricia Ramos Oliveira,[a] and Katrin Skoruppa[a]

Purpose: A wealth of studies has assessed the diagnostic value of the nonword repetition task (NWRT) for the detection of developmental language disorder (DLD) in the clinical context of speech and language therapy, first in monolingual children and, more recently, in bilingual children. This review article reviews this literature systematically and conducts a meta-analysis on the discriminative power of this type of task in both populations.
Method: Three databases were used to select articles based on keyword combinations, which were then reviewed for relevance and methodological rigor based on internationally recognized checklists. From an initial pool of 488 studies, 46 studies were selected for inclusion in the systematic review, and 35 of these studies could be included in a meta-analysis.
Results: Most of the articles report significant discrimination between children with and without DLD in both monolingual and bilingual contexts, and the meta-analysis shows a large mean effect size. Three factors (age of the child, linguistic status, and language specificity of the task) yielded enough quantitative data for further exploration. Subgroups analysis shows variance in effect sizes, but none of the three factors, neither their interactions, were significant in a metaregression. We discuss how other, less explored factors (e.g., nature of the stimuli, scoring methods) could also contribute to differences in results. Sensitivity and specificity analyses reported in 33 studies confirmed that, despite possible effect size differences, the diagnostic accuracy of the NWRT is generally near thresholds considered to be discriminatory. It generally increases when it is combined with other tasks (e.g., parental questionnaire).
Conclusions: This review indicates that the NWRT is a promising diagnostic tool to identify children with DLD in monolingual and bilingual contexts with a large mean effect size. However, it seems necessary to choose the precise NWRT materials based on the children's language background and to complement the assessment sessions with other tools in order to ensure diagnosis and to obtain complete language profile of the child.
Supplemental Material: https://doi.org/10.23641/asha.15152370

The ability to adequately repeat nonwords relies on many processes such as auditory perception, phonological encoding and assembly, storage of phonological representations, motor planning, and the ability to articulate (Coady & Evans, 2008). Thus, it is unsurprising that many studies have shown that nonword repetition task (NWRT) is a good clinical indicator of developmental language disorders, which may affect these processes (see, for instance, the recent literature review by Coady & Evans, 2008). The accuracy with which phonologically plausible nonwords are repeated is a relatively reliable index of language level (Szewczyk et al., 2018) and a clinical marker of developmental language disorder (DLD; Boerma et al., 2015; Conti-Ramsden et al., 2001). Beyond general performance differences between children with and without DLD (Archibald, 2008), the presence and nature of errors in NWRT seem to reflect difficulties in short-term verbal memory and phonological accuracy (Ferré & Dos Santos, 2015; Li'el et al., 2019). Indeed, verbal short-term memory is often reported to be affected by language difficulties (Gray, 2004).

Compared to other assessments used in speech and language therapy (SLT), NWRT has the advantage of not

[a]Institut des sciences logopédiques, Pierre-à-Mazel 7, 2000 Neuchâtel, Université de Neuchâtel, Suisse
Correspondence to Salomé Schwob: salome.schwob@unine.ch

Editor-in-Chief: Stephen M. Camarata
Editor: Filip Smolik
Received September 18, 2020
Revision received February 11, 2021
Accepted May 9, 2021
https://doi.org/10.1044/2021_JSLHR-20-00552

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

directly involving the test subject's linguistic knowledge (e.g., lexicon and sentence structure). It is more independent from linguistic experience, since it assesses the ability to process new information (Archibald, 2008). Thus, this task can already be administered to very young children (Guiberson & Rodríguez, 2015) and has also been praised as a promising tool for children from multilingual backgrounds with very limited exposure to the test language (e.g., Boerma et al., 2015; Chiat, 2015).

In a global context where multiple languages coexist, and where the number of individuals speaking several languages continues to increase (Grosjean, 2015), the question of the correct language assessment for these individuals is a central concern. In addition to the many interindividual differences between children, disparities in language development in a context of bilingualism depend on further factors. For instance, both quantity and quality of exposure crucially impact the development of the bilingual child's language skills (Paradis, 2019), which can also be influenced by the migration path, social and motivational factors (Di Meo et al., 2014; Paradis et al., 2011). Given this multitude of aspects influencing the level of linguistic knowledge in bilingual children, establishing bilingual developmental norms is not straightforward (Pearson, 2013) and the diagnosis of DLD in bilingual children is often characterized as the greatest challenge for the present-day SLT clinic (Armon-Lotem, 2012). Indeed, it is often difficult for clinicians to differentiate between persistent language difficulties caused by DLD and transient difficulties due to insufficient exposure to one language (Camilleri & Law, 2007). The language difficulties of children falling into these two scenarios are strikingly similar: Typically developing (TD) children beginning to learn a new language often make morphosyntactic errors traditionally associated with DLD in monolingual children (Armon-Lotem, 2012), for instance, omissions of clitic objects in French (Tuller et al., 2018). Therefore, the identification of DLD in bilingual children requires the construction of a new set of tools, of which NWRT seems to be an important part (Boerma & Blom, 2017). However, it is important to note some controversies on this issue. Although NWRT scores are less affected by the child's prior language knowledge, some studies show that they are not completely free from language specificities. Indeed, a child is more likely to correctly repeat nonwords created on the basis of the phonological characteristics of their own language. Several studies (e.g., Boerma et al., 2015; Kohnert et al., 2006) have found that children's scores depend largely on their experience in the language in which the NWRT was created, leading some authors to advocate the construction of a quasi-universal NWRT (e.g., Chiat, 2015). Other authors (e.g., De Almeida et al., 2016; Thordardottir & Brandeker, 2013) found no influence of language experience on the task scores used.

Aims of the Article

In view of the numerous studies and the various controversies on the subject (see also Coady & Evans, 2008), the primary objective of this article is to provide a general overview of the current state of evidence regarding NWRT and DLD diagnosis, both in monolingual and bilingual children. Coady and Evans's (2008) literature review and Estes et al.'s (2007) meta-analysis already provided first overviews of the use of the task with monolingual children, but to our knowledge, this article is the first systematic review and meta-analysis that specifically focuses on bilingual children with DLD.

In particular, we will attempt to answer the following research questions:

1. Does the NWRT reveal significant differences in performance between children with TD versus DLD, in a monolingual versus bilingual context?
   a. What is the mean effect size of the studies?
   b. What are the sensitivity and specificity values of the NWRT found by the studies in the review?
2. What are the internal and external characteristics of children that may influence their NWRT scores and the discriminating power of the task?
   a. Does language status (monolingual vs. bilingual) influence the discriminating power of the task and the effect size?
   b. Does the language specificity of the task (language specific vs. quasi-universal) influence the effect size?
   c. Is the age of the children a moderator factor of effect size?
   d. Which other factors do studies report to influence performance on this task?
3. Which other tasks correlate with NWRT scores?
4. Which other tasks, in combination with NWRT, improve the discriminative power and thus facilitate the distinction between children with TD versus DLD?

In order to zoom in on the most relevant age range for DLD diagnosis, we decided to focus on studies using data from children up to the age of 8;11 (years;months). However, despite the fact the diagnosis of DLD is typically not made until around the fourth year of life (Bishop et al., 2017; Norbury et al., 2016), we decided not to use a lower age limit and to include studies that have focused on children at risk for DLD and late talkers, generally defined as producing less than 50 words and having no word combinations at 24 months of age (Desmarais et al., 2008; Pearson, 2013) in our qualitative part.

Method

Article Selection

A literature search was carried out using three databases: PsycINFO, Scopus, and ASHA (American Speech-Language-Hearing Association). The PsycINFO and Scopus databases were selected because of the large number of publications available, and ASHA, because of its specificity

to SLT. In order to conduct an exhaustive search of the existing literature, the following two combinations of English keywords were entered into the three databases: (a) nonword repetition, language impairment or language disorders, and assessment; and (b) nonword repetition, language impairment or language disorders, diagnosis. The same search was also conducted with the terms in French and in German, but no results were found in all three databases. For each database, and for all combinations of keywords, the bibliographic search was conducted by two researchers in order to ensure the reliability of the results. In case of disagreement, the results were reviewed jointly. The search was conducted between November 2019 and June 2020. In order to exclude articles that did not correspond to the interests of this review, criteria were used to mask the following publications: publications prior to 2000; publications whose language of publication was neither English, French, nor German; publications whose population consisted exclusively of children over the age of 8;11 (or, by extension, adolescents/adults).

The flow chart (see Figure 1) details the exact number of articles selected at each stage of the literature search. At this first stage of the selection process, 488 references were identified from the three databases and 16 from external sources. After removal of duplicates, a total of 226 studies were retained. Two of the authors of this study then reviewed the titles and abstracts of the articles to ensure their relevance. In case of disagreement, the studies were reviewed jointly. The following studies were then excluded: (a) systematic reviews and meta-analysis, (b) studies whose research design ultimately did not use NWRT, and (c) studies whose population consisted of children with disorders associated with DLD (e.g., deafness, attention deficit disorder) or of children who did not fit clinical diagnostic criteria for DLD or late talkers (e.g., "at-risk" children with scores below the 50th percentile, which was deemed as too broad a criterion). A total of 74 articles were then included. A critical evaluation process was then conducted using the INESS recommendations (Martin et al., 2013) to identify quality studies, as well as the PRISMA (Moher et al., 2009) checklist. Then, the methodological part and the section concerning the results of the studies were scrutinized. A robust and sufficiently described methodological framework was a prerequisite for our systematic review. We therefore verified that the studies (a) had a sample size of more than 10 children; that (b) the groups of participants were described, defined, and matched; that (c) the NWRT used was described (e.g., stimuli); that (d) the data analysis methodology was presented; and finally that (e) inferential statistical analyses had been performed and described. Each article was evaluated on the basis of these criteria by two authors, so as to ensure reliability among judges in the final selection of articles actually included in this review.

