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

Linnéa Öberg and Ute Bohnacker *

Department of Linguistics & Philology, Uppsala University, P.O. Box 635, SE-75126 Uppsala, Sweden; linