

Accepted version of the following article:

Polišenská, K., Chiat, S., Szewczyk, J., Antonijevic, S., Blom, E., Boerma, T., Bohnacker, U., Chan, A., Chondrogianni, V., Fu, N. C., Gatt, D., Grech, H., Jezek, M., Kapalková, S., Kunnari, S., Maier, J., Mayer-Crittenden, C., Öberg, L., Schwob, S., Skoruppa, K., Tabone, N., Verhagen, J., & White, M. (2025).

Evaluation of the Crosslinguistic Nonword Repetition Test: Evidence From a Large and Diverse Secondary Data Set. *Journal of Speech, Language, and Hearing Research*, 1–21.

https://doi.org/10.1044/2025_JSLHR-25-00158

This version of the article has been accepted for publication in *Journal of Speech, Language, and Hearing Research* following peer review. The final published version is available at

https://doi.org/10.1044/2025_JSLHR-25-00158

Evaluation of the Crosslinguistic Nonword Repetition Test:

Evidence from a large and diverse secondary dataset

Kamila Polišenská^{1,2}, Shula Chiat¹, Jakub Szewczyk³, Stanislava Antonijevic⁴, Elma Blom⁵, Tessel Boerma⁶, Ute Bohnacker⁷, Angel Chan^{8,9}, Vasiliki Chondrogianni¹⁰, Nga Ching Fu⁸, Daniela Gatt¹¹, Helen Grech¹¹, Magdalena Jezek^{12,13}, Svetlana Kapalková¹⁴, Sari Kunnari¹⁵, Juliane Maier¹⁶, Chantal Mayer-Crittenden¹⁷, Linnéa Öberg⁷, Salomé Schwob¹⁸, Katrin Skoruppa¹⁸, Nadine Tabone¹¹, Josje Verhagen¹⁹, Michelle White²⁰

¹ Department of Language and Communication Science, City St George's, University of London, London, United Kingdom

² Division of Psychology, Communication and Human Neuroscience, The University of Manchester, Manchester, United Kingdom

³ Institute of Psychology, Jagiellonian University, Kraków, Poland

⁴ School of Health Sciences, National University of Ireland, Galway, Ireland

⁵ Department of Education and Pedagogy, Utrecht University, Utrecht, The Netherlands

⁶ Institute for Language Sciences, Utrecht University, Utrecht, The Netherlands

⁷ Department of Linguistics & Philology, Uppsala University, Uppsala, Sweden

⁸ Department of Language Science and Technology, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR

⁹ Research Centre for Language, Cognition, and Neuroscience, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR

¹⁰ School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, United Kingdom

¹¹ Department of Human Communication Sciences and Disorders, Faculty of Health Sciences, University of Malta, Msida, Malta

¹² Institute of Neurology of Senses and Language, Hospital of St. John of God, Linz, Austria

¹³ Research Institute for Developmental Medicine, Johannes Kepler University, Linz, Austria

¹⁴ Department of Communication Disorders, Comenius University, Bratislava, Slovakia

¹⁵ Research Unit of Logopedics, University of Oulu, Oulu, Finland

¹⁶ Institut für Deutsch als Fremdsprachenphilologie, Heidelberg University, Heidelberg, Germany

¹⁷ School of Speech-Language Pathology, Laurentian University, Sudbury, Ontario, Canada

¹⁸ Institut des sciences logopédiques, Université de Neuchâtel, Neuchâtel, Switzerland

¹⁹ Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam, The Netherlands

²⁰ Department of General Linguistics, Stellenbosch University, Stellenbosch, South Africa

Correspondence should be addressed to Kamila Polišenská, Department of Allied Health, City St George's, University of London, Northampton Square, London EC1V 0HB, United Kingdom
kamila.polisenska@city.ac.uk <https://orcid.org/0000-0001-7405-6689>

Statements and Declarations: No potential competing interest was reported by the authors.

Abstract

Purpose: The aim of this study was to evaluate the crosslinguistic validity of the Crosslinguistic Nonword Repetition test (CL-NWR) based on a large multi-country sample, by investigating factors related to language ability, as well as potential confounds.

Method: The data consisted of CL-NWR scores from children aged 37-165 months, collected by 18 research teams across 15 countries. Item-level analysis was employed to examine any non-desirable effects of gender, socioeconomic status, bilingual status and the amount of exposure to the test language, as well as desirable effects of age, item length, and clinical status (children categorized as typically developing [TD], with developmental language disorder [DLD], or with reported language concerns [LC], respectively). Subsamples were used to evaluate the consistency of findings across three time points and between different versions of the CL-NWR.

Results: Bayesian analysis provided strong evidence for the effects of age, item length, and clinical status on CL-NWR performance, as well as consistency across time points. In contrast, there was weak or no evidence for effects of gender, socioeconomic status, bilingual status, amount of exposure or test version. Additionally, there were two interactions between: i) item length and clinical status, suggesting that children with DLD found longer nonwords disproportionately more challenging than TD children, ii) age and clinical status, with the gap between TD and LC groups narrowing with age.

Conclusions: The CL-NWR was unaffected by environmental and demographic factors that often influence language assessments, including some nonword repetition tests. Performance was driven by factors reflecting language abilities. This makes the CL-NWR a unique and valuable tool for language assessment contributing to the identification of DLD in diverse linguistic, social, and geographical contexts.

Keywords: bilingualism; assessment; developmental language disorder

Developmental language disorder (DLD) affects approximately 7% of school-aged children, significantly impairing their ability to understand and produce language (Norbury et al., 2016). Children with DLD have been found to perform below typically developing (TD) peers on nonword repetition (NWR) tasks. This has led to extensive interest in NWR as a potential indicator of DLD, and one which is of particular interest for the assessment of bilingual children: since NWR is less dependent on knowledge of the language and less influenced by prior language knowledge compared to other assessments, poor NWR performance has the potential to indicate DLD in bilingual children independently of their exposure to each language. The Crosslinguistic Nonword Repetition Test (CL-NWR) was specifically developed for linguistically diverse children, including bilingual and monolingual children with varied language input. This paper aims to evaluate CL-NWR through analysis of a multi-sample crosslinguistic dataset.

The CL-NWR framework (Chiat, 2015; <https://www.bi-sli.org/cl-nonword-repetition>) provides a set of nonwords that are compatible with the diverse phonologies and lexical phonologies of the world's languages and would not disadvantage children with limited exposure to the language of testing. By design, the CL-NWR tests can be used with children whatever language(s) they speak, whether they were exposed to the language from birth or later, and whether it is their dominant or non-dominant language. While there has been considerable progress in evaluating the CL-NWR task by individual research teams on specific language versions of the CL-NWR (e.g., Boerma et al., 2015; Boerma & Blom, 2021, Chiat & Polišenská, 2016; Fu, Chan, et al., 2024; Fu, Chen, et al., 2024; Hamdani et al., 2025; Öberg & Bohnacker, 2022, 2024; White, 2021), no study has compared performance across different teams working with different populations. To further evaluate the cross-linguistic validity of the tool, and hence its potential as a culturally and linguistically inclusive assessment, this study draws together data from linguistically and geographically diverse samples of children collected by 18 research teams in 15 countries. We use this unique dataset to investigate the effects of two sets of factors on NWR performance:

- 1) Environmental factors that may disadvantage certain groups of children, leading to inaccurate conclusions about their language abilities (*socioeconomic status, bilingual status/amount of exposure to the language of testing, and test version*). Evidence of their impact on NWR performance would indicate biases in the test and undermine crosslinguistic applicability, so we will refer to them as ‘non-desirable’ factors.
- 2) Factors that have been found to influence performance on language-specific NWR tasks and that are indicative of children’s phonological processing, memory and development (*item length, age, clinical status, predictiveness across time*). These factors relate to abilities involved in processing and acquiring language, which the task seeks to assess. We will therefore refer to them as ‘desirable factors’.

There is now a vast literature on nonword repetition as a potential indicator of DLD, with studies conducted in many different countries using many different tests. This is the first study to bring together nonword repetition data from independent research studies using a unified test with children in geographically and linguistically diverse settings. As such, it is the first to evaluate whether a unified test can make a valid contribution to clinical identification of DLD regardless of children’s language experience, and whether this is an aim worth pursuing. While nonword repetition should always be considered with other sources of evidence, the contribution of a unified crosslinguistic test would be particularly valuable for assessment of bilingual children with variable experience of each language, and children speaking a language for which no formal assessments are available.

Non-desirable factors known to affect language tasks

Bilingualism and amount of exposure: While some studies of NWR have not found significant differences between monolingual and bilingual groups (e.g., Lee et al., 2013; Lee & Gorman, 2013; Thordardottir & Juliusdottir, 2013), others have found such differences (e.g., Cockcroft, 2016; Engel

de Abreu, 2011; Engel de Abreu et al., 2013; Kohnert et al., 2006; Messer et al., 2010; Windsor et al., 2010). Similarly, findings on amount of exposure in bilingual children have varied, with some reporting that tests were neutral to exposure (e.g., Thordardottir & Brandeker, 2013; Thordardottir, 2014; Scheidnes, 2020) while others report significant relations between NWR performance and exposure to the test language (Haman et al., 2017; Parra et al., 2011). These differences in outcome may well arise from differences in NWR tests administered, with some being language-independent while others are more language-specific and therefore susceptible to benefits of language experience and knowledge, and some including more complex phonological structures which are likely to be more language-specific. The CL-NWR, on the other hand, was designed to be compatible with diverse languages, maximally free of experience with a specific language, and therefore valid for crosslinguistic use.

Socioeconomic status (SES): Most studies that have addressed SES have found no effects on NWR performance (Farabolini et al., 2021; Farmani et al., 2018; Kalnak et al., 2014; Meir & Armon-Lotem, 2017; Polišenská & Kapalková, 2014; Sundström et al., 2018). Law et al. (2011) found that a sample of socially disadvantaged children performed in line with norms on a standardized test of NWR, in contrast to their low performance on assessments of receptive and expressive language. Chiat and Polišenská (2016) administered the CL-NWR and language-specific NWR tasks and found both to be free of effects of SES, although the effects on the language-specific task approached significance. As the CL-NWR was designed to minimize the contribution of specific language experience, the effect of SES should be negligible. However, this prediction needs to be evaluated empirically as none of the other studies that have employed CL-NWR investigated SES effects directly.

Test version: Where findings on NWR have been inconsistent across studies, differences have often been attributed to differences in the tests administered, particularly with respect to language-specificity of test items. The CL-NWR always comprises 16 items of 2-5 syllables, all with CV

structure. To be compatible with different languages, the CL-NWR offers options for each item, providing alternatives in case any one option is a real word in the testing language, or contains consonants that are not part of the language's phonetic inventory (Chiat, 2015). This ensures that the task remains appropriate as well as structurally consistent regardless of the language being tested, making it a flexible tool for cross-linguistic language assessments, but results in different versions of the test depending on the option selected for each item (e.g., English version: *sipula*, Dutch version: *zibula*). In addition, items selected for each version were recorded using phonetic realizations appropriate to the test language (e.g., Cantonese CL-NWR was recorded by a native speaker of Cantonese). Though closely matched in item content, differences between test versions could be a factor in performance with implications for crosslinguistic applicability. Since two of our contributing research teams administered two different versions of the CL-NWR (British English and Dutch; Swedish and Arabic) to their participants, we are able to investigate the effect of test version within-subject. If test version has a notable effect on performance within-subject, the sources of difference would need to be considered, and findings on one version (for example, on effects of bilingualism or DLD) could not be generalized to another version or to a different population. On the other hand, if two versions produce very similar results within-subject, generalization of findings for one version and use of one version with other populations may be valid.

Desirable factors known to affect NWR performance

Item length: Effects of length are taken to be a measure of phonological memory. Length was the only item factor that was applicable across the diversity of human languages and hence compatible with the aims of the CL-NWR; key factors of phonological complexity and phonotactic probability, tapping into phonological processing and representations, are language-specific and therefore incompatible with a crosslinguistic test. To our knowledge, almost every study to date that has investigated length in NWR has found significant effects on performance, and significant interactions

with clinical groups showing stronger effects of length (e.g., Ahufinger et al., 2021; Boerma et al., 2015; Dispaldro, Leonard & Deevy, 2013; Graf Estes et al., 2007).

Age: Developmental increases have been found consistently across different NWR tasks, age ranges, languages and populations (in monolingual children, e.g., Spanish: Guiberson & Rodriguez (2016); English: Gathercole et al., 1994); Slovak: Polišenská & Kapalková (2014); Cantonese: Stokes et al. (2006), and in bilingual children, Duncan & Paradis, 2016). Meta-analyses and systematic reviews of differences between TD and DLD groups (e.g., Graf-Estes et al., 2007; Schwob et al., 2021) have reported no effect of age on differentiation.

Clinical status: A key motivation for the clinical use of NWR is its potential as a quick assessment helping to identify DLD and guide support for children, with a particular advantage for children with diverse language backgrounds. Two meta-analyses/systematic reviews have addressed the diagnostic accuracy of NWR in bilingual (Ortiz, 2021) and both bilingual and monolingual children (Schwob et al., 2021). Schwob et al. (2021) reported that the mean effect size across all studies was large, with DLD groups consistently performing below TD groups. Overall, the included studies reported higher specificity, reflecting the higher accuracy in correctly classifying TD children, and a larger effect size for the monolingual groups than for the bilingual groups. Ortiz also reported significant variation in classification accuracy across included studies; the type of task used affected discriminatory power, with quasi-universal tasks providing a higher mean effect size than language-specific tasks, and the poorer discriminant results in bilingual groups arising mainly in studies that used language-specific tasks. Quasi-universal tasks are those designed to be compatible with phonologies of different languages; they are deemed 'quasi-universal' on the grounds that language-specific influences cannot be entirely eliminated (Chiat, 2015). Studies using the CL-NWR, designed to be compatible with diverse phonologies, have reported mixed results, with some finding differences between TD children and children with DLD (e.g., Boerma et al., 2015; Fu, Chan, et al., 2024; Fu, Chen, et al.,

2024), while others report no significant differences (e.g., Öberg & Bohnacker, 2022, 2024). The observed differences are likely due to how DLD is operationalized, the inherent heterogeneity of DLD, and small sample sizes.

Predictiveness across time: Studies that have administered NWR tests at two or more timepoints have found performance to increase and significantly correlate across time (Boerma & Blom, 2021; Chiat & Roy, 2013; Gathercole et al., 1994; Melby-Lervåg et al., 2012; Næss et al., 2015; White, 2021). Such findings offer further evidence that the administered tests measure specific and developing skills. Two studies contributing to our unique dataset repeated the CL-NWR at three timepoints, allowing us to evaluate the stability of performance.

Findings on the effects of the above two sets of factors (desirable and non-desirable), based on diverse samples and versions of the CL-NWR tests, will indicate the extent to which the CL-NWR is a valid and clinically informative measure of skills regardless of children's language background and experience. This will contribute to the development of screening tools for language disorders that can be used across different linguistic and cultural settings. Such tools can contribute to early identification of DLD, enabling timely intervention.

Aims and research questions

The current study collates CL-NWR datasets from 18 research teams in 15 countries to assess the crosslinguistic validity of the CL-NWR, by investigating the extent to which performance is affected by 'non-desirable' and 'desirable' factors in this large and diverse sample.

Research questions related to non-desirable factors:

RQ1: Is CL-NWR accuracy affected by bilingual status, amount of exposure to the language of the test recording, or SES?

Hypothesis: The task will show negligible effects of language background and experience as reflected by variation in bilingual status, amount of exposure or SES, and therefore be bias-free. Children from disadvantaged/minority groups are at a higher risk of being misdiagnosed if clinical tools are not adapted to their specific needs and backgrounds. If no SES/bilingualism effects are found, as we would expect, this would confirm that the CL-NWR task is inclusive.

RQ2: In the subsample of children who completed two distinct versions of the CL-NWR tasks, is accuracy affected by CL-NWR test version?

Hypothesis: Based on the design of the CL-NWR, we expect accuracy not to be substantially affected by the test version used. This outcome would suggest that the different language versions of the task are comparable in terms of measuring a child's NWR performance.

Research questions related to desirable factors:

RQ3: Is CL-NWR accuracy affected by (i) children's chronological age and (ii) item length?

Hypothesis: The CL-NWR will be age-sensitive, and therefore informative about development. Longer nonwords will lead to lower repetition performance. Additionally, we do not anticipate that any participant-related factors identified as robust predictors in RQ1 will interact with length.

RQ4: In the subsample of children who were followed longitudinally, does CL-NWR accuracy at T1 and T2 reliably predict CL-NWR accuracy at T3?

Hypothesis: Accuracy at T1 and T2 will predict CL-NWR accuracy at T3, taking into account participant factors identified as predictive in RQ1. This would indicate that NWR performance captures a stable ability and that earlier performance can be used to predict future accuracy in this task.

RQ5: Is CL-NWR accuracy affected by clinical status?

Hypothesis: Given the overall findings of the two systematic reviews (Schwob et al., 2021 and Ortiz, 2021) and some findings on the CL-NWR specifically, we expect CL-NWR performance to differ between TD and clinical groups (see Method for definitions of clinical groups within this study). However, the majority of the NWR tests reviewed in the systematic reviews are either language-specific or, in the case of some quasi-universal tests, include language-specific items (dos Santos & Ferré, 2018), and such items may enhance discrimination between TD and clinical groups (as found with monolingual children, Graf-Estes et al., 2007). To be compatible with diverse phonologies, CL-NWR uses a limited range of phonological features compared to more language-specific NWR tasks and items, and this may affect the magnitude of difference between TD and clinical groups. In addition, our analysis differs from previous studies because it collates multiple independent samples that vary in the criteria used to identify clinical groups, whereas within-study analysis of TD/DLD group differences is based on a single criterion or set of criteria for allocation to the DLD group. The heterogeneity in the operationalisation of DLD evident in previous research is also apparent in our datasets (see Method section for details and see Discussion).

RQ6: Is there a length by clinical status interaction?

Hypothesis: Clinical groups will show disproportionately reduced accuracy with increasing item length compared to TD children.

RQ7: How accurately does the CL-NWR distinguish between children with DLD and their TD peers in different age bands, and what are the optimal age-specific thresholds for identifying DLD risk?

Hypothesis: The CL-NWR will show acceptable classification accuracy by age bands, with optimal thresholds increasing with age, reflecting developmental changes in task performance.

The current study will implement Bayesian statistical approaches. While these are increasingly being used in research on neurodevelopmental disorders, they are not yet as common as traditional frequentist methods. Language development/disorders research often involves small samples, reflecting the difficulties in recruiting participants. This can limit the power of traditional statistical methods and lead to difficulties with interpreting null results. Bayesian methods mitigate this limitation, by providing a framework for directly estimating the probability of a hypothesis given the data, rather than relying on p-values, which can be misleading in small samples. Unlike frequentist approaches, they allow for the incorporation of prior knowledge, which helps stabilize estimates and improve inference when data are limited. Additionally, Bayesian methods offer a clearer interpretation of null findings by quantifying the degree of evidence for or against a hypothesis, rather than simply failing to reject the null. This allows researchers to draw more nuanced conclusions, reducing uncertainty and the risk of misinterpretation in studies with small samples.

Method

Participants

Anonymized datasets from 18 teams spanning 15 countries and 17 CL-NWR tests were included. Each data controller confirmed that the collection of their data complied with requirements for ethical approval in their institution, and that secondary data analysis for this study entitled *A Crosslinguistic Comparison of Performance on the Crosslinguistic Nonword Repetition Test* fell within the remit of the original consent provided by participants. As the current study involved secondary data analysis only, with no access to identifiable information or original materials such as scoring sheets or recordings, separate ethical approval for the secondary analysis was not required.

Table 1 provides an overview of the datasets submitted by the 18 research teams, indicating the country of participants, the language version of the CL-NWR used, the number of participants (N), along with the mean ages (M) and standard deviations (SD) for each group included in the study. Since the outcome variable in the analyses is accuracy of item repetition rather than total score, we

include the number of datapoints for each sample (including different test versions and time points where more than one was administered), in addition to the number of children.

Table 1. Participant demographics.

Country	Language version	N participants	N datapoints	Age			
				Min	Max	Mean	SD
Austria‡	German	156	2,434	52	65	58.8	3.3
Canada*	English	44	704	56	73	64.4	4
Canada*	French	84	2,799	54	109	83	15
Finland‡	Finnish	98	1,568	48	84	66.4	10.9
Germany*,‡	German	126	2,016	43	165	98.9	30.3
Greece	Greek	60	960	48	71	59	7.2
Hong Kong*	Cantonese	38	608	96	142	115.8	11.7
Ireland*, ‡	English	101	1,616	66	136	82.0	10.2
Malta Team 1‡	Language-neutral	18	288	37	80	55	13.2
Malta Team 2	Maltese	100	1,599	43	70	61.8	6.1
Netherlands Team 1‡	Dutch	203	3,100	59	81	68.2	4.6
Netherlands Team 2*	Dutch	250	11,357	54	116	82	12
Singapore	Mandarin	36	576	42	76	57.6	10.3
Slovakia*, ‡	Slovak	99	1,584	48	94	67.1	10.4
South Africa	English	34	1,632	60	80	70.5	4.6
Sweden*	Arabic	109	1,744	48	96	72.9	13.5
Sweden*	Swedish	109	1,744	48	96	73	13.4
Switzerland*	Swedish	53	848	60	95	78.1	11.3
UK England‡	Dutch	97	1,552	41	131	75.1	15.2
UK England‡	English	243	3,888	37	131	71	14.9
UK Scotland*	English	52	832	73	98	88.5	6.9

Note. * Team included children with DLD. ‡ Team included children with LC. Min = minimum; Max = maximum

The uniqueness and strength of our study is that it includes children across diverse language contexts that vary in numbers and typology of language(s) used, and in the nature of bilingualism.

Some samples come from countries or regions that are bilingual and where being exposed to two or

more languages is the norm (e.g., Malta, Singapore) or typical of many communities (e.g., Canada, South Africa); some come from countries that have a dominant language and included bilingual children exposed to a specific language combination (e.g., Greek/Albanian or Greek/Russian in Greece, English/Scottish Gaelic in Scotland, Arabic/Swedish in Sweden, Portuguese/French in Switzerland); others from countries with a dominant language included bilingual children exposed to heterogenous language combinations (Finland, the Netherlands, UK-England, Germany, Austria, Ireland). In some of these samples, certain minority languages were more common than others (e.g., Turkish/Arabic/Kurdish in Germany; Turkish/Tarifit Berber/Moroccan Arabic in the Netherlands; Polish/Lithuanian in Ireland). Other samples spanned a large number of minority languages (e.g., around 40 different languages in the Austrian and UK samples).

The majority of the research teams administered parental questionnaires obtaining background information about their samples. The tools differed across teams and there were some missing data (see section on Analysis Plan for further details). Amount of exposure appeared to be the most readily available parameter across the parent questionnaires administered by our research teams, in line with a recent review by Kaščelan et al. (2022). Their study examined 48 questionnaires documenting children's bilingual experiences, identifying 32 overarching constructs. Among these, exposure was the most consistently represented construct, featuring in 96% of the questionnaires. We aimed to obtain a rough measure of language exposure to the language of the CL-NWR recording in order to find out if the amount of exposure to the language of testing is a robust predictor of nonword repetition accuracy. From the questionnaire data available, teams were asked to provide an estimate of children's exposure to the language of testing, i.e. language of the recording of CL-NWR, on a percentage scale where 100 would describe the situation of monolingual children tested in their own language (e.g., a monolingual Dutch child who is only exposed to Dutch and is tested with a Dutch CL-NWR version), and 0 would describe a situation of a monolingual or bilingual child never exposed to the test language (e.g., a bilingual Portuguese/French child tested with a Swedish CL-NWR version) or a child newly arrived in a country and not yet exposed to the language of testing

(e.g., a Polish-speaking child in the early stages of acquiring English in the UK tested on the English CL-NWR version). The rationale for eliciting percentage ratings was that this format was judged to be simplest for translating the different types of quantitative and qualitative data collected by teams into a numerical rating. The primary goal was to enable a straightforward and consistent method of aggregation across data types. Moreover, the use of percentage ratings aligns with the approach adopted in other published questionnaires such as Parents of Bilingual Children Questionnaire (PaBiQ) by Tuller (2015). Use of this measure is in line with views of bilingualism as a continuous variable (e.g., Baum & Titone, 2014; de Bruin, 2019; Luk & Bialystok, 2013).

The background questionnaires were also used to obtain information about maternal education as a proxy for SES in the sample. Since our research was international, we needed a classification that would be applicable and permit comparison across different countries. We adopted the International Standard Classification of Education scale (UNESCO Institute for Statistics, 2011) and collapsed some categories resulting in the following five: 0 – *no education* ('less than primary' for educational attainment), 1 – *primary education*, 2 – *secondary education*, 3 – *post-secondary/further/non-university education*, 4 – *university education*. Again, the categories were similar to those in PaBiQ (Tuller, 2015), but we provided an additional category of 'no education' as it could not be assumed that all parents in a study with a very diverse sample would have had access to education.

Classification of children's clinical status posed challenges because there is no gold standard for identifying DLD in bilingual children, and different standards, labels and tools apply across different educational and clinical settings internationally (e.g., Bishop et al., 2017; Law et al., 2019). For the purposes of this study, we based clinical classification of participants on a combination of information from the background questionnaires, place of recruitment and results of language testing where this was carried out, as follows:

- Typically developing (TD): assigned if children were not recruited by clinicians; no concerns had been raised about their language; they had no diagnosis of DLD; and if they were tested

with language assessments, they did not meet DLD criteria as set out by individual research teams.

- Developmental Language Disorder (DLD): assigned if children had a clinical diagnosis of DLD and/or, if tested with language assessments, they met the research team's DLD criteria.

We included an intermediate category, for children who did not meet the criteria for DLD but there was evidence of concerns about their language:

- Language concerns (LC): assigned if children were recruited by clinicians or had raised LC; did not have a clinical diagnosis of DLD and/or, if tested with language assessments, they did not meet the research team's DLD criteria.

Materials

Each research team presented a version of the Crosslinguistic Nonword Repetition test which was constructed and audio-recorded following The Crosslinguistic Nonword Repetition framework (CL-NWR; Chiat, 2015). Sixteen items were selected from this framework, four at each length between 2 and 5 syllables, all with consonant-vowel (CV) syllable structure. Chiat (2015) describes in detail the rationale for creating the items and the selection process for individual test versions. The test items used by individual teams are provided in the Appendix C. As can be seen, some items were selected more often than others. In line with the framework, each team recorded the items and all but three embedded their audio files in a PowerPoint presentation. The format of presentation varied. The majority of the teams (11 out of 18) used the bead format described in Polišenská and Kapalková (2014) in which nonwords are presented as part of a game in which children are asked to help the researcher reconstruct a necklace by repeating a magic word (i.e. nonword item). Two teams used an alien, two used talking parrot/animals to present the items, and three used audio presentation without visual support. Nonwords were presented in a fixed randomized order, and the order was randomized by each research team.

Procedure and scoring

Children were tested individually, and their responses were audio recorded for scoring. Scoring was carried out by each team independently and reliability measures are presented for each dataset where available in Table 2. The reliability is generally strong, with most agreement percentages/ICC values well within acceptable ranges for robust measurements. The majority of the teams followed the scoring method described in Chiat and Polišenská (2016). Whole-item scoring was chosen over other types of scoring based on previous findings (e.g., Boerma et al., 2015) and has since then received support in a systematic review conducted by Schwob et al. (2021) which reports that the majority of the studies reviewed used whole-item scoring, and more importantly, that whole-item scoring showed better sensitivity than percentage of phonemes correct. On a practical level, whole-item scoring is also faster and more suitable for clinical purposes.

Table 2. Interrater reliability for the CL-NWR across the research teams.

Country	Percentage of Sample Scored by 2nd Rater	Reliability Metric
Austria	Not reported	No information available
Canada	20%	97.6% agreement
Finland	10%	99.2% agreement
Germany	50.5%	K = .997
Greece	17%	ICC = .961
Hong Kong	26.3%	ICC = .95 (95% CI [.82, .99])
Ireland	17%	ICC = .984 (95% CI [.951, .995])
Malta Team 1	10%	93.9% agreement
Malta Team 2	10%	93.4% agreement
Netherlands Team 1	10%	89% agreement
Netherlands Team 2	75%	ICC = .97
Singapore	27%	ICC = .99
Slovakia	10%	ICC = .971 (95% CI [.884, .993])
South Africa	10%	Correlation $r = .786$
Sweden	15%	97.92% agreement
Switzerland	15%	100% agreement
UK (England)	22%	ICC = .97
UK (Scotland)	10%	95.5% agreement

Note. ICC = intraclass correlation coefficient; CI = confidence interval.

In the CL-NWR, whole items are correct if they contain *all* and *only* the segments in the target in the correct order. Hence, *omissions*, *substitutions*, and *additions* are scored as errors. Segments are correct if they fall within the target segmental category. Even if they are phonetically distorted, they are scored as correct provided they are perceived as closer to the target category than any neighbouring category. Changes in prosody are not penalized, but categorical changes in vowel length are scored as errors. Replacement of a full vowel with schwa is scored as incorrect, e.g., [ˈluˈmi.gə] for [ˈluˈmi.gə]. Allowances are made for: 1) *Immature speech*: segmental substitutions that are *relatively consistent* in the child's productions and are characteristic of immature speech, for example, stopping of fricatives, fronting of velar stops. 2) *Accent/dialectal variation*: Segmental substitutions that are consistent with the child's accent/dialect. 3) *Intermediate realizations of targets*: consonants that are borderline between voiced and voiceless (e.g., [s/z]), the consonant is scored wrong if it is judged to be *definitely* on the wrong side of the continuum. However, in some teams (Netherlands, Canada, Sweden and Austria) repetitions with only additions were considered correct on the grounds that there was no loss of information. Previous research indicates that additions are a small proportion of errors (Burke & Coady, 2015; Edwards & Lahey, 1998; and see Analysis Plan regarding methodological variation).

A subsample of children received the CL-NWR test multiple times. The children in the UK sample and Swedish sample received two different versions of the CL-NWR. Dutch and British English CL-NWR versions were administered to the UK sample within the same session; Swedish and Arabic CL-NWR versions were administered to the Swedish sample in two sessions about a week apart. The order of the tests was counterbalanced. Two teams administered their CL-NWR tests at three time points. The gaps between testing times for children were 4-5 months for South Africa, and 11-12 months for most children in the Netherlands sample.

Analysis Plan

The data used in this study is available on the Open Science Framework (OSF) at

https://osf.io/2wu8r/?view_only=1e31708359f545d6b53c2da6ca727eac

Missing data. This is a large-scale project with data collected independently by several research teams, across different countries, which operate with different clinical labels and vary in the number and type of assessment tools available. In addition, data collection took place in different settings, and sometimes as part of a wider research project. As a result, and as is often the case in large population studies, some participant background information was missing. In order to address our research questions, we evaluate some models using subsets of data that included all relevant predictors, and others including all datasets and limiting variables to those available for all datasets. If a variable turned out not to be predictive in the subset of data that included the variable (e.g., if maternal education was not a credible predictor in all datasets that measured it), we would exclude it from the final analyses to maximize sample and maintain representativeness across all the research teams.

The specific variables entering each analysis will be described next to the results of corresponding analyses. The Bayesian models were fitted using the *brms* package in R (Bürkner, 2017). All categorical predictors were deviation coded (mean 0, difference between conditions 1), while the continuous predictors were standardized. We modelled the maximal random effects structure. For all predictors, we used regularizing Gaussian priors (mean: 0, SD: 2). Models were fitted using 21000 sampling iterations (not including warmup), collected across 12 chains.

In Bayesian analysis, group differences are not assessed through binary significance testing but through the estimation of the magnitude and uncertainty of effects. Interpretation centres on the credible interval (in our case, 95%, but it could be any other interval), which indicates the range within which the true effect is likely to lie with a given probability. This approach allows for a more graded and informative interpretation of evidence, reflecting both the direction and strength of effects without reducing conclusions to a dichotomy of “significant” or “not significant”, and instead

focusing more on the estimation of the effect size and its reliability (e.g., weak or non-existent vs. strong).¹

We first investigate the effects of participant- and item-related factors on NWR scores in TD children in order to establish whether the non-desirable factors (bilingualism, amount of exposure, gender, SES) and desirable factors (chronological age, item length) affect performance (RQs 1 and 3). To account for variation in the number of data points across countries and language versions, we employed mixed-effects models with participants, items, and research teams included as random effects, ensuring robust handling of unequal group sizes.

Another set of questions (RQ5/6) addressed differences due to clinical status (desirable factor). The analyses focusing on DLD children and children with LC only took into account children from research teams that tested children from both compared groups (TD and DLD, or TD and children with LC). In this way, children were tested in the same languages, using the same tools and by the same researchers, minimizing factors that could spuriously affect the group effect. Any of the participant-related factors found to substantially impact NWR scores in TD children will be controlled at the next step; this ensures that any effects of clinical status are not confounded by variables such as age, gender, bilingualism, SES or language exposure.

The effects of different CL-NWR versions on NWR accuracy in two samples where two different CL-NWR versions were administered (Dutch and English in the UK; Swedish and Arabic in Sweden), will be addressed with two types of analyses (RQ2). The first one involves separate models for UK and Swedish data, where we check whether the language and the order of test administration affect the accuracy. This model looks for sources of potential differences between the tests. The second type of analysis looks at commonalities, i.e. correlations of test results from two language versions aggregated by participant.

¹ Alongside the main parameter estimates, diagnostic measures such as Rhat and effective sample size (ESS) are reported to assess the reliability of the Bayesian analysis. Rhat indicates whether the model has converged properly, with values close to 1 suggesting that the estimates are stable and trustworthy. ESS reflects how much useful, independent information the model has drawn from the data when estimating a parameter. Although these values are not used to interpret the meaning of the results, they provide important reassurance that the statistical model has performed well and that the findings can be considered robust.

Longitudinal data from two teams (Netherlands, South Africa) who administered the CL-NWR at three time points allow us to investigate whether CL-NWR accuracy from the T3 can be reliably predicted using CL-NWR accuracy data from T1 and T2 (RQ4). To express the strength of the link between previous tests (T1 and T2) and the test at T3, a model which averaged scores over participants (having a by-participant logit score in each testing stage) will be run.

Results

Non-desirable factors:

The first research question estimated the fixed effects of bilingual status, amount of exposure to the language of the test recording, and maternal education, while controlling for children's age and gender. The random variables included item (i.e. nonword), participant and research team. As discussed in the analysis plan, to maximize the number of data points, we evaluated several models using subsets of data that included all relevant predictors. The first model included children's age, bilingualism status, amount of exposure, maternal education and gender as fixed effects and item, participant and research team as random effects. In subsequent models, we eliminated exposure and maternal education as not contributing to the model and containing a lot of missing data, thus significantly constraining the pool of participants (please see Appendix A and Appendix B for the intermediate models). The final model includes all other predictors regardless of their significance, as they did not have many missing values. The final model (see Table 3) evaluated the effect of children's age, bilingualism status, and gender as fixed effects. From these remaining variables, only children's age had a robust effect on CL-NWR accuracy: older children repeated the nonwords more accurately (see *Desirable factors* below). The estimated effect for bilingual status was weak. Extended models presented in Appendix A and B, which included additional covariates, provided further support for this interpretation. Across these more strictly controlled models, no consistent evidence emerged for an effect of bilingualism, suggesting that this variable did not exhibit a stable or robust association with task performance.

Table 3. Final model estimating the effects of participant factors.

Parameter	Estimate	Est. Error	95% CI (lower)	95% CI (upper)	Rhat	Bulk ESS	Tail ESS
Intercept	1.32	0.24	0.85	1.79	1.00	2,421	6,500
Age	0.50	0.14	0.22	0.78	1.00	9,098	12,170
Bilingual status	-0.32	0.13	-0.59	-0.07	1.00	13,872	14,699
Gender	0.13	0.10	-0.09	0.32	1.00	15,447	15,279

Note. CI = Credible Interval; Rhat = Convergence Diagnostic; Bulk ESS = Effective Sample Size (bulk); Tail ESS = Effective Sample Size (tail). Random effects were estimated for research team (n=17), nonword (n=79), and participant (n=1,293).

RQ2: Test versions

The effect of test version was addressed in datasets from two research teams who administered two different versions of the CL-NWR to their sample. This enabled us to investigate the effect of test version within-subject and independently of research team. The models run to address this question evaluated the fixed effects of the language version and the order in which the versions were administered, in addition to length. The models were analysed separately for each dataset (i.e., UK subsample with Dutch/English versions; Swedish sample with Swedish/Arabic version), as it was only appropriate to compare the effect of the test version within the same research team sample. The results were similar: Both in the UK and Swedish samples, neither language nor order had a notable effect (credible intervals in these effects included 0; see Table 4 and Table 5). However, in the Swedish sample, most of the posterior distribution mass for language was above zero, suggesting a trend toward lower accuracy for the Arabic version (as reported in Öberg's thesis, 2020). Turning to commonalities, correlational analyses showed that the language versions across both samples (aggregated by participant) are significantly positively associated: UK $r = .77, p < .0001$, Sweden $r = .63, p < .0001$.

Table 4. Comparison of two CL-NWR test versions in the UK sample.

Predictor	Estimate	Standard Error	95% CI (lower)	95% CI (upper)	Rhat	Bulk ESS	Tail ESS
Intercept	1.94	0.25	1.46	2.44	1.00	2,594	4,315
Length	-1.77	0.22	-2.23	-1.34	1.00	2,966	5,065
Language	-0.19	0.27	-0.73	0.34	1.00	4,401	7,456
Order	0.07	0.13	-0.20	0.33	1.00	11,598	8,421

Note. CI = Credible Interval; Rhat = Convergence Diagnostic; Bulk ESS = Effective Sample Size (bulk);

Tail ESS = Effective Sample Size (tail). Random effects were estimated for nonword (n=31), and participant (n=97).

Table 5. Comparison of two CL-NWR test versions in the Swedish sample.

Predictor	Estimate	Standard Error	95% CI (lower)	95% CI (upper)	Rhat	Bulk ESS	Tail ESS
Intercept	1.50	0.24	1.02	1.98	1.00	2,479	4,625
Length	-1.67	0.21	-2.08	-1.24	1.00	3,275	4,274
Language	-0.35	0.20	-0.74	0.05	1.00	7,657	8,278
Order	0.13	0.20	-0.26	0.52	1.00	9,193	8,645

Note. CI = Credible Interval; Rhat = Convergence Diagnostic; Bulk ESS = Effective Sample Size (bulk);

Tail ESS = Effective Sample Size (tail). Random effects were estimated for nonword (n=28), and participant (n=108).

Desirable factors:

RQ3: Item length and chronological age

The models investigating non-desirable factors took age into account and found this to be the only robust factor. We evaluated the effect of nonword length on repetition accuracy, controlling for age. The results replicated the effect of age, and showed that longer nonwords were much more difficult to repeat. However, these effects appeared additive as their interaction was negligible (see Table 6).

This indicates that the effect of age on repetition accuracy is consistent regardless of the length of the nonword, and vice versa.

Table 6. Model estimating the effect of length, age and their interaction.

Parameter	Estimate	Est. Error	95% CI (lower)	95% CI (upper)	Rhat	Bulk ESS	Tail ESS
Intercept	1.33	0.22	0.89	1.77	1.00	5,009	9,183
Age	0.50	0.12	0.26	0.74	1.00	10,073	12,550
Length	-1.15	0.15	-1.44	-0.86	1.00	6,088	10,231
Age:Length	0.04	0.05	-0.07	0.14	1.00	17,567	15,305

Note. CI = Credible Interval; Rhat = Convergence Diagnostic; Bulk ESS = Effective Sample Size (bulk);

Tail ESS = Effective Sample Size (tail). Random effects were estimated for research team (n=18), nonword (n=79), and participant (n=1,422).

RQ4: Predictiveness across time

Two research teams (The Netherlands, South Africa) administered the CL-NWR at three time points (T1-3), enabling us to investigate within-subject how NWR performance changes with age and the consistency of performance across time. One way to address this question is to evaluate how well the total scores at T3 were predicted by scores at T1 and T2. This analysis was conducted on typically developing children only, on the single item level. The model (see Table 7) confirmed an effect of research team, child's age, and importantly, previous CL-NWR score. It also confirmed a very weak effect of the distance between testing stages – the longer the distance the better the scores at T3. There was no notable interaction between previous accuracy and the time between testing stages. Overall, this analysis shows that T3 scores can be estimated based on age alone, but accuracy greatly improves when incorporating the child's score on the same item at previous time points. In other words, each child has their own trajectory of acquiring the skills/knowledge necessary to repeat each nonword.

Table 7. Model estimating the predictiveness of time on CL-NWR accuracy.

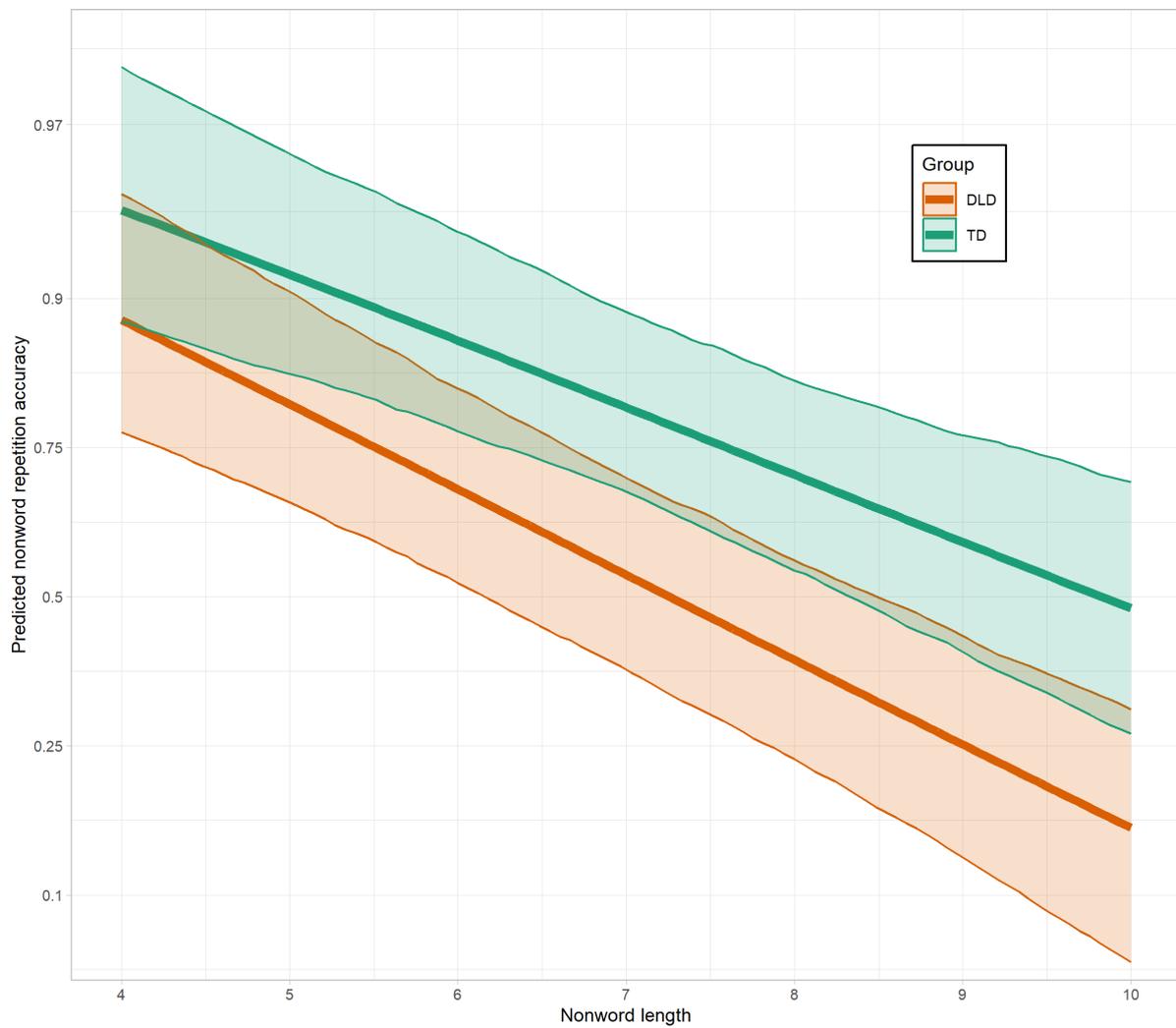
Predictor	Estimate	Standard Error	95% CI (lower)	95% CI (upper)	Rhat	Bulk ESS	Tail ESS
Intercept	1.82	0.29	1.25	2.39	1.00	5,343	7,094
Distance between testing times	0.24	0.11	0.03	0.46	1.00	3,956	6,569
Age	0.27	0.12	0.04	0.51	1.00	3,368	6,097
Accuracy at T1/T2	0.68	0.14	0.42	0.96	1.00	8,522	7,964
Research Team	2.00	0.62	0.81	3.22	1.00	4,119	5,516
Distance:Accuracy	-0.09	0.12	-0.34	0.13	1.00	9,482	8,671

Note. CI = Credible Interval; Rhat = Convergence Diagnostic; Bulk ESS = Effective Sample Size (bulk); Tail ESS = Effective Sample Size (tail). Random effects were estimated for nonword (n=31), and participant (n=147).

RQ5 & RQ6: Clinical status

In further analyses we evaluated the clinical potential of CL-NWR by looking at differences between the repetition accuracy of TD children versus children with language concerns (LC) and children with DLD, adding factors identified as robust in RQ1 (age) and RQ3 (length) and their interactions. In the analysis focusing on children with DLD (557 TD children, 284 DLD children, spread across 9 research teams), we found that children with DLD scored substantially below TD children (see Table 8). Additionally, there was weak evidence of an interaction with length, suggesting that children with DLD were finding longer nonwords disproportionately more difficult than TD children (see Figure 1).

Figure 1. Interaction of nonword length and clinical status.



Note: Length is presented in terms of number of phonemes (rather than syllables, each of which contained CV, i.e., two phonemes)

Table 8. Model estimating the effect of Age, Length, Clinical status (TD/DLD) and their interactions.

Predictor	Estimate	Standard Error	95% CI (lower)	95% CI (upper)	Rhat	Bulk ESS	Tail ESS
Intercept	0.82	0.30	0.22	1.40	1.00	2,937	5,443
Age	0.49	0.07	0.36	0.63	1.00	6,048	6,540
Length	-1.29	0.20	-1.67	-0.87	1.00	3,707	6,108
TD DLD	-1.24	0.39	-2.02	-0.44	1.00	6,455	7,146
Age:TD DLD	0.11	0.10	-0.12	0.30	1.00	7,697	7,264
Length:TD DLD	-0.32	0.15	-0.61	0.00	1.00	7,708	7,984

Note. CI = Credible Interval; Rhat = Convergence Diagnostic; Bulk ESS = Effective Sample Size (bulk); Tail ESS = Effective Sample Size (tail). Random effects were estimated for research team (n=9), nonword (n=67), and participant (n=841).

The second analysis compared TD children with those identified as having LC (879 TD children, 148 LC children, spread across 9 research groups). It revealed that children flagged by teachers, parents, or speech and language therapists as potentially having language difficulties had lower scores on the NWR test (see Table 9). However, unlike children with DLD, those with LC did not show disproportionate effects on longer items. The results were also indicative of an interaction between age and clinical group (TD/LC), with the effect of age being even more pronounced in children with LC. The cross-sectional data suggest that the gap between the groups was larger at younger ages and reduced as age increased (See Figure 2).

Figure 2. Interaction between age and clinical group (TD/LC).

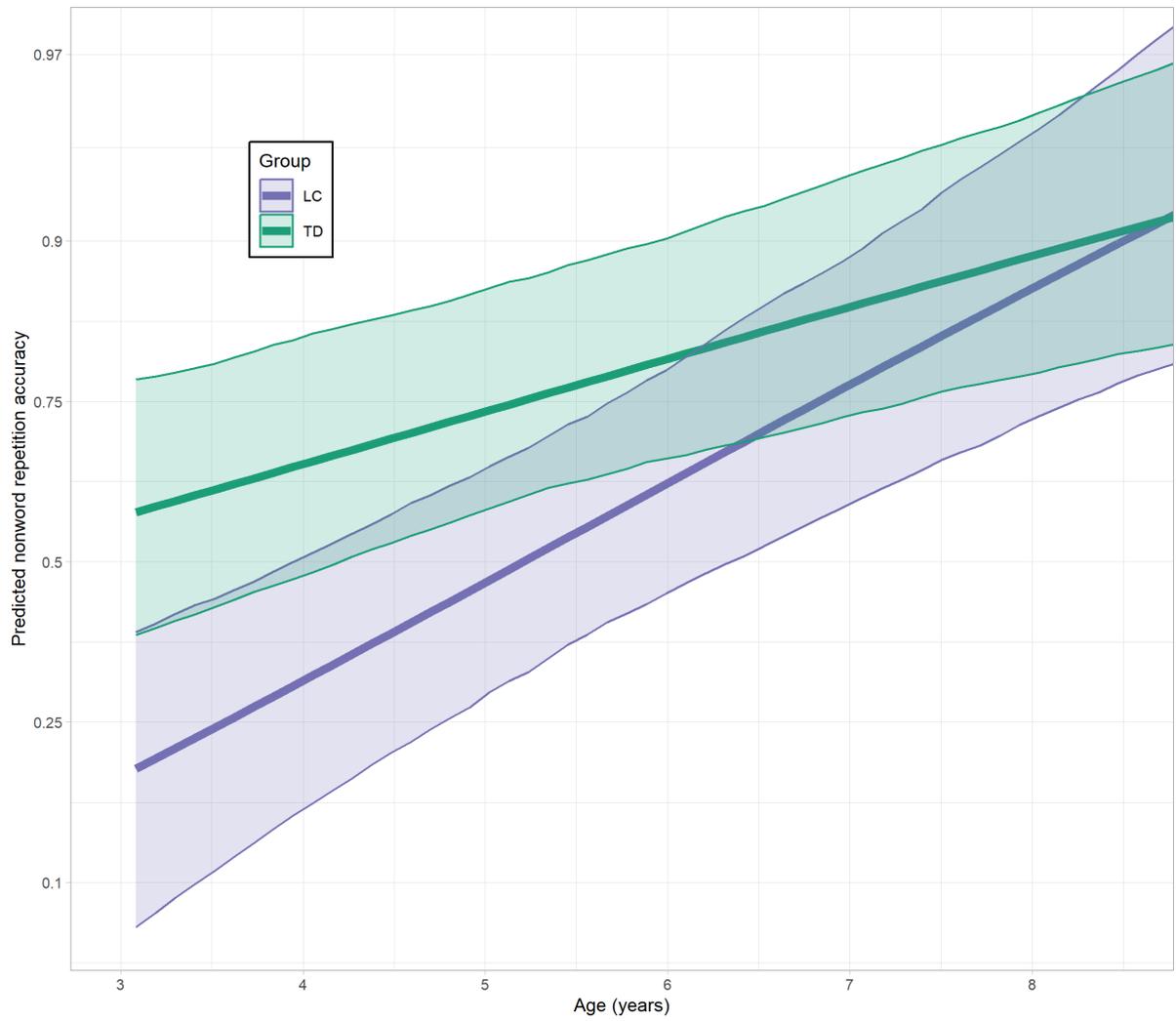


Table 9. Model estimating the effect of Age, Length, Clinical status (TD/LC) and their interactions.

Predictor	Estimate	Standard Error	95% CI (lower)	95% CI (upper)	Rhat	Bulk ESS	Tail ESS
Intercept	0.94	0.37	0.21	1.67	1.00	2,051	4,086
Age	0.78	0.13	0.51	1.03	1.00	3,918	5,882
Length	-1.27	0.21	-1.67	-0.84	1.00	3,039	4,501
TD LC	-0.92	0.30	-1.57	-0.35	1.00	4,149	5,628
Age:TD LC	0.48	0.22	0.06	0.94	1.00	4,790	5,837
Length:TD LC	0.09	0.24	-0.39	0.57	1.00	5,574	6,867

Note. CI = Credible Interval; Rhat = Convergence Diagnostic; Bulk ESS = Effective Sample Size (bulk); Tail ESS = Effective Sample Size (tail). Random effects were estimated for the research team (n=9), nonword (n=63), and participant (n=1,027).

RQ7: Classification accuracy by age bands

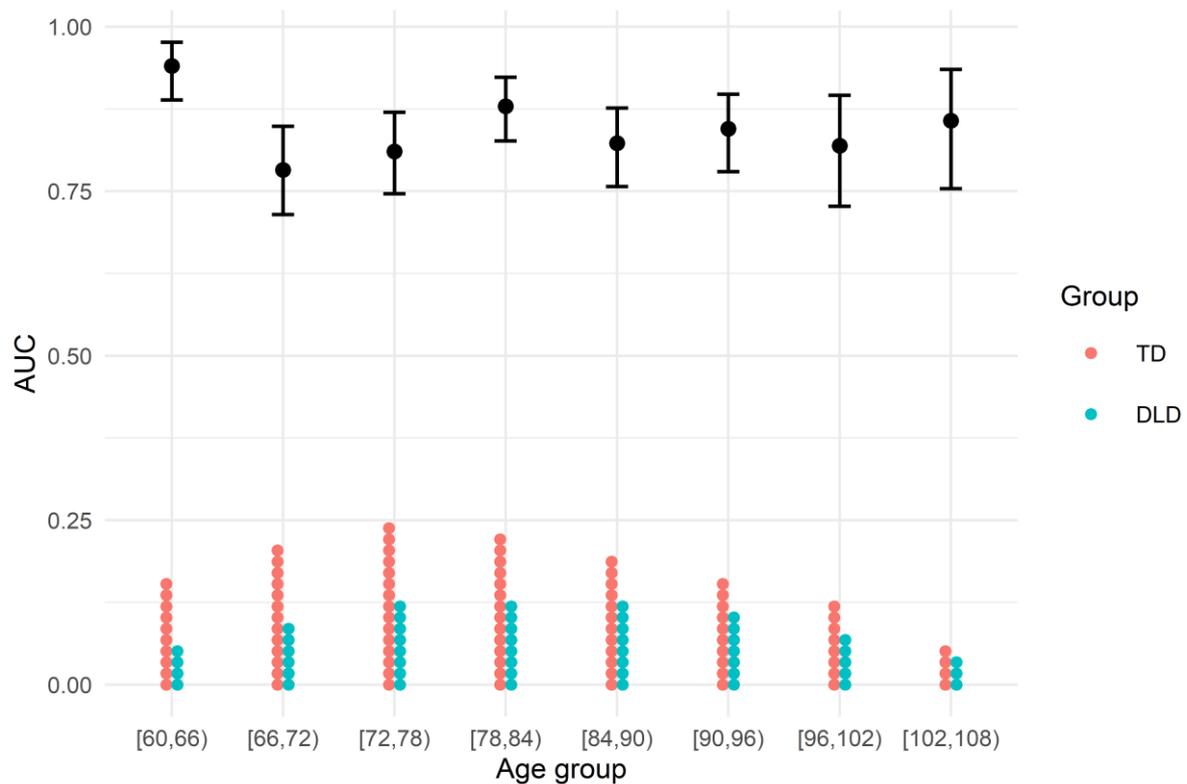
Table 10 presents threshold scores, sensitivity, and specificity values for distinguishing children with DLD from TD peers using CL-NWR performance, based on age-specific cut-offs applied across 6-month bands from 5;0 to 8;11 years. Analyses were restricted to age bands that included at least 20 children in both the DLD and TD groups. Sensitivity remained consistently high (≥ 0.75), while specificity ranged from 0.62 to 0.89, with lower values observed in the younger groups. Optimum threshold scores increased with age, reflecting developmental gains in nonword repetition performance. Figure 3 displays the corresponding area under the curve (AUC) values by age bands. As a general guideline, AUC above 0.90 is considered to reflect high classification accuracy, values between 0.70 and 0.90 indicate moderate accuracy, values from 0.50 to 0.70 suggest low accuracy, and an AUC of 0.50 reflects chance-level performance (e.g., Fisher et al., 2003). All AUC estimates exceeded 0.75, with most above 0.80, indicating good overall classification accuracy of the test. These results show that CL-NWR is a sensitive and specific test for DLD across all tested age groups, when used with age-adjusted cut-offs.

Table 10. Sensitivity and Specificity of DLD Classification by Age Band.

Age Band in Months	Threshold Score	Sensitivity	Specificity
60-65	7	0.89	0.89
66-71	7	0.84	0.62
72-77	7	0.86	0.65
78-83	8	0.82	0.81
84-89	8	0.8	0.75
90-95	9	0.75	0.81
96-101	10	0.76	0.78
102-107	10	0.81	0.76

Note. Optimal threshold scores and their associated sensitivity and specificity are reported for each 6-month age band. Sensitivity represents the true positive rate (correct identification of children with DLD), and specificity reflects the true negative rate (correct classification of TD children). The threshold indicates the score below (and including) at which a child would be considered at risk (e.g., a threshold of 7 means that scores from 0 to 7 indicate DLD risk).

Figure 3. Classifier AUC scores for distinguishing TD children from those with DLD based on CL-NWR performance, plotted by age groups.



Note. Higher AUC values indicate better classification accuracy. Lines represent 95% bootstrapped confidence intervals. Dots at the bottom indicate the number of children in each age group (1 dot = 10 children).

Discussion

This study investigated the effects of participant- and item-related factors on repetition accuracy in the Crosslinguistic Nonword Repetition (CL-NWR) task, a tool designed to assess language-related abilities of linguistically diverse children. Uniquely, it utilised a large, multi-sample dataset collected by 18 research teams across 15 countries. To evaluate the CL-NWR's validity crosslinguistically and across different social and geographical contexts, the study investigated (1) the effects of factors that are potential confounds with children's abilities, including SES, bilingualism, amount of exposure to the language of testing, and test version (referred to as *non-desirable factors* in this study); and (2)

the effects of factors that demonstrate the task is informative about children's abilities and their development, including age, item length, predictiveness of performance across time and, most importantly, clinical status (referred to as *desirable factors* in this study).

Non-desirable factors

Our findings indicated that the CL-NWR was not affected by the 'non-desirable' environmental factors we investigated, indicating that it is free of biases that are observed in other types of language assessments and limit the interpretation of test results when used with certain populations. Maternal education, used as a proxy for SES, did not contribute to CL-NWR accuracy, aligning with the lack of SES effects reported in other studies of nonword repetition (e.g., Farabolini et al., 2021; Meir & Armon-Lotem, 2017; Polišenská & Kapalková, 2014; Sundström et al., 2018). Nor was accuracy on the CL-NWR affected by the amount of exposure to the language of testing, or whether a child was exposed to one or more languages. This finding contrasts with the significant effects of bilingualism reported in some studies of language-specific nonword repetition tests (e.g., Cockcroft, 2016; Engel de Abreu, 2011; Engel de Abreu et al., 2013; Messer et al., 2010; Kohnert, Windsor & Yim, 2006; Windsor et al., 2010), though is consistent with previous findings on quasi-universal tests including but not limited to the CL-NWR (Chiat & Polišenská, 2016; de Almeida et al., 2017; Grimm, 2022). The negligible effects of demographic and environmental factors such as SES, exposure, and bilingualism on the CL-NWR support its potential to distinguish limited nonword repetition ability from limited experience of the language of testing. This is particularly beneficial because it ensures that the assessment probes targeted abilities independently of environmental factors, contributing a useful tool for equitable clinical practice in diverse populations.

Additionally, we found little evidence that CL-NWR performance was affected when different test versions were administered to the same children (i.e. Dutch/British English in the UK sample; Swedish/Arabic in the Swedish sample). Both models addressing this issue found length to be the only robust predictor. While the model of the Swedish data was consistent with weak differences in

scores for the Swedish and Arabic versions of the test, the generally comparable results for both pairs of tests support the crosslinguistic validity of the CL-NWR, implying that, in the absence of a version created for a child's own population, another available version would produce reliable results. However, this implication should be treated with caution since it is based on just two pairs of tests, with each pair administered to just two of our contributing samples (total n=205 children) (see section on limitations and future research).

Desirable factors

The robust effect of age replicates previous research and indicates that skills evaluated by the CL-NWR are still developing over the age range of the combined samples (37-165 months). This is also evident in the analysis of longitudinal data collected at three time points by teams in the Netherlands and South Africa. The finding that variance at T3 was predicted by scores at T1 and T2 indicates that the CL-NWR measures a stable underlying construct that persists across time. It also suggests that the CL-NWR captures children's developmental trajectories, such that a child's accuracy at a particular age is more precisely predicted if their accuracy at earlier age(s) is known.

Like age, length was a predictor found to affect CL-NWR accuracy across our multiple analyses. Replicating the length effect seen in established tests provides evidence of validity. Our data showed that while the effect of age and the effect of length were both robust, they were also additive and did not interact in the typical sample.

Given the purpose of the CL-NWR, evaluating the effect of identified language difficulties on accuracy is key. Our study set out to identify the clinical potential of CL-NWR in two clinical groups. The LC group referred to children about whom concerns have been expressed but who did not have a diagnosis of DLD; it also included those who had a clinical diagnosis of DLD, but when assessed independently, did not meet criteria for DLD. The 'DLD group' designated children who had a clinical diagnosis of DLD, and those who met test criteria for DLD with or without a clinical diagnosis. Our findings demonstrated the potential of the CL-NWR to distinguish between TD and clinical groups,

with TD children outperforming those from the clinical groups (TD > DLD, TD > LC). As in previous analyses, repetition accuracy was sensitive to age and nonword length.

The classification analysis demonstrated that the CL-NWR has strong potential as a screening tool for identifying children at risk of DLD. As shown in Table 10, sensitivity remained consistently high across all age bands. This indicates low risk of missing DLD cases. Specificity was more variable, particularly in younger children, suggesting a greater likelihood of false positives during earlier stages of development. Optimal threshold scores increased with age, reflecting expected developmental improvements in CL-NWR performance. These findings are further supported by the AUC values presented in Figure 3, where all age bands yielded values above 0.75, and most exceeded 0.80, indicating good discriminatory power when age-specific cut-offs were applied.

In screening contexts, higher sensitivity is especially important, as it ensures that most children with DLD are identified for further assessment, even at the cost of some over-identification. This is particularly critical given the developmental risks associated with missed cases. Although classification performance varied slightly across age bands, the overall pattern reinforces the potential of the CL-NWR as a population-level screener. However, its application as a standalone diagnostic tool remains limited and should be supplemented with broader clinical evaluation. The nature of the interactions varied across the LC and DLD samples. In the TD/DLD sample, there was no evidence of an interaction between age and DLD, indicating that age affected both groups similarly. In contrast, the TD/LC sample showed weak evidence for an interaction, with the effect of age being more pronounced in LC children. Specifically, there were larger differences between TD and LC children at younger ages, which diminished in the older children. The interaction between clinical group (DLD or LC) and nonword length also revealed opposite patterns suggesting that the LC group may be less vulnerable. Children with DLD had disproportionately lower accuracy as nonword length increased, replicating previous reports (Ahufinger et al., 2021; Boerma et al., 2015; Dispaldro et al., 2013; Graf Estes et al., 2007). However, no such pattern was observed in the LC group, indicating that nonword length did not impact their accuracy in the same way and that they behaved more similarly

to the TD group in this respect. The narrowing gap between the LC and TD groups, together with the similar effects of length on accuracy, suggest that difficulties of children who qualify as LC (as defined in this study) may be transitory, perhaps because their difficulties involve aspects of language processing that are noticeable and concerning to parents but more peripheral and more likely to resolve (e.g., speech production difficulties) than those of the DLD group. Longitudinal evidence is needed to determine whether the rate of catch-up is higher in children with LC than those with a diagnosis, and if so, to investigate differences in their language profiles.

Strengths, limitations and future research

One of the key strengths of this study was the diversity and size of the sample. The dataset included an impressive 43,449 individual datapoints (i.e., scores for individual nonword items) equivalent to responses from just under 2,000 children, spanning 15 countries across four continents, with children being exposed to a wide range of languages. Importantly, the CL-NWR task demonstrated robustness, with accuracy unaffected by external demographic factors.

A further strength lies in our approach to analyses of the dataset. An advantage of using Bayesian analyses is their ability to interpret null effects, providing a probabilistic framework that quantifies evidence for the absence of an effect. This is particularly novel compared to previous individual studies using the CL-NWR, which often relied on smaller samples and may have struggled to draw meaningful conclusions about null results. Additionally, while some questions necessarily relied on smaller subsamples (e.g., 841 children for the DLD analysis; 1027 for the LC analysis; 147 for the predictiveness analysis), item-level analyses allowed us to work effectively with these smaller samples and handle missing data without compromising the reliability of the results.

Nonetheless, some limitations of our sample should be acknowledged. While this was very large and spanned children from a wide range of geographical, cultural and demographic backgrounds, it is likely that children from extremely disadvantaged backgrounds were not represented. Recruiting these children is a common and ongoing challenge in the field. Including

such groups could provide valuable insights into how extreme disadvantage might influence CL-NWR performance, further enhancing our understanding of its applicability.

While the sensitivity and specificity values observed in this study were generally acceptable, analyses were limited to age bands that included at least 20 children in both the DLD and TD groups. As a result, the findings may not generalize to age groups with lower representation. Future research should aim to broaden the age range examined. It would also be valuable to explore sensitivity and specificity by bilingual status. Although bilingualism showed only weak effects in this study and is unlikely to meaningfully influence CL-NWR performance at group level, it remains important to evaluate its potential impact in clinical settings at an individual level. Large-scale studies could extend this work by systematically examining classification accuracy across both age and language background to further refine the tool's applicability and diagnostic utility.

The designation of the DLD sample in our analysis was based on classifications provided by contributing studies, which varied in their criteria. Some studies relied on clinical diagnoses, while others used formal assessments and applied cut-offs that were not specified for the purposes of our study. While this introduces a degree of heterogeneity, such variability is more the norm than the exception in clinical practice, both within countries and even more so in international contexts. As such, our sample may be seen as representative of the diversity encountered in clinical settings and across research studies. Despite this variation in DLD identification, our analysis revealed a substantial effect of DLD on CL-NWR performance. This consistency across differing diagnostic practices suggests a degree of robustness and strengthens the case for the clinical utility of the task. Replication using more standardized and explicitly defined DLD criteria would nonetheless be a valuable next step to further validate these findings. In particular, the respectable levels of sensitivity and specificity found in the 4-6-year-old subsample are promising and warrant further investigation.

An additional consideration for future research is the need for greater consistency in scoring conventions across sites. Although we accounted for between-team variation by including research team as a random effect, small differences in scoring criteria (e.g., allowing for additions) may still

introduce noise. Promoting standardised guidelines and scoring procedures will be important for enhancing comparability of CL-NWR results in cross-site applications.

While our finding of similar performance on different test versions supports their crosslinguistic validity, as pointed out above, this was based on just two samples and two pairs of tests. Recall that the difference between test versions lies in the selection of targets from available options and in the phonetic realization of their constituent consonants and vowels. Comparison of more test versions in more diverse samples is therefore needed before conclusions can be drawn about the effects of these factors. However, test versions also varied in the quality of speech and recordings, potentially confounding effects of differences in target items or their phonetic realization. To eliminate such extraneous factors, we have developed a single set of CL-NWR materials for universal use. These include recordings of all CL-NWR items, produced by a phonetician using vowel and consonant realizations designed to be as neutral across languages as possible (Chiat et al., 2020). All test materials, including the PowerPoint presentation with language-neutral recordings of test items, instructions for administration and scoring, and guidelines for interpretation, are available on request from the first author.

These standardized materials afford greater consistency in test administration across studies and settings. In so doing, they lay foundations for addressing more ambitious questions regarding the crosslinguistic validity of the CL-NWR: whether performance is affected by country of testing, ambient language(s), or phonological typology of ambient language(s), and hence whether universal norms are justified or if norms need to be derived for particular communities.

Conclusion

Nonword repetition tasks hold a unique advantage in the assessment of children's language-related abilities due to their minimal reliance on prior language knowledge. The Cross-Linguistic Nonword Repetition (CL-NWR) task builds on this strength by offering test items specifically designed to accommodate a wide range of lexical phonologies. An analysis of data from 18 CL-NWR studies

conducted across 15 countries revealed that repetition accuracy was influenced by age, item length, and clinical status, indicating that it accesses language-related abilities and their development. Conversely, accuracy was unaffected by gender, bilingualism, amount of exposure, or maternal education (used as a proxy for SES). This highlights its potential advantage over language-specific tests, which can inadvertently disadvantage children with limited exposure to the test language and whose performance might overlap with that of clinical groups.