A total of 46 articles were thus selected and reviewed in their entirety, of which 35 were used to conduct the meta-analysis. The remaining articles were excluded from the meta-analysis because they either did not contain the information needed to calculate effect sizes, or the clinical population was not clearly identified as DLD, but at-risk children or late talkers.

Analysis

To conduct the meta-analysis, we opted for the Comprehensive Meta-Analysis software (Bornstein et al., 2013). We calculated an effect size (Cohen's *d*) for each sample by subtracting each DLD group's mean nonword repetition score from its TD group's mean score and dividing the difference by the mean within-group (pooled) standard deviation. With reference to Cohen (1988), the effect size is considered small when $d = .20$, medium when $d = .50$, and large when $d = .80$. All effect sizes have been corrected for bias in the estimation of the sample size. We therefore applied the formula of Hedges and Becker (1986) for the correction factor. A homogeneity test of effects was used to determine whether all effects were from the same population. Since our data are heterogeneous (the I^2 index showed a large heterogeneity of 79.4%), the overall effect size was calculated using random effects models. In addition, subgroups analysis (language specificity of the task and linguistic status) and a metaregression with moderators (age of children, language specificity of the task, and linguistic status) was applied to determine whether study characteristics explain the variation in effect size between groups. To test for publication bias, we created a funnel plot and we conducted an Egger's regression test.

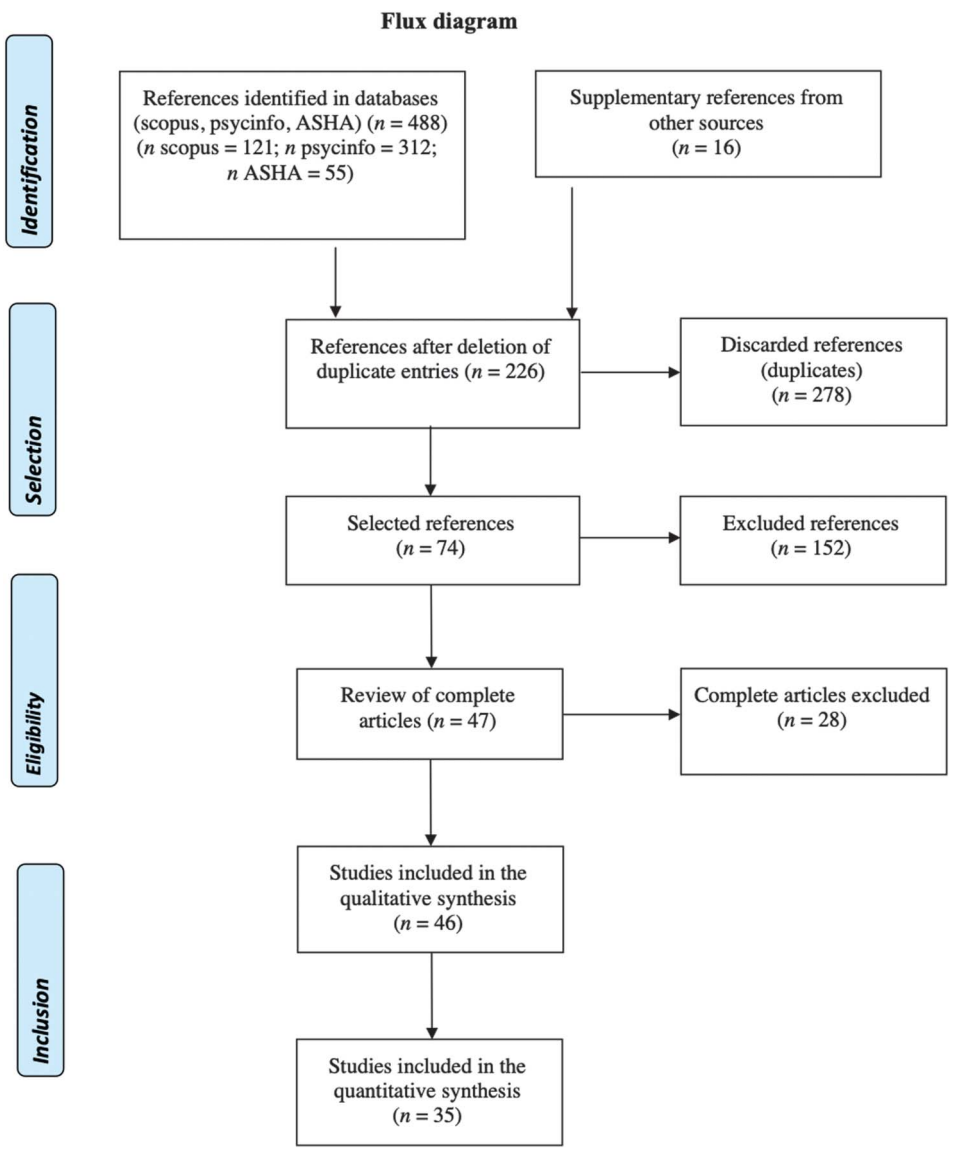
Results and Discussion

We will first report some general characteristics of the studies included in our systematic review and meta-analysis. We will then present our quantitative meta-analysis on group differences between children with TD and DLD and perform a mean effect size calculation and check for publication bias. We will then conduct subgroups analysis and metaregression to understand what contributes to the variance in effect sizes. We will continue by presenting other possible influencing factors qualitatively. Finally, we will examine the discriminative values (sensitivity and specificity) of this task in monolingual and bilingual assessment contexts, as well as correlations and combinations with other language tasks.

General Characteristics of the Included Studies

Some basic methodological features (test languages and stimuli, age ranges, and clinical status of the participants) of each study are summarized in Table 1 below. It is worth noting that the majority (32 of the 46 studies) involved an exclusively monolingual population, five involved an exclusively bilingual population, and nine involved a comparison between a monolingual and a bilingual population. Table 1 also shows the different language profiles of the children participating in the studies identified. We note that the most studied language is English and that other languages are examined in several papers (e.g., French and Dutch in particular). Other languages and profiles (e.g., bilingualism)

Figure 1. PRISMA flux diagram illustrating the literature research steps and quantitative results.



are also studied (e.g., Vietnamese monolinguals in the study by Pham & Ebert, 2020).

Quantitative Meta-Analysis

First, we conducted an overall meta-analysis of effect sizes. A forest plot summarizing these results is depicted in Figure 2, which shows the estimated effect sizes and the 95% confidence interval (CI) for each study. All individual effect sizes can be found in the Supplemental Material S1. The mean effect size across the studies is large $g = 1.57$ (95% CI [1.37, 1.72], $p < .001$), indicating that the children with DLD consistently performed much poorer than children with TD.

We then conducted subgroups analysis to observe whether effect sizes vary with the linguistic status of the children (monolingual or bilingual or with the language specificity of the task (language-specific or quasi-universal). Table 2 shows slight variations in mean effect size across subgroups. Specifically, the effect size is larger for the monolingual subgroups (1.61) than for the bilingual subgroups (1.36). Similarly, the use of the quasi-universal task provides a higher mean effect size (1.73) than the language-specific task (1.52).

To examine the statistical significance of these variations, we performed a metaregression. Including three moderators often reported in the literature as influencing factor (language status, age, and language specificity of the task) for which sufficient quantitative data were available (> 10

Table 1. Clinical and methodological characteristics and general aims of the studies on nonword repetition task (NWRT) included in this review (n = 46).

Study	Sample size	Age	Linguistic profile	NWRT used
Weismer et al. (2000)	n = 581	7;1–8;11 (years;months)	English monolinguals	Dollaghan & Campbell (1998)
Rodekohr & Haynes (2001)	n = 40	7;0–7;3	English monolinguals	Dollaghan & Campbell (1998)
Conti-Ramsden (2003)	n = 64	4;4–5;10	English monolinguals	Gathercole & Baddeley (1996)
Conti-Ramsden & Hesketh (2003)	n = 64	DLD: 4;4–5;10/LT: 2;4–3;7	English monolinguals	Gathercole & Baddeley (1996)
Gray (2003)	n = 44	4;0–5;11	English monolinguals	Gathercole et al. (1994)
Gray (2004)	n = 40	4;0–5;11	English monolinguals	Dollaghan & Campbell (1998)
Horohov & Oetting (2004)	n = 54	5–7 years	English monolinguals	Montgomery (1995)
Washington & Craig (2004)	n = 81	3;10–4;1	English monolinguals	Dollaghan & Campbell (1998)
Thal et al. (2005)	n = 64	4;0–4;6	English monolinguals	Dollaghan & Campbell (1998)
Bortolini et al. (2006)	n = 22	3;7–5;6	Italian monolinguals	own task
Gray (2006)	n = 106	3–6 years	English monolinguals	Dollaghan & Campbell (1998)
Oetting & Cleveland (2006)	n = 83	4;0–6;0	English monolinguals	Dollaghan & Campbell (1998)
Chiat & Roy (2007)	n = 483	2;0–4;0	English monolinguals	"PRT" (Roy & Chiat, 2004)
Oetting et al. (2008)	n = 95	4;0–6;0	English monolinguals	Dollaghan & Campbell (1998)
Archibald & Joanisse (2009)	n = 400	5;3–9;4	English monolinguals	Dollaghan & Campbell (1998)
Jones et al. (2010)	n = 36	5;7–6;7	English monolinguals	Gathercole & Baddeley (1996) and own task
Deevy et al. (2010)	n = 76	4;1–5;11	English monolinguals	Dollaghan & Campbell (1998)
Gutiérrez-Clellen & Simon-Cereijido (2010)	n = 144	3;0–7;0	English–Spanish bilinguals	Dollaghan & Campbell (1998) and own task
Thordardottir et al. (2011)	n = 92	4;1–5;9	French monolinguals	Courcy (2000)
Petrucelli et al. (2012)	n = 95	5;0–5;8	English monolinguals	Gathercole & Baddeley (1996)
Kapalková et al. (2013)	n = 32	3–5 years	Slovak monolinguals	own task
McKean et al. (2013)	n = 50	3;0–5;11	English monolinguals	own task
Guiberson & Rodríguez (2013)	n = 44	3;0–5;10	Spanish monolinguals	Ebert et al. (2008)
Dispaldro et al. (2013)	n = 34	3;11–5;8	Italian monolinguals	Dispaldro et al. (2009, 2011)
Paradis et al. (2013)	n = 178	4;10–8;7	bilinguals English–other languages	"CTOPP" (Wagner et al., 1999)
Stokes et al. (2006)	n = 44	DLD: 4;2–7;7/same-age TD: 4;1–6;9/younger TD: 2;11–3;6	Cantonese monolinguals	own task
Thordardottir & Brandeker (2013)	n = 140	4;4–5;9	French monolinguals, bilinguals French–other languages	Courcy (2000), Thordardottir et al. (2011), and Gathercole et al., (1994)
Tuller et al. (2013)	n = 29	5–7 years	French–Arabian bilinguals	French LITMUS
Topbaş et al. (2014)	n = 60	3;6–8;3	Turkish monolinguals	own task
Guiberson & Rodríguez (2015)	n = 65	2;0–2;11	Spanish monolinguals	own task
Boerma et al. (2015)	n = 120	4;6–7;3	Dutch monolinguals, bilinguals Dutch–other languages	Rispens & Baker (2012), Dutch LITMUS
Ferré & Dos Santos (2015)	n = 67	5;2–8;5	French monolinguals, French–Arabian/English bilinguals	French LITMUS

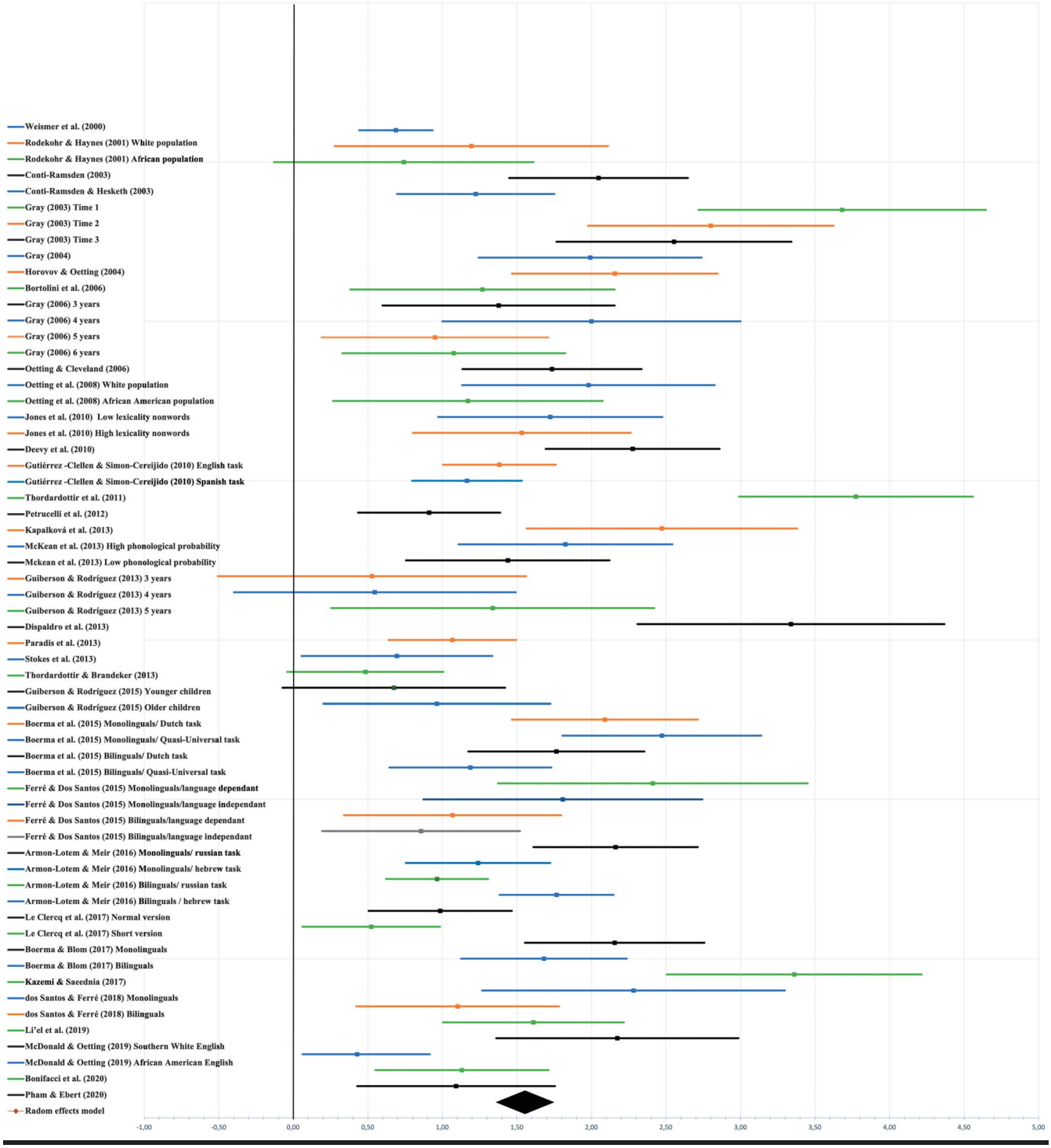
(*table continues*)

Table 1. *(Continued).*

Study	Sample size	Age	Linguistic profile	NWRT used
De Almeida et al. (2016)	*n* = 136	5;4–8;11	French monolinguals, French Portuguese/Arabian/Turkish bilinguals	French LITMUS and "Palpa-P" (Castro et al., 2007)
Armon-Lotem & Meir (2016)	*n* = 231	5;5–6;8	Hebrew and Russian monolinguals, Hebrew–Russian bilinguals	Shortened version of Armon-Lotem & Chiat (2012)
De Almeida et al. (2017)	*n* = 137	5;4–8;11	French monolinguals, French Portuguese/Arabian/Turkish bilinguals	French LITMUS and "Palpa-P" (Castro et al., 2007)
Le Clercq et al. (2017)	*n* = 88	5;4–8;10	Dutch monolinguals	Normal version and shortened version of Rispens & Baker (2012)
Boerma & Blom (2017)	*n* = 132	5;0–6;0	Dutch monolinguals, Dutch–Berber/Moroccan/Turkish bilinguals	Quasi-universal LITMUS (Chiat, 2015)
Hodges et al. (2017)	*n* = 52	2;1–2;11	English monolinguals	"TENR" (Stokes & Klee, 2009) and "MITT" (Hodges et al., 2017)
Kazemi & Saeednia (2017)	*n* = 51	3;7–5;3	Persian monolinguals	Afshar et al. (2013)
Marini et al. (2017)	*n* = 293	2;2–2;11	Italian monolinguals	"BVL" (Marini et al., 2015)
Tuller et al. (2018)	*n* = 227	5;6–8;11	French and German monolinguals, French/German–Arabian/Portuguese/Turkish bilinguals	French and German LITMUS
Dos Santos & Ferré (2018)	*n* = 67	5;2–8;5	French monolinguals, French/Arabian/English bilinguals	French LITMUS
Li'el et al. (2019)	*n* = 61	5;0–6;0	bilinguals English–other languages	"CTOPP" (Wagner et al., 1999)
McDonald & Oetting (2019)	*n* = 106	4;11–6;2	English monolinguals	Dollaghan & Campbell (1998)
Pham & Ebert (2020)	*n* = 104	5;2–6;2	Vietnamese monolinguals	Pham et al. (2018)
Bonifacci et al. (2020)	*n* = 55	6;4–7;4	bilinguals Italian–other languages	"PVN 5–11" (Bisiacchi et al., 2005)

Note. *N* = 46, listed by order of publication date. DLD = developmental language disorders; LT = late talkers; TD = typic development.

Figure 2. Forest plot. *N* =35.



studies). However, as can be concluded from the results in Table 3, the three moderators do not significantly influence effect size as assumed, neither do their interactions. Thus, although there may be small trends with respect to language status and the language specificity of the stimuli in the subgroup analysis, these should be interpreted with caution as they are not statistically confirmed, which may of course be due to the small and unequal sample size of the studies.

It may seem surprising at first that age does not have a significant impact, but it is worth noting that, here, we do

Table 2. Mean effect size and subgroups analysis.