Based on findings from this multilingual study, we suggest that the CL-NWR shows promise as a tool for assessing children irrespective of their linguistic background, may be useful in contexts where language-specific assessments and assessors fluent in the child's language are unavailable, and can serve as an indicator of potential risk for language difficulties including DLD. While, like other nonword repetition tasks, it is insufficient to definitively confirm or rule out language disorders, it offers valuable insights into children's language-related abilities and can guide further assessment.

According to Bao et al. (2024), effective screening tools are crucial for the timely identification of children at risk of DLD. However, screeners that are not designed for linguistically diverse or bilingual children carry an increased risk of over- or under-identification in these populations. A nonword repetition task designed to be crosslinguistic and dialect-neutral offers a promising solution. By targeting core linguistic processing abilities rather than language-specific features, such a tool reduces bias and may thereby improve accuracy in identifying children who require further assessment and support.

Acknowledgements

This study is a result of collaboration that took place within the COST Action IS0804 ‘Language Impairment in a Multilingual Society: Linguistics Patterns and the Road to Assessment’ (www.bili.org). We would like to express our gratitude to all children who participated in the study as well as their parents/caregivers, teachers and clinicians. Our appreciation for help in data collection/acquisition goes to Nikoleta Barnová, Solveig Chilla, Yvonne Fitzmaurice, Rima Haddad, Cornelia Hamann, Daniel Holzinger, Monika Janíková, Shiyun Mao, Mary Pat O’Malley, Ora Oudgenoeg-Paz, Alexandra Polatidou, and Rianne van den Berghe.

Data Availability Statement

The data used in this study is available on the Open Science Framework (OSF) at https://osf.io/2wu8r/?view_only=1e31708359f545d6b53c2da6ca727eac

References

- Ahufinger, N., Berglund-Barraza, A., Cruz-Santos, A., Ferinu, L., Andreu, L., Sanz-Torrent, M., & Evans, J. L. (2021). Consistency of a nonword repetition task to discriminate children with and without developmental language disorder in Catalan–Spanish and European Portuguese speaking children. *Children*, 8(2), 85. doi.org/10.3390/children8020085
- Bao, X., Komesidou, R., & Hogan, T. P. (2024). A review of screeners to identify risk of Developmental Language Disorder. *American Journal of Speech-Language Pathology*, 33(3), 1548-1571. doi.org/10.1044/2023_AJSLP-23-0028
- Baum, S., & Titone, D. (2014). Moving toward a neuroplasticity view of bilingualism, executive control, and aging. *Applied Psycholinguistics*, 35(5), 857-894. doi.org/10.1017/S0142716414000174
- Bishop, D. V., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & Klee, T. M. (2017). CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development. Phase 2. *Journal of Child Psychology and Psychiatry*, 58(10), 1068-1080. doi.org/10.1111/jcpp.12721
- Boerma, T. D., & Blom, W. B. T. (2021). Crosslinguistic nonword repetition and narrative performance over time: A longitudinal study on 5- to 8-year-old children with diverse language skills. In S. Armon-Lotem, & K. Grohmann (Eds.), *Language Impairment in Multilingual Settings: LITMUS in action across Europe* (pp. 302-328). John Benjamins.
- 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. doi.org/10.1044/2015_JSLHR-L-15-0058
- Burke, H. L., & Coady, J. A. (2015). Nonword repetition errors of children with and without specific language impairments (SLI). *International Journal of Language & Communication Disorders*, 50(3), 337-346. doi.org/10.1111/1460-6984.12136

- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1), 1–28. doi.org/10.18637/jss.v080.i01
- Chiat, S. (2015). Nonword Repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 125–150). Bristol: Multilingual Matters.
- Chiat, S., & Polišenská, K. (2016). A framework for crosslinguistic nonword repetition tests: Effects of bilingualism and socioeconomic status on children's performance. *Journal of Speech, Language, and Hearing Research*, 59(5), 1179-1189. doi.org/10.1044/2016_JSLHR-L-15-02
- Chiat, S., Polišenská, K., Yanushevskaya, I., Antonijevic, S. (2020). Crosslinguistic Nonword Repetition Test: Language-Neutral Version. Available from corresponding author.
- Chiat, S., & Roy, P. (2013). Early predictors of language and social communication impairments at ages 9–11 years: A follow-up study of early-referred children. *Journal of Speech, Language, and Hearing Research*, 56 (6), 1824-1836. [doi.org/10.1044/1092-4388\(2013/12-0249\)](https://doi.org/10.1044/1092-4388(2013/12-0249))
- Cockcroft, K. (2016). A comparison between verbal working memory and vocabulary in bilingual and monolingual South African school beginners: Implications for bilingual language assessment. *International Journal of Bilingual Education and Bilingualism*, 19(1), 74-88. doi.org/10.1080/13670050.2014.964172
- 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. doi.org/10.1075/lab.15019.alm
- de Bruin, A. (2019). Not all bilinguals are the same: A call for more detailed assessments and descriptions of bilingual experiences. *Behavioral Sciences*, 9(3), 33. doi.org/10.3390/bs9030033
- 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. [doi.org/10.1044/1092-4388\(2012/11-0304](https://doi.org/10.1044/1092-4388(2012/11-0304)
- Dos Santos, C., & Ferré, S. (2018). A nonword repetition task to assess bilingual children's phonology. *Language Acquisition*, 25(1), 58-71. doi.org/10.1080/10489223.2016.1243692
- Duncan, T. S., & Paradis, J. (2016). English language learners' nonword repetition performance: The influence of age, L2 vocabulary size, length of L2 exposure, and L1 phonology. *Journal of Speech, Language, and Hearing Research*, 59(1), 39-48. doi.org/10.1044/2015_JSLHR-L-14-0020
- Edwards, J., & Lahey, M. (1998). Nonword repetitions of children with specific language impairment: Exploration of some explanations for their inaccuracies. *Applied Psycholinguistics*, 19(2), 279-309. doi.org/10.1017/S0142716400010079
- Engel de Abreu, P. M. (2011). Working memory in multilingual children: Is there a bilingual effect? *Memory*, 19(5), 529-537. doi:[10.1080/09658211.2011.590504](https://doi.org/10.1080/09658211.2011.590504)
- Engel de Abreu, P. M., Baldassi, M., Puglisi, M. L., & Befi-Lopes, D. M. (2013). Cross-linguistic and cross-cultural effects on verbal working memory and vocabulary: Testing language-minority children with an immigrant background. *Journal of Speech, Language, and Hearing Research*, 56(2), 630-42. doi: [10.1044/1092-4388\(2012/12-0079\)](https://doi.org/10.1044/1092-4388(2012/12-0079)
- Farabolini, G., Rinaldi, P., Caselli, M. C., & Cristia, A. (2021). Non-word repetition in bilingual children: the role of language exposure, vocabulary scores and environmental factors. *Speech, Language and Hearing*, 25(3), 283–298. doi.org/10.1080/2050571X.2021.1879609
- Farmani, H., Sayyahi, F., Soleymani, Z., Labbaf, F.Z., Talebi, E., Shourvazi, Z. (2018). Normalization of the non-word repetition test in Farsi-speaking children. *Journal of Modern Rehabilitation*, 12, 217–224. doi: [10.32598/JMR.V12.N4.217](https://doi.org/10.32598/JMR.V12.N4.217)
- Fischer, J.E., Bachmann, L.M. & Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Medicine*, 29, 1043–1051. <https://doi.org/10.1007/s00134-003-1761-8>

- Fu, N. C., Chen, S., Polišenská, K., Chan, A., Kan, R., & Chiat, S. (2024). Nonword repetition in children with Developmental Language Disorder: Revisiting the case of Cantonese. *Journal of Speech, Language, and Hearing Research*, 67(6), 1772-1784. doi.org/10.1044/2024_JSLHR-22-00397
- Fu, N. C., Chan, A., Chen, S., Polišenská, K., & Chiat, S. (2024). Revisiting nonword repetition as a clinical marker of developmental language disorder: Evidence from monolingual and bilingual L2 Cantonese. *Brain and Language*, 257, 105450. doi: [10.1016/j.bandl.2024.105450](https://doi.org/10.1016/j.bandl.2024.105450)
- 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. doi.org/10.1080/09658219408258940
- Graf Estes, K., Evans, J. and Else-Quest, N. M. (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, 177-195. doi:[10.1044/1092-4388\(2007/015\)](https://doi.org/10.1044/1092-4388(2007/015))
- Grimm, A. (2022). The use of the LITMUS quasi-universal nonword repetition task to identify DLD in monolingual and early second language learners aged 8 to 10. *Languages*, 7(3), 218. doi.org/10.3390/languages7030218
- Guiberson, M., & Rodríguez, B. L. (2016). Nonword repetition in Spanish-speaking toddlers with and without early language delays. *Folia Phoniatrica et Logopaedica*, 67(5), 253-258. doi.org/10.1159/000442745
- Haman, E., Wodniecka, Z., Marecka, M., Szewczyk, J., Białecka-Pikul, M., Otwinowska, A., Mieszkowska, K., Łuniewska, M., Kořak, J., Miękiř, A., Kacprzak, A., Banasik, N., & Foryś-Nogala, M. (2017). How does L1 and L2 exposure impact L1 performance in bilingual children? Evidence from Polish-English migrants to the United Kingdom. *Frontiers in Psychology*, 8, 1444. [doi: 10.3389/fpsyg.2017.01444](https://doi.org/10.3389/fpsyg.2017.01444)
- Hamdani, S., Chan, A., Kan, R., Chiat, S., Gagarina, N., Haman, E., ... & Armon-Lotem, S. (2025). Identifying developmental language disorder (DLD) in multilingual children: A case study

- tutorial. *International Journal of Speech-Language Pathology*, 1-15.
doi.org/10.1080/17549507.2024.2326095
- Kalnak, N., Peyrard-Janvid, M., Forssberg, H., & Sahlén, B. (2014). Nonword repetition—a clinical marker for specific language impairment in Swedish associated with parents' language-related problems. *PloS one*, 9(2), e89544. doi: [10.1371/journal.pone.0089544](https://doi.org/10.1371/journal.pone.0089544)
- Kaščelan, D., Prévost, P., Serratrice, L., Tuller, L., Unsworth, S., & De Cat, C. (2022). A review of questionnaires quantifying bilingual experience in children: Do they document the same constructs? *Bilingualism: Language and Cognition*, 25(1), 29-41.
doi.org/10.1017/S1366728921001152
- 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. doi.org/10.1111/j.1540-5826.2006.00204.x
- Law, J., Tulip, J., & Beckermann, E. (2019). The development of the practitioner survey. In J. Law, C. McKean, C-A. Murphy, & E. Thordardottir (Eds.), *Managing Children with Developmental Language Disorder: Theory and Practice Across Europe and Beyond* (pp. 30-55). London: Routledge.
- Law, J., McBean, K., & Rush, R. (2011). Communication skills in a population of primary school-aged children raised in an area of pronounced social disadvantage. *International Journal of Language & Communication Disorders*, 46(6), 657–664. doi: [10.1111/j.1460-6984.2011.00036.x](https://doi.org/10.1111/j.1460-6984.2011.00036.x)
- Lee, H. J., Kim, Y. T., & Yim, D. (2013). Non-word repetition performance in Korean-English bilingual children. *International Journal of Speech-Language Pathology*, 15(4), 375-382.
doi.org/10.3109/17549507.2012.752866
- Lee, S. A. S., & Gorman, B. K. (2013). Nonword repetition performance and related factors in children representing four linguistic groups. *International Journal of Bilingualism*, 17(4), 479-495.
doi.org/10.1177/136700691243830

- Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology*, 25(5), 605–621.
doi.org/10.1080/20445911.2013.795574
- Meir, N., & Armon-Lotem, S. (2017). Independent and combined effects of socioeconomic status (SES) and bilingualism on children’s vocabulary and verbal short-term memory. *Frontiers in Psychology*, 8, 1442. doi.org/10.3389/fpsyg.2017.01442
- Melby-Lervåg, M., Lervåg, A., Lyster, S.-A. H., Klem, M., Hagtvet, B., & Hulme, C. (2012). Nonword-repetition ability does not appear to be a causal influence on children’s vocabulary development. *Psychological Science*, 23(10), 1092-1098. doi.org/10.1177/0956797612443833
- Messer, M. H., Leleman, P. P., Boom, J., & Mayo, A. Y. (2010). Phonotactic probability effect in nonword recall and its relationship with vocabulary in monolingual and bilingual preschoolers. *Journal of Experimental Child Psychology*, 105(4), 306-323.
doi.org/10.1016/j.jecp.2009.12.006
- Næss, K. A. B., Lervåg, A., Lyster, S. A. H., & Hulme, C. (2015). Longitudinal relationships between language and verbal short-term memory skills in children with Down syndrome. *Journal of Experimental Child Psychology*, 135, 43-55. doi.org/10.1016/j.jecp.2015.02.004
- 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. *Journal of Child Psychology and Psychiatry*, 57(11), 1247-1257. doi.org/10.1111/jcpp.12573
- Öberg, L. (2020). Words and nonwords: Vocabulary and phonological working memory in Arabic-Swedish-speaking 4–7 -year-olds with and without a diagnosis of Developmental Language Disorder. [Doctoral thesis, Uppsala University].
<https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-421590>

- Öberg, L., & Bohnacker, U. (2022). Non-word repetition and vocabulary in Arabic-Swedish-speaking 4–7-year-olds with and without Developmental Language Disorder. *Languages*, 7(3), 204. doi.org/10.3390/languages7030204
- Öberg, L., & Bohnacker, U. (2024). Beyond language scores: How language exposure informs assessment of nonword repetition, vocabulary and narrative macrostructure in bilingual Turkish/Swedish children with and without Developmental Language Disorder. *Children*, 11(6), 704. doi.org/10.3390/children11060704
- Ortiz, J. A. (2021). Using nonword repetition to identify language impairment in bilingual children: A meta-analysis of diagnostic accuracy. *American Journal of Speech-Language Pathology*, 30(5), 2275-2295. doi.org/10.1044/2021_AJSLP-20-00237
- Parra, M., Hoff, E., & Core, C. (2011). Relations among language exposure, phonological memory, and language development in Spanish–English bilingually developing 2-year-olds. *Journal of Experimental Child Psychology*, 108(1), 113-125. doi.org/10.1016/j.jecp.2010.07.011
- Polišenská, K., & Kapalková, S. (2014). Improving child compliance on a computer-administered nonword repetition task. *Journal of Speech, Language and Hearing Research*, 57 (3), 1060-1068. [doi.org/10.1044/1092-4388\(2013/13-0014](https://doi.org/10.1044/1092-4388(2013/13-0014)
- Scheidnes, M. (2020). Sentence repetition and non-word repetition in early total French immersion. *Applied Psycholinguistics*, 41(1), 107-131. doi.org/10.1017/S0142716419000420
- Schwob, S., Eddé, L., Jacquin, L., Leboulanger, M., Picard, M., Oliveira, P. R., & Skoruppa, K. (2021). Using nonword repetition to identify Developmental Language Disorder in monolingual and bilingual children: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, 64(9), 3578–3593. doi.org/10.1044/2021_JSLHR-20-00
- Stokes, S. F., Wong, A. M., 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. [doi.org/10.1044/1092-4388\(2006/019](https://doi.org/10.1044/1092-4388(2006/019)

Sundström, S., Löfkvist, U., Lyxell, B., & Samuelsson, C. (2018). Prosodic and segmental aspects of nonword repetition in 4-to 6-year-old children who are deaf and hard of hearing compared to controls with normal hearing. *Clinical Linguistics & Phonetics*, 32(10), 950-971.

doi.org/10.1080/02699206.2018.1469671

Thordardottir, E. (2014). The typical development of simultaneous bilinguals: Vocabulary, morphosyntax and language processing in two age groups of Montreal preschoolers. In T. Grüter & J. Paradis, J. (Eds.), *Input and experience in bilingual development* (pp. 141-160). John Benjamins Publishing Company.

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. doi.org/10.1016/j.jcomdis.2012.08.002

Tuller, L. (2015). Clinical use of parental questionnaires in multilingual contexts. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 299–328). Bristol: Multilingual Matters.

UNESCO Institute for Statistics. (2011). *International Standard Classification of Education ISCED 2011*. Available at: <https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf>

White, M. J. (2021). Phonological working memory and non-verbal complex working memory as predictors of future English outcomes in young ELLs. *International Journal of Bilingualism*, 25(1), 318-337. doi.org/10.1177/1367006920948136

Windsor, J., Kohnert, K., Lobitz, K. F., & Pham, G. T. (2010). Cross-language nonword repetition by bilingual and monolingual children. *American Journal of Speech-Language Pathology*, 19 (4), 298-310. [doi.org/10.1044/1058-0360\(2010/09-0064\)](https://doi.org/10.1044/1058-0360(2010/09-0064))

Appendix A. Table of intermediate model of demographic factors.

Predictor	Estimate	Standard Error	95% CI (lower)	95% CI (upper)	Rhat	Bulk ESS	Tail ESS
Intercept	1.42	0.31	0.80	2.04	1.00	2,440	5,615
Age	0.40	0.07	0.28	0.55	1.00	7,817	8,451
Bilingual status	-0.11	0.27	-0.67	0.40	1.00	7,895	9,811
Maternal education	0.14	0.07	0.01	0.27	1.00	9,679	11,102
Amount of exposure	0.14	0.17	-0.20	0.47	1.00	8,848	11,355
Gender	0.29	0.12	0.05	0.53	1.00	10,081	11,516

Note. CI = Credible Interval; Rhat = Convergence Diagnostic; Bulk ESS = Effective Sample Size (bulk);

Tail ESS = Effective Sample Size (tail). Random effects were estimated for research team (n=12),

nonword (n=69), and participant (n=767).

Appendix B. Table of intermediate model of demographic factors without amount of exposure.

Predictor	Estimate	Standard Error	95% CI (lower)	95% CI (upper)	Rhat	Bulk ESS	Tail ESS
Intercept	1.59	0.31	0.97	2.21	1.00	2,909	6,523
Age	0.52	0.15	0.23	0.84	1.00	7,326	10,798
Bilingual status	-0.28	0.22	-0.75	0.13	1.00	9,924	10,361
Maternal education	0.11	0.08	-0.07	0.26	1.00	10,644	10,591
Gender	0.20	0.15	-0.10	0.49	1.00	11,237	11,794

Note. CI = Credible Interval; Rhat = Convergence Diagnostic; Bulk ESS = Effective Sample Size (bulk);

Tail ESS = Effective Sample Size (tail). Random effects were estimated for research team (n=12),

nonword (n=69), and participant (n=804).

Appendix C. Nonword items used across the research teams.

Team: Language version	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5	ITEM6	ITEM7	ITEM8	ITEM9	ITEM10	ITEM11	ITEM12	ITEM13	ITEM14	ITEM15	ITEM16
Austria	zipu	tula	naki	lumi	zipula	panuti	nalidu	luniga	zipalita	nugitala	gazulumi	litisagu	sipunakila	tulikazumu	maluzikupa	litapimudi
Canada-English	zibu	dula	nagi	lumi	sipula	bamudi	malitu	lumiga	zipalida	mukitala	kasulumi	lidisaku	sipumakila	duligasumu	maluziguba	litapimuti
Canada-French	zibu	dula	nagi	lumi	sipula	bamudi	malitu	lumiga	zipalida	mukitala	kasalumi	lidisaku	sipumakila	duligasumu	maluziguba	litapimuti
Finland	sipu	lita	maki	nuli	sipula	pamuti	nalitu	lunika	sipalita	nukitala	kasulumi	litisaku	sipumakila	tulikasumu	malusikupa	litapimuti
Germany	zipu	tula	naki	lumi	zipula	panuti	nalidu	luniga	zipalita	nugitala	gazulumi	litisagu	sipunakila	tulikazumu	maluzikupa	litapimudi
Greece	zibu	lida	nagi	luni	sipula	bamudi	nalitu	luniga	zibalita	mukidala	kasulumi	litzaku	sibumagila	tulikasumu	malusiguba	lidapimuti
Hong Kong	sibu	dula	magi	lumi	sipula	bamudi	malidu	lumiga	sipalida	mugidala	gasulumi	litisaku	sipumagila	duligasumu	malusikuba	lidabimudi
Ireland	zibu	dula	nagi	lumi	sipula	bamudi	malitu	lumiga	zipalida	mukitala	kasulumi	lidisaku	sipumakila	duligasumu	maluziguba	litapimuti
Malta Team 1	zibu	dula	nagi	muli	sipula	bamudi	malitu	lumiga	sipalita	mukitala	kasulumi	lidisaku	sipumakila	duligasumu	malusikupa	litapimuti
Malta Team 2	zibu	dula	nagi	lubi*	sipula	bamudi	malitu	lumiga	zipalida	mukitala	kasulumi	lidisaku	sipumakila	duligasumu	maluziguba	litapimuti
Netherlands Team 1	sibu	lita	naki	nuli	zibula	bamudi	nalidu	lumika	zibalita	nukitala	kazulumi	litisaku	sibunakila	tulikasumu	maluzikuba	lidabimudi
Netherlands Team 2	sibu	lita	naki	nuli	zibula	bamudi	nalidu	lumika	zibalita	nukitala	kazulumi	litisaku	sibunakila	tulikasumu	maluzikuba	lidabimudi
Singapore	sibu	dula	nagi	luni	sibula	panudi	malitu	lunika	sipalita	mugidala	kasulumi	litisaku	sipumakila	dulikasumu	malusikuba	lidabimudi
Slovakia	zibu	lita	nagi	luni	sipula	bamudi	malidu	lumiga	zipalita	mukitala	kasulumi	litisaku	sipumakila	tulikazumu	maluziguba	lidabimuti
South Africa	zibu	dula	nagi	lumi	sipula	bamudi	malitu	lumiga	zipalida	mukitala	kasulumi	lidisaku	sipumakila	duligasumu	maluziguba	litapimuti
Sweden-Arabich	zibu	lita	naki	muli	sibula	banudi	nalitu	limika*	sibalita	mukidala	kasulumi	lidizaku	sibunakila	dulikasumu	maluzikuba	lidabimudi
Sweden-Swedish	sibu	dula	nagi	luni	sipula	banudi	malitu	limika*	sibalita	mukidala	gasulumi	lidisaku	sipunakila	tuligasumu	malusiguba	lidapimuti
Switzerland	zibu	dula	nagi	lumi	sipula	bamudi	malitu	lumiga	zipalida	mukitala	kasulumi	lidisaku	sipumakila	duligasumu	maluziguba	litapimuti
UK England: Dutch	sibu	lita	naki	nuli	zibula	bamudi	nalidu	lumika	zibalita	nukitala	kazulumi	litisaku	sibunakila	tulikasumu	maluzikuba	lidabimudi
UK England: English	zibu	dula	nagi	lumi	sipula	bamudi	malitu	lumiga	zipalida	mukitala	kasulumi	lidisaku	sipumakila	duligasumu	maluziguba	litapimuti
UK Scotland: English	zibu	dula	nagi	lumi	sipula	bamudi	malitu	lumiga	zipalida	mukitala	kasulumi	lidisaku	sipumakila	duligasumu	maluziguba	litapimuti

Note. Items marked by * deviated from the Crosslinguistic nonword repetition framework (Chiat, 2015)