Task/group	Hedges's *g*	Standard error	Variance	Lower limit	Upper limit	*Z* value	*p* value
Language-specific	1.52	0.09	0.01	1.33	1.71	15.51	$p < .001^{***}$
Quasi-universal	1.73	0.18	0.03	1.36	2.10	9.15	$p < .001^{***}$
Bilinguals	1.36	0.10	0.01	1.16	1.57	12.95	$p < .001^{***}$
Monolinguals	1.61	0.11	0.01	1.37	1.84	13.64	$p < .001^{***}$
Overall	1.57	0.08	0.00	1.40	1.73	18.19	$p < .001^{***}$

Note. Some studies were counted multiple times because they detailed their results according to several tasks or groups of children (e.g., age, language status). See details in Supplemental Material S1. For language specificity of the task, $N = 59$ (language-specific task, $n = 53$; quasi-universal task, $n = 6$). One study (looking at two groups of children, monolingual and bilingual) could not be clearly identified as having used a language-specific or quasi-universal task, that is, Dos Santos and Ferré (2018), those who used tasks that included language-dependent and language-independent items. For linguistic status of the children, $N = 61$ (bilingual children, $n = 13$, monolingual children, $n = 48$).

$^{***}p < .001$.

not examine the influence of age on the children's scores per se, but on performance differences between children with DLD and TD of the same age. Furthermore, the NWRT stimuli, often taken from test batteries, are generally chosen to fit with the age range of the children studied, and many tasks contain (sometimes age-specific) abortion criteria in order to make them maximally discriminative for all ages. Thus, the discriminative power of NWRT does not seem to vary across the age range examined (2;00–8;11), and we can therefore confirm that NWRT is a reliable indicator of DLD across development in young children.

Finally, to test for publication bias, we created a funnel plot (in the Supplemental Material S2). We noted missing data mostly on the left side of the figure, indicating that studies with small effect sizes or with nonsignificant data are missing. The Egger's regression was statistically significant ($t = 4.13$, $p < .001$), indicating a publication bias. We can think of several explanations for this fact: First, NWRT is a well-established and robust clinical task used in many test batteries that are designed to maximize the effect of DLD; thus, studies not finding such an effect at least in monolingual children would be very hard to publish. Second, despite our efforts to contact authors in the field and to participate in relevant conferences, we were not able to find and include unpublished gray literature on the topic. However, it should be noted that the main point of our review was not to establish whether NWRT can be used to distinguish TD and DLD children, which has been done before, but to compare how it fares for monolingual and bilingual children as well as analyze the influence of other factors.

Qualitative Analysis

In the following, we will qualitatively assess other factors that may influence NWR scores, but that could not be included in the meta-analysis because of lack of data. These factors are mentioned by few studies, and often the details (means and standard deviations) relating to them do not appear in the articles. Again, we focus specifically on the influence of these factors on the discriminative power of the task, rather than scores per se.

Stimuli Characteristics

Referring to Table 1 (presented earlier), we note that different tasks were used. Several studies used language-specific tasks like Dolloghan and Campbell's (1998) task or the NWRT developed by Gathercole et al. (1994) and Gathercole and Baddeley (1996). Some studies created their own NWRT (e.g., Bortolini et al., 2006) or drew on other

Table 3. Moderators analysis.

Covariate	Coefficient	Standard error	95% Lower	95% Upper	*Z* value	*p* value
Intercept	1.94	0.50	0.95	2.93	3.85	$p < .001^{***}$
Linguistic status	0.14	0.23	–0.31	0.61	0.62	$p = .53$
Task type	0.08	0.42	–0.75	0.92	0.21	$p = .83$
Linguistic Status × Task Type	0.44	0.59	–0.71	1.60	0.75	$p = .45$
Age	–0.00	0.00	–0.02	0.00	–1.24	$p = .21$

Note. Some studies were counted multiple times because they detailed their results according to several tasks or groups of children (e.g., age, language status). See details in Supplemental Material S1. For language specificity of the task, $N = 59$ (language-specific task, $n = 53$; quasi-universal task, $n = 6$). One study (looking at two groups of children, monolingual and bilingual) could not be clearly identified as having used a language-specific or quasi-universal task, that is, Dos Santos and Ferré (2018), those who used tasks that included language-dependent and language-independent items. For linguistic status of the children, $N = 61$ (bilingual children, $n = 13$, monolingual children, $n = 48$). For age of children, $N = 61$. Test of the model: $Q = 3.95$, $df = 4$, $p = .41$.

$^{***}p < .001$.

tasks previously designed or derived from test batteries (for details, refer to Table 1). A few studies have also used Chiat's (2015) quasi-universal task and combinations of language-dependent and language-independent items (e.g., Dos Santos & Ferré, 2018).

Different characteristics of the NWRT stimuli have been shown to influence children's scores. Turns out, 46 studies have focused on syllable structure (Dos Santos & Ferré, 2018; Ferré & Dos Santos, 2015; Hodges et al., 2017; Jones et al., 2010; Marini et al., 2017). This varies widely among the different studies. Some studies used a single syllabic structure (CV or CVC), while others combined different syllabic structures within nonwords (e.g., CV + CVC, CV + CCV). Dos Santos and Ferré (2018) found that, in the presence of complex consonant groups, the performance of DLD children decreases in both monolinguals and bilinguals. Children with DLD have more difficulties than children with TD when the structures are complex (Jones et al., 2010). Dos Santos and Ferré also specified that the complexity of items can be influenced by whether or not they resemble the child's language. Therefore, on the basis of the LITMUS task, these authors create stimuli that are dependent on and independent of the language spoken by the child. The results show that with the exception of monolingual children with DLD, none of the groups had a significant difference between the two types of stimuli of the test. Other authors such as Hodges et al. (2017) have focused on the acquisition of phonemes in relation to syllabic structure. They found that the effect of complexity appears only when the initial consonants are those acquired early and not late. This is valid for children with DLD and with TD, even if for the latter the effect is less pronounced.

Nonword length is another factor that is often considered important. In the studies reviewed here, it varies between one and five syllables. Some authors have chosen not to go beyond four syllables, given the age of the children in their sample (e.g., Jones et al., 2010). Of the 46 studies in our literature review, 18 found an effect of the length of nonword (Boerma et al., 2015; Bortolini et al., 2006; Chiat & Roy, 2007; Dispaldro et al., 2013; Guiberson & Rodríguez, 2013; Guiberson & Rodríguez, 2015; Horohov & Oetting, 2004; Jones et al., 2010; Marini et al., 2017; McDonald & Oetting, 2019; McKean et al., 2013; Oetting & Cleveland, 2006; Petruccelli et al., 2012; Rodekohr & Hayne, 2001; Thal et al., 2005; Thordardottir & Brandeker, 2013; Topbaş et al., 2014; Weismer et al., 2000). In general, both children with TD and children with DLD have more difficulty with NWRT as stimulus length increases (Bortolini et al., 2006; Chiat & Roy, 2007; Dispaldro et al., 2013; Guiberson & Rodríguez, 2013; Horohov & Oetting, 2004; Jones et al., 2010; Marini et al., 2017; Oetting & Cleveland, 2006; Petruccelli et al., 2012; Rodekohr & Haynes, 2001; Thal et al., 2005; Topbaş et al., 2014). Boerma et al. (2015) noted that monolingual and bilingual children with DLD perform less well than those with TD on all syllable lengths. Some authors (Dispaldro et al., 2013; McKean et al., 2013; Petruccelli et al., 2012) claimed that there is a significant difference between children with DLD and TD only when the length of nonwords increases, reflecting the limited short-term verbal memory capacity of children with DLD. One study found that children with DLD were significantly less accurate than children with TD on nonwords with two and three syllables of simple structures (Jones et al., 2010). The study by Petruccelli et al. (2012) identified that late-talking children's scores of fall between the scores of children with TD and children with DLD. Children with DLD and late-talking children have more difficulties than children with TD, especially on three- and four-syllable nonwords (Weismer et al., 2000). Finally, some authors claim that, for preschool children, it is not be relevant to use four- and five-syllable nonwords in the NWRT because of floor effects (Guiberson & Rodríguez, 2015).

Furthermore, five of the 46 studies examined possible effects of the wordlikeness of the nonwords used (Chiat & Roy, 2007; Gray, 2003; Hodges et al., 2017; Jones et al., 2010; McKean et al., 2013). McKean et al. (2013) found a wordlikeness effect for children with TD and with DLD. Jones et al. (2010) confirmed these results, although the effect is much stronger for children with DLD than for children with TD. Hodges et al. (2017) observed this, but with late-talking children instead of children with DLD. Gray (2003) reflected that the wordlikeness effect may signal a greater influence of prior language knowledge. Finally, lexical status seems to influence performance as well, as shown by Chiat and Roy (2007) in their study comparing repetition of words and nonwords. Chiat and Roy (2007) also looked at prosody. They noted that accented syllables are almost never omitted. Children with DLD omitted accented syllables only in long nonwords (bisyllables and trisyllables). However, children with DLD omit nonaccented syllables that follow accented syllables more easily, but this is not the case for children with TD.

Scoring Method

Different scoring methods were used, depending on the tasks and on authors' preferences (e.g., Gray, 2003, 2006). Two main scoring methods were identified in the review studies. First, a relatively straightforward count of percentage of items correct (PIC) was used by 21 studies (Archibald & Joanisse, 2009; Armon-Lotem & Meir, 2016; Boerma & Blom, 2017; Bonifacci et al., 2020; Chiat & Roy, 2007; Conti-Ramsden, 2003; Conti-Ramsden & Hesketh, 2003; De Almeida et al., 2016; De Almeida et al., 2017; Gray, 2003; Gray, 2006; Guiberson & Rodríguez, 2015; Horohov & Oetting, 2004; Jones et al., 2010; Li'el et al., 2019; Marini et al., 2017; Paradis et al., 2013; Petruccelli et al., 2012; Thal et al., 2005; Tuller et al., 2013, 2018).

Second, 12 studies in the review count the percentage of phonemes correct (PPC; Deevy et al., 2010; Gray, 2004; Gutiérrez-Clellen & Simon-Cereijido, 2010; Hodges et al., 2017; McDonald & Oetting, 2019; Oetting & Cleveland, 2006; Oetting et al., 2008; Rodekohr & Haynes, 2001; Thordardottir & Brandeker, 2013; Thordardottir et al., 2011; Washington & Craig, 2004; Weismer et al., 2000). The precise way in which the PPC is calculated can influence the method's discriminating performance. For example, Deevy et al. (2010) compared two PPC methods (e.g., by

treating phonemes out of inventory either as errors or as unscorable) and observed differences in their percentages of sensitivity and specificity.

Eight out of 46 studies combined different scoring methods: For example, Topbaş et al. (2014) combined PIC and PPC and added the percentage of vowels correct. McKean et al. (2013) calculated the number of incorrect phonemes in each nonword, and Dos Santos and Ferré (2018) also looked at the number of errors present in each item. These three studies then linked this score to the PIC or PPC. A final study conducted in Cantonese (Stokes et al., 2006) attempted to assign a point for each correct nucleus and consonant. This shows that, depending on the language, in this case Cantonese, the scoring method can differ for linguistic reasons. Bortolini et al. (2006) combined the PIC and the number of errors per segment. Finally, Kapalková et al. (2013) and Pham and Ebert (2020) combined several measures (e.g., between three and five depending on the study). They observed that all the measures used made it possible to dissociate children with DLD from those with TD. A few studies (Boerma et al., 2015; Deevy et al., 2010; Dispaldro et al., 2013; Guiberson & Rodríguez, 2013; Le Clercq et al., 2017; Pham & Ebert, 2020) used several scoring methods (including PIC and PPC) to compare them. The authors found that both methods discriminated between children with TD and DLD. However, the PIC had better sensitivity (e.g., 97% in Boerma et al., 2015; 71% in Guiberson & Rodríguez, 2013; and 100% in Dispaldro et al., 2013) in contrast to the PPC (63%, 48%, and 94%, respectively). Furthermore, Boerma et al. (2015) found that PIC was more accurate for monolingual children (in the quasi-universal NWRT), whereas no difference was found between PPC and PIC for bilingual children. Le Clercq et al. (2017) observed similar results for the effectiveness of PIC and PPC. Finally, PIC scoring is faster and therefore seems interesting for the clinic (Dispaldro et al., 2013; Pham & Ebert, 2020).

Error Analysis

As noted above, beyond a generally lower performance on NWRT, children with DLD may also make different errors than children with TD. Kapalková et al. (2013) found more phonological errors in Slovak children with DLD, including some atypical errors such as nasal and alveolar assimilations. In a study in which children were considered at risk for DLD (children were referred to speech language therapy because they were suspected of having a DLD), older at-risk children scored similarly to younger children with TD, omitting syllables more often than children with TD of their age (Chiat & Roy, 2007). However, only few studies make use of a complete error analysis, which is very time consuming, and none of them analyzed its potential to increase overall diagnostic accuracy.

Sociocultural Characteristics

Two studies (Chiat & Roy, 2007; Washington & Craig, 2004) found no effect of gender and socioeconomic status on children's performance in the NWRT. Study by Boerma et al. (2015) also found no effect of socioeconomic status on NWRT scores. In addition, Chiat and Roy (2007) found that there was no effect of day care type and languages spoken at home on scores.

Discriminative Value of the NWRT

Going beyond group differences, we will now turn to a clinically highly relevant parameter, the discriminative power of the NWRT to detect DLD in individuals among monolingual and bilingual children. We note that 33 out of 46 studies, summed up in the Supplemental Material S3, report the sensitivity and specificity of the NWRT. We note large methodological differences in the pathology thresholds adopted (sometimes expressed as scores, percentile, standard deviation) and, in the tasks, scoring methods and inclusion criteria of DLD used. The discriminative value of the NWRT also varies largely, with sensitivity ranging from 40% to 100% and specificity from 48% to 100%. We note that the majority of studies consider NWRT to be accurate since they have sensitivity and specificity percentages between 80% and 100% (e.g., Plante & Vance, 1994). However, 14 studies obtained at least one result considered to be diagnostically inaccurate. Seven of these studies involved a monolingual population. We note that some authors (e.g., Conti-Ramsden, 2003) have attempted to vary the threshold score (based on receiver operating characteristic curves; Dunn, 2014) to be reached but the percentages did not improve. It should be noted that the study by Thal et al. (2005) carried out on a monolingual population obtained a low sensitivity percentage (40%) for the NWRT alone, but this result improves significantly when the NWRT is combined with other tools. This point will be discussed more fully in later sections. Among the seven studies reporting poor discriminant results in a bilingual population, the majority used language-specific tasks (e.g., Dos Santos & Ferré, 2018; Gutiérrez-Clellen & Simon-Cereijido, 2010), sometimes influencing the subject's ability to repeat according to their experience in the language of the task (Boerma et al., 2015; Kohnert et al., 2006). The quasi-universal LITMUS task (Chiat, 2015) seems promising because it provides results that are still considered accurate in bilinguals. With this bilingual population, the deviation thresholds sometimes have to be modified in order to find the optimal discriminating values for children with and without DLD (e.g., Thordardottir, 2015). Indeed, Armon-Lotem and Meir (2016) also noted that, depending on the language of the tasks used, sensitivity and specificity scores can sometimes improve (e.g., the case for Hebrew but not for Russian). Finally, three studies (Deevy et al., 2010; Guiberson & Rodríguez, 2013; Pham & Ebert, 2020) noted differences in sensitivity and specificity depending on the scoring method. It is therefore not possible to conclude whether it is better to use an NWRT in the language of the society and/or family. Indeed, the skills of the bilingual child differ according to the time and tasks proposed (Gutiérrez-Clellen & Simon-Cereijido, 2010). Moreover, the latter can sometimes be designed with

various methodological properties (Dos Santos & Ferré, 2018). In order to obtain the best diagnostic accuracy, it therefore seems appropriate to carry out NWRTs in both of the child's languages or to use a quasi-universal task (Tuller et al., 2018).

NWRT and Other Assessment Tools

Correlations Between NWRT and Other Measures

Some authors reported correlations between NWRT scores and other verbal and nonverbal measures (summary presented in the Supplemental Material S4). For example, some studies observed a significant correlation between NWRT scores and vocabulary measures in English, Spanish, Italian, and French (e.g., Thordardottir & Brandeker, 2013). However, two further studies, Stokes et al. (2006) and Thordardottir et al. (2011), found no correlation between vocabulary measures and NWRT scores, neither for children with TD nor with DLD who speak Cantonese and French. Thordardottir and Brandeker (2013) and De Almeida et al. (2017) also found significant correlations between NWRT scores and other measures from standardized French and English language tests tapping into morphosyntax (sentence repetition) and phonology (word repetition). In addition to the aforementioned correlation with vocabulary, Horohov and Oetting (2004) reported significant correlations between NWRT scores and standardized measures of English syntax (syntax comprehension). Finally, Paradis et al. (2013) observed correlations between NWRT and tasks testing inflectional morphology in English for the group of children with TD. These authors, however, observed no correlation between these standardized test measures and NWRT scores for the group of children with DLD.

A number of studies put NWRT in relation with other tasks requiring repetition. Dispaldro et al. (2013) noted a correlation between the ability to repeat nonwords and the ability to repeat words in Italian. De Almeida et al. (2016, 2017) and Thordardottir and Brandeker (2013) observed correlations between children's scores on NWRT and a sentence repetition task in bilinguals with TD and DLD in French, Portuguese, and English. Conversely, the study by Stokes et al. (2006), focusing on Cantonese, found no correlation between NWRT and sentence repetition. Petruccelli et al. (2012), on the other hand, found no significant correlation between NWRT scores and scores on number and word recall in English. Gray (2003) noted correlations between the NWRT task and a number repetition task (assessed at different points in time) and other standardized language assessments in English. Other research has investigated links between NWRT and nonverbal tasks, with conflicting results. Oetting et al. (2006, 2008) noted a correlation between NWRT scores and a nonverbal reasoning test, whereas Boerma et al. (2015) found no correlation between nonverbal IQ and NWRT.

Several authors are also interested in possible correlations with parental observations. Tuller et al. (2018) found that high NWRT scores were most strongly correlated with positive early language development, as established through analysis of a parental questionnaire. Similarly, De Almeida et al. (2016, 2017) noted a correlation between NWRT and a language impairment risk index calculated through a parental questionnaire.

To sum up, NWRT seems to be generally correlated to other assessment measures, in particular those concerning oral language, such as vocabulary, morphosyntax, and early developmental indexes. However, the data are heterogeneous and, to date, no firm generalizations can be made regarding the link between NWRT and other verbal and nonverbal measures.

Combining NWRT With Other Measures

Nine out of 46 studies have addressed the value of combining NWRT with other empirical tools to increase diagnostic accuracy for DLD (see in the Supplemental Material S5). For example, NWRT can be combined with a sentence repetition task (Armon-Lotem & Meir, 2016), a comprehension test (Oetting & Cleveland, 2006), or a receptive vocabulary test (Thordardottir et al., 2011[1]). Boerma and Blom (2017) also obtained an increase in discriminant value by associating the NWRT with a narrative task, but only in the group of bilingual children. Bortolini et al. (2006), in their study of Italian, found that sensitivity and specificity improve when NWRT is combined with morphosyntactic tasks, such as third-person plural inflection and a clitic object task. Four studies (Boerma & Blom, 2017; Bonifacci et al., 2020; Li'el et al., 2019; Paradis et al., 2013) showed that the combination of NWRT with a parental questionnaire reporting early language milestones and parental concern can improve the accuracy of diagnosis in different groups of children (monolinguals–bilinguals/children with TD or DLD). Thus, combining NWRT with other tasks can be an effective means to increase diagnostic accuracy, especially in situations where the discriminative value NWRT is not sufficient on its own.

General Discussion and Conclusions

The aim of this work was to report and synthesize recent results on the use of NWRT in monolingual and bilingual SLT assessments. The various studies show that, in general, children with DLD (or at risk of developing DLD) score significantly lower than children with TD, which can be related, among other things, to impairment of their short-term verbal memory. Beyond these memory deficits, children with DLD often have poorer phonological abilities and have more difficulty developing fine and accurate phonological representations (McKean et al., 2013). We also note that children with DLD are less accurate in repetition and make more errors (Jones et al., 2010). The phonological errors of

[1]The study by Thordardottir et al. (2011) is included in Supplemental Material S5, but the new sensitivity and specificity scores with combined tasks are not described in the reference article.

these children are sometimes atypical, such as nasal and alveolar assimilations (Kapalková et al., 2013). However, given the number of NWR tasks used in the studies and the associated methodological differences, it is difficult to determine what exactly each type of NWRT measures (verbal short-term memory vs. phonological accuracy). We note that children with clearly diagnosed DLD show much more difficulties with all NWRT types than their peers without DLD. The large effect size found in our meta-analysis confirms that the NWRT appears to be a robust and recommendable tool for detecting DLD in monolingual and bilingual children.

In order to explore why our meta-analysis data were heterogeneous, we examined the influence of three factors known to influence NWRT performance, linguistic status, language specificity of the stimuli, and age. Indeed, subgroups analysis showed that effect sizes appear to vary according to the linguistic status of the children (with larger effects for monolinguals) and the language specificity of the task (with larger effects for quasi-universal tasks). However, these factors were not significant moderators in a metaregression, which may be due to the small and unequal number of studies included. In addition, with respect to age, discriminative power does not seem to vary by age in children up to age 8;11, who represent the main target group for DLD assessment. However, it is possible that an inclusion of older children (for instance, in the context of reading difficulties) would yield an influence of the age factor.

Other child-internal factors were analyzed qualitatively such as sociocultural, socioeconomic, gender, and day care attendance, and they do not appear to have an impact on children's scores (e.g., Chiat & Roy, 2007). It appears then that the NWRT is relatively transparent to the child's characteristics and could be used with a relatively heterogeneous population.

We also had to address several well-known factors relating to stimuli and task characteristics qualitatively, due to the limited number of studies that report this type of details. We noted that an effect of syllabic length of the nonword items has been demonstrated many times, in particular on the performance of children with DLD (e.g., Boerma et al., 2015; Thordardottir & Brandeker, 2013). Although analyzed systematically in fewer studies, an effect of syllabic complexity has also been shown (e.g., Dos Santos & Ferré, 2018). Wordlikeness and prosody have been studied infrequently and can also be important (e.g., Chiat & Roy, 2007). We were also able to highlight differences in the scoring method. The two most commonly used methods are PIC (counting correct items) and PPC (counting correct phonemes). Conveniently, it appears that PIC is faster, easier to use, and more discriminating (Le Clercq et al., 2017). We can therefore think that these factors, which are more complex to analyze quantitatively, could constitute other interesting moderators to deal with. In particular, Estes et al., (2007) had shown that the length of stimuli was an important moderator of effect size.

Despite significant methodological differences (e.g., type of task and stimuli, age of children, analysis), we noted that they unanimously mention the NWRT's high discriminative power for detection of DLD in monolingual and bilingual children. Most studies reported acceptable sensitivity and specificity (over 80% following Plante & Vance, 1994). The studies that obtained poor discriminative accuracy mainly concern language-specific tasks, where the scores obtained by children may depend on their language experience in the language of the task (Boerma et al., 2015; Kohnert et al., 2006). One solution for this issue may be the use of a quasi-universal task, like the one devised by Chiat (2015), which still allows for good discrimination. However, we noted that some of the poor discriminative accuracy scores also come from studies in a monolingual population (e.g., Conti-Ramsden, 2003). We can therefore hypothesize that methodological choices regarding the design of certain tasks can also contribute to low accuracy percentages.

Finally, it also appears that performance in NWRT is frequently correlated with that on other tasks and that it could be combined with other tools to increase diagnostic accuracy. Several studies associated NWRT with other tasks such as sentence repetition, expressive and receptive lexical tasks, narrative tasks, or a parental questionnaire, which all increased diagnostic accuracy to a certain extent. In absence of a clearly superior combination, clinician may be able to choose associated tasks depending on the child's profile of difficulty, the family's request, and what is available in a specific language. It should also be kept in mind that the NWRT highlights phonological and short-term verbal memory difficulties in children with DLD (Ferré & Dos Santos, 2015; Gray, 2004, Li'el et al., 2019). However, some children with DLD may not have phonological difficulties, making the complementary use of other tools in assessment sessions essential (Kapalková et al., 2013). This recommendation fits well with the long-standing clinical tradition to combine different diagnostic sources in order to establish with relative certainty the actual presence of DLD in a child (Bishop & McDonald, 2009), whether monolingual or bilingual. Moreover, combining tasks focused on the evaluation of procedural skills, as the NWRT, with tests assessing the child's declarative knowledge (e.g., lexical, grammatical, narrative skills), is also important for guiding intervention.

To conclude, we note that the NWRT is a tool that has been found to be effective in the studies reviewed and thus deserves a fixed place in language assessments of monolingual and bilingual children. Of course, as already mentioned, the type of task and stimuli need to be carefully chosen to fit the age and linguistic profile of the child, and NWRT should be complemented with other tools in order to obtain the most complete language profile possible and to corroborate any diagnostic indications from NWRT by different sources.

Acknowledgments

We would like to thank our colleague, Nuzhat, for her revision.

References

Afshar, M. R., Gorbani, A., Jalilevand, N., & Kamali, M. (2013). Providing the non-word repetition test and determining its validity and reliability and comparing phonological working memory in 4 to 6 Farsi-speaking normal and SSD children in Tehran City. *Journal of Research in Rehabilitation Sciences, 9*(5), 899–911.

Archibald, L. (2008). The promise of nonword repetition as a clinical tool. *Canadian Journal of Speech-Language Pathology and Audiology, 32*(1), 21–28. https://cjslpa.ca/files/2008_CJSLPA_Vol_32/No_01_1-68/Archibald_CJSLPA_2008.pdf

Archibald, L., & Joanisse, M. (2009). On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language, and Hearing Research, 52*(4), 899–914. https://doi.org/10.1044/1092-4388(2009/08-0099)

Armon-Lotem, S. (2012). Introduction: Bilingual children with SLI—The nature of the problem. *Bilingualism: Language and Cognition, 15*(1), 1–4. https://doi.org/10.1017/S1366728911000599

Armon-Lotem, S., & Chiat, S. (2012). How do sequential bilingual children perform on non-word repetition tasks? In A. K. Biller, E. Y. Chung, & A. E. Kimball (Eds.), *Proceedings of the 36th annual Boston University Conference on Language Development* (pp. 53–62). Cascadilla Press.

Armon-Lotem, S., & Meir, N. (2016). Diagnostic accuracy of repetition tasks for the identification of specific language impairment (SLI) in bilingual children: Evidence from Russian and Hebrew. *International Journal of Language & Communication Disorders, 51*(6), 715–731. https://doi.org/10.1111/1460-6984.12242

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

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

Bisiacchi, P. S., Cendron, M., Gugliotta, M., Tressoldi, P. E., & Vio, C. (2005). *BVN 5–11: Batteria di valutazione neuropsicologica per l'età evolutiva* [BVN 5–11: Neuropsychological assessment battery for the developmental age]. Centro Studi Erickson.

Boerma, T., & Blom, E. (2017). Assessment of bilingual children: What if testing both languages is not possible. *Journal of Communication Disorders, 66*(1), 65–76. https://doi.org/10.1016/j.jcomdis.2017.04.001

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning Dutch as a second language. *Journal of Speech, Language, and Hearing Research, 58*(6), 1747–1760. https://doi.org/10.1044/2015_JSLHR-L-15-0058

Bonifacci, P., Atti, E., Casamenti, M., Piani, B., Porrelli, M., & Mari, R. (2020). Which measures better discriminate language minority bilingual children with and without developmental language disorder? A study testing a combined protocol of first and second language assessment. *Journal of Speech, Language, and Hearing Research, 63*(6), 1898–1915. https://doi.org/10.1044/2020_JSLHR-19-00100

Bornstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2013). *Comprehensive meta-analysis Version 3*. Biostat.

Bortolini, U., Arfé, B., Caselli, C., Degasperi, L., Deevy, P., & Leonard, L. (2006). Clinical markers for specific language impairment in Italian: The contribution of clitics and non-word repetition. *International Journal of Language & Communication Disorders, 41*(6), 695–712. https://doi.org/10.1080/13682820600570831

Camilleri, B., & Law, J. (2007). Assessing children referred to speech and language therapy: Static and dynamic assessment of receptive vocabulary. *Advances in Speech Language Pathology, 9*(4), 312–322. https://doi.org/10.1080/14417040701624474

Castro, S. L., Caló, S., Gomes, I., Kay, J., Lesser, R., & Coltheart, M. (2007). *Provas de avaliação da linguagem e da afasia em português* [Language and aphasia assessment tests in Portuguese]. CEGOC.

Chiat, S. (2015). Non-word repetition. In S. Armon-Lotem, K. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 95–122). Multilingual Matters. https://doi.org/10.21832/9781783093137-008

Chiat, S., & Roy, P. (2007). The preschool repetition test: An evaluation of performance in typically developing and clinically referred children. *Journal of Speech, Language, and Hearing Research, 50*(2), 429–443. https://doi.org/10.1044/1092-4388(2007/030)

Coady, J., & Evans, J. (2008). Uses and interpretations of non-word repetition tasks in children with and without specific language impairments (SLI). *International Journal of Language & Communication Disorders, 43*(1), 1–40. https://doi.org/10.1080/13682820601116485

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Conti-Ramsden, G. (2003). Processing and linguistic markers in young children with specific language impairment (SLI). *Journal of Speech, Language, and Hearing Research, 46*(5), 1029–1037. https://doi.org/10.1044/1092-4388(2003/082)

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *The Journal of Child Psychology and Psychiatry, 42*(6), 741–748. https://doi.org/10.1111/1469-7610.00770

Conti-Ramsden, G., & Hesketh, A. (2003). Risk markers for SLI: A study of young language-learning children. *International Journal of Language & Communication Disorders, 38*(3), 251–263. https://doi.org/10.1080/1368282031000092339

Courcy, A. (2000). *Conscience phonologique et apprentissage de la lecture* [Phonological awareness and reading acquisition] (Unpublished doctoral dissertation, Université de Montréal).

De Almeida, L., Ferré, S., Morin, E., Prévost, P., dos Santos, C., Tuller, L., & Zebib, R. (2016). L'identification d'enfants bilingues avec trouble spécifique du langage en France. *SHS Web of Conferences, 27,* 10005. https://doi.org/10.1051/shsconf/20162710005

De Almeida, L., Ferré, S., Morin, E., Prévost, P., Dos Santos, C., Tuller, L., Zebib, R., & Barthez, M.-A. (2017). Identification of bilingual children with specific language impairment in France. *Linguistic Approaches to Bilingualism, 7*(3–4), 331–358. https://doi.org/10.1075/lab.15019.alm

Deevy, P., Weil, L. W., Leonard, L. B., & Goffman, L. (2010). Extending use of the NRT to preschool-age children with and without specific language impairment. *Language, Speech, and Hearing Services in Schools, 41*(3), 277–288. https://doi.org/10.1044/0161-1461(2009/08-0096)

Desmarais, C., Sylvestre, A., Meyer, F., Bairati, I., & Rouleau, N. (2008). Systematic review of the literature on characteristics of late-talking toddlers. *International Journal of Language &*

Communication Disorders, 43(4), 361–389. https://doi.org/10.1080/13682820701546854

Di Meo, S., Sanson, C., Simon, A., Bossuroy, M., Rakotomala, L., Rezzoug, D., Serre, G., Baudet, T., & Moro, M. R. (2014). Le bilinguisme des enfants migrants. In H. Bijleveld, F. Estienne, & F. Vander Linden (Eds.), *Multilinguisme et orthophonie. Réflexions et pratiques à l'heure de l'Europe* (pp. 149–171). Elsevier Masson.

Dispaldro, M., Benelli, B., Marcolini, S., & Stella, G. (2009). Real-word repetition as a predictor of grammatical competence in Italian children with typical language development. *International Journal of Language & Communication Disorders, 44*(6), 941–961. https://doi.org/10.3109/13682820802491794

Dispaldro, M., Deevy, P., Altoé, G., Benelli, B., & Leonard, L. B. (2011). A cross-linguistic study of real-word and non-word repetition as predictors of grammatical competence in children with typical language development. *International Journal of Language & Communication Disorders, 46*(5), 564–578. https://doi.org/10.1111/j.1460-6984.2011.00008.x

Dispaldro, M., Leonard, L. B., & Deevy, P. (2013). Real-word and nonword repetition in Italian-speaking children with specific language impairment: A study of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research, 56*(1), 323–336. https://doi.org/10.1044/1092-4388(2012/11-0304)

Dollaghan, C., & Campbell, T. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41*(5), 1136–1146. https://doi.org/10.1044/jslhr.4105.1136

Dos Santos, C., & Ferré, S. (2018). A nonword repetition task to assess bilingual children's phonology. *Language Acquisition, 25*(1), 58–71. https://doi.org/10.1080/10489223.2016.1243692

Dunn, G. (2014). Statistics in psychiatry. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 1136–1138). Springer Science & Business Media.

Ebert, K. D., Kalanek, J., Cordero, K. N., & Kohnert, K. (2008). Spanish nonword repetition: Stimuli development and preliminary results. *Communication Disorders Quarterly, 29*(2), 67–74. https://doi.org/10.1177/1525740108314861

Estes, K. G., Evans, J., & Else-Quest, N. (2007). Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 50*(1), 177–195. https://doi.org/10.1044/1092-4388(2007/015)

Ferré, S., & Dos Santos, C. (2015). Comment évaluer la phonologie des enfants bilingues. *Revue de linguistique et de didactique des langues, 51*(1), 11–34. https://doi.org/10.4000/lidil.3678

Gathercole, S. E., & Baddeley, A. D. (1996). *The children's test of nonword repetition*. The Psychological Corp.

Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory, 2*(2), 103–127. https://doi.org/10.1080/09658219408258940

Gray, S. (2003). Diagnostic accuracy and test–retest reliability of nonword repetition and digit span tasks administered to preschool children with specific language impairment. *Journal of Communication Disorders, 36*(2), 129–151. https://doi.org/10.1016/S0021-9924(03)00003-0

Gray, S. (2004). Word learning by preschoolers with specific language impairment. *Journal of Speech, Language, and Hearing Research, 47*(5), 1117–1132. https://doi.org/10.1044/1092-4388(2004/083)

Gray, S. (2006). The relationship between phonological memory, receptive vocabulary, and fast mapping in young children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 45*(5), 955–969. https://doi.org/10.1044/1092-4388(2006/069)

Grosjean, F. (2015). *Parler plusieurs langues: Le monde des bilingues.* Albin Michel.

Guiberson, M., & Rodríguez, B. L. (2013). Classification accuracy of nonword repetition when used with preschool-age Spanish-speaking children. *Language, Speech, and Hearing Services in Schools, 44*(2), 121–132. https://doi.org/10.1044/0161-1461(2012/12-0009)

Guiberson, M., & Rodríguez, B. L. (2015). Nonword repetition in Spanish-speaking toddlers with and without early language delays. *Folia Phoniatrica et Logopaedica, 67*(5), 253–258. https://doi.org/10.1159/000442745

Gutiérrez-Clellen, V., & Simon-Cereijido, G. (2010). Using nonword repetition tasks for the identification of language impairment in Spanish–English speaking children: Does the language of assessment matter. *Learning Disabilities Research & Practice, 25*(10), 48–58. https://doi.org/10.1111/j.1540-5826.2009.00300.x

Hedges, L. V., & Becker, B. J. (1986). Statistical methods in the meta-analysis of research on gender difference. In S. H. Hyde & M. C. Linn (Eds.), *The psychology of gender: Advances through meta-analysis,* (pp. 14–50). John Hopkins University Press.

Hodges, R., Munro, N., Baker, E., McGregor, K., & Heard, R. (2017). The monosyllable imitation test for toddlers: Influence of stimulus characteristics on imitation, compliance and diagnostic accuracy. *International Journal of Language & Communication Disorders, 52*(1), 30–45. https://doi.org/10.1111/1460-6984.12249

Horohov, J. E., & Oetting, J. B. (2004). Effects of input manipulations on the word learning abilities of children with and without specific language impairment. *Applied Psycholinguistics, 25*(1), 43–65. https://doi.org/10.1017/S0142716404001031

Jones, G., Tamburelli, M., Watson, S. E., Gobet, F., & Pine, J. M. (2010). Lexicality and frequency in specific language impairment: Accuracy and error data from two nonword repetition tests. *Journal of Speech, Language, and Hearing Research, 53*(6), 1642–1655. https://doi.org/10.1044/1092-4388(2010/09-0222)

Kapalková, S., Polisenská, K., & Vicenová, Z. (2013). Non-word repetition performance in Slovak-speaking children with and without SLI: Novel scoring methods. *International Journal of Language & Communication Disorders, 48*(1), 78–89. https://doi.org/10.1111/j.1460-6984.2012.00189.x

Kazemi, Y., & Saeednia, S. (2017). The clinical examination of non-word repetition tasks in identifying Persian-speaking children with primary language impairment. *International Journal of Pediatric Otorhinolaryngology, 93*(1), 7–12. https://doi.org/10.1016/j.ijporl.2016.11.028

Kohnert, K., Windsor, J., & Yim, D. (2006). Do language-based processing tasks separate children with language impairment from typical bilinguals. *Learning Disabilities Research & Practice, 21*(1), 19–29. https://doi.org/10.1111/j.1540-5826.2006.00204.x

Le Clercq, C. M. P., Van Der Schroeff, M. P., Rispens, J. E., Ruytjens, L., Goedegebure, A., Van Ingen, G., & Franken, M.-C. (2017). Shortened nonword repetition task (NWR-S): A simple, quick, and less expensive outcome to identify children with combined specific language and reading impairment. *Journal of Speech, Language, and Hearing Research, 60*(8), 2241–2248. https://doi.org/10.1044/2017_JSLHR-L-16-0060

Li'el, N., Williams, C., & Kane, R. (2019). Identifying developmental language disorder in bilingual children from diverse linguistic backgrounds. *International Journal of Speech-Language*

Pathology, 21(6), 613–622. https://doi.org/10.1080/17549507.2018.1513073

Marini, A., Marotta, L., Bulgheroni, S., & Fabbro, F. (2015). *Batteria per la valutazione del linguaggio in bambini dai 4 ai 12 anni*. Giunti O.S.

Marini, A., Ruffino, M., Sali, M. E., & Molteni, M. (2017). The role of phonological working memory and environmental factors in lexical development in Italian-speaking late talkers: A one-year follow-up study. *Journal of Speech, Language, and Hearing Research, 60*(12), 3462–3473. https://doi.org/10.1044/2017_JSLHR-L-15-0415

Martin, V., Renaud, J., & Dagenais, P. (2013). *Les normes de production des revues systématiques: Guide méthodologique*. National Institute of Excellence in Health and Social Services.

McDonald, J. L., & Oetting, J. B. (2019). Nonword repetition across two dialects of English: Effects of specific language impairment and nonmainstream form density. *Journal of Speech, Language, and Hearing Research, 62*(5), 1381–1391. https://doi.org/10.1044/2018_JSLHR-L-18-0253

McKean, C., Letts, C., & Howard, D. (2013). Developmental change is key to understanding primary language impairment: The case of phonotactic probability and nonword repetition. *Journal of Speech, Language, and Hearing Research, 56*(5), 1579–1594. https://doi.org/10.1044/1092-4388(2013/12-0066)

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine, 6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

Montgomery, J. (1995). Examination of phonological working memory in specifically language impaired children. *Applied Psycholinguistics, 16*(4), 355–378. https://doi.org/10.1017/S0142716400065991

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *The Journal of Child Psychology and Psychiatry, 57*(11), 1247–1257. https://doi.org/10.1111/jcpp.12573

Oetting, J. B., & Cleveland, L. H. (2006). The clinical utility of nonword repetition for children living in the rural south of the US. *Clinical Linguistics & Phonetics, 20*(7–8), 553–561. https://doi.org/10.1080/02699200500266455

Oetting, J. B., Cleveland, L. H., & Cope, R. F. (2008). Empirically derived combinations of tools and clinical cutoffs: An illustrative case with a sample of culturally/linguistically diverse children. *Language, Speech, and Hearing Services in Schools, 39*(1), 44–53. https://doi.org/10.1044/0161-1461(2008/005)

Paradis, J. (2019). English second language acquisition from early childhood to adulthood: The role of age, first language, cognitive, and input factors. In M. Brown & B. Dailey (Eds.), *Proceedings of the 43rd annual Boston University Conference on Language Development* (pp. 11–26). Cascadilla Press. http://www.lingref.com/bucld/43/BUCLD43-02.pdf

Paradis, J., Genesee, F., & Crago, M. B. (2011). *Dual language development & disorders: A handbook on bilingualism and second language learning* (2nd ed.). Brookes.

Paradis, J., Schneider, P., & Duncan, T. S. (2013). Discriminating children with language impairment among English-language learners from diverse first-language backgrounds. *Journal of Speech, Language, and Hearing Research, 56*(3), 971–981. https://doi.org/10.1044/1092-4388(2012/12-0050)

Pearson, B. Z. (2013). Distinguishing the bilingual as a late talker from the late talker who is bilingual. In L. Rescorla & P. S. Dale (Eds.), *Late talkers: Language development, interventions, and outcomes* (pp. 67–91). Brookes.

Petruccelli, N., Bavin, E. L., & Bretherto, L. (2012). Children with specific language impairment and resolved late talkers: Working memory profiles at 5 years. *Journal of Speech, Language, and Hearing Research, 55*(6), 1690–1703. https://doi.org/10.1044/1092-4388(2012/11-0288)

Pham, G., & Ebert, K. D. (2020). Diagnostic accuracy of sentence repetition and nonword repetition for developmental language disorder in Vietnamese. *Journal of Speech, Language, and Hearing Research, 63*(5), 1521–1536. https://doi.org/10.1044/2020_JSLHR-19-00366

Pham, G., Ebert, K. D., Dinh, K. T., & Dam, Q. (2018). Nonword repetition stimuli for Vietnamese-speaking children. *Behavior Research Methods, 50*(1), 1311–1326. https://doi.org/10.3758/s13428-018-1049-0

Plante, E., & Vance, R. (1994). Selection of preschool language tests. *Language, Speech, and Hearing Services in Schools, 25*(1), 15–24. https://doi.org/10.1044/0161-1461.2501.15

Rispens, J., & Baker, A. (2012). Nonword repetition: The relative contributions of phonological short-term memory and phonological representations in children with language and reading impairment. *Journal of Speech, Language, and Hearing Research, 55*(3), 683–694. https://doi.org/10.1044/1092-4388(2011/10-0263)

Rodekohr, R. K., & Haynes, W. O. (2001). Differentiating dialect from disorder: A comparison of two processing tasks and a standardized language test. *Journal of Communication Disorders, 34*(3), 255–272. https://doi.org/10.1016/S0021-9924(01)00050-8

Roy, P., & Chiat, S. (2004). A prosodically controlled word and nonword repetition task for 2- to 4-year-olds: Evidence from typically developing children. *Journal of Speech, Language, and Hearing Research, 47*(1), 223–234. https://doi.org/10.1044/1092-4388(2004/019)

Stokes, S. F., & Klee, T. (2009). The diagnostic accuracy of a new test of early nonword repetition for differentiating late talking and typically developing children. *Journal of Speech, Language, and Hearing Research, 52*(4), 872–882. https://doi.org/10.1044/1092-4388(2009/08-0030)

Stokes, S. F., Wong, A. M.-Y., Fletcher, P., & Leonard, L. B. (2006). Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. *Journal of Speech, Language, and Hearing Research, 49*(2), 219–236. https://doi.org/10.1044/1092-4388(2006/019)

Szewczyk, J. M., Marecka, M., Chiat, S., & Wodniecka, Z. (2018). Nonword repetition depends on the frequency of sublexical representations at different grain sizes: Evidence from a multifactorial analysis. *Cognition, 179,* 23–36. https://doi.org/10.1016/j.cognition.2018.06.002

Thal, D. J., Miller, S., Carlson, J., & Vega, M. M. (2005). Nonword repetition and language development in 4-year-old children with and without a history of early language delay. *Journal of Speech, Language, and Hearing Research, 48*(6), 1481–1495. https://doi.org/10.1044/1092-4388(2005/103)

Thordardottir, E. (2015). Proposed diagnostic procedures for use in bilingual and cross-linguistic contexts. In S. Armon-Lotem, K. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 331–359). Multilingual Matters. https://doi.org/10.21832/9781783093137-014

Thordardottir, E., & Brandeker, M. (2013). The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. *Journal of Communication Disorders, 46*(1), 1–16. https://doi.org/10.1016/j.jcomdis.2012.08.002

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

Topbaş, S., Kaçar-Kütükçü, D., & Kopkalli-Yavuz, H. (2014). Performance of children on the Turkish nonword repetition test: Effect of word similarity, word length, and scoring. *Clinical Linguistics & Phonetics, 28*(7–8), 602–616. https://doi.org/10.3109/02699206.2014.927003

Tuller, L., Abboud, L., Ferré, S., Fleckstein, A., Prérévost, P., dos Santos, C., Scheidnes, M., & Zebib, R. (2013). Specific language impairment and bilingualism: Assembling the pieces. In C. Hamann & E. Ruigendijk (Eds.), *Language acquisition and development: Proceedings of GALA*. Cambridge Scholars Publishing. https://www.unige.ch/fapse/logopedie/files/8314/2245/9669/12_article_2_P._Prevost.pdf

Tuller, L., Hamann, C., Chilla, S., Ferré, S., Morin, E., Prevost, P., dos Santos, C., Abed Ibrahim, L., & Zebib, R. (2018). Identifying language impairment in bilingual children in France and in Germany. *International Journal of Language & Communication Disorders, 53*(4), 888–904. https://doi.org/10.1111/1460-6984.12397

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (1999). *Comprehensive test of phonological processing*. Pro-Ed.

Washington, J. A., & Craig, H. K. (2004). A language screening protocol for use with young African American children in urban settings. *American Journal of Speech-Language Pathology, 13*(4), 329–340. https://doi.org/10.1044/1058-0360(2004/033)

Weismer, S. E., Tomblin, J. B., Zhang, X., Buckwalter, P., Chynoweth, J. G., & Jones, M. (2000). Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language, and Hearing Research, 43*(4), 865–878. https://doi.org/10.1044/jslhr.4304.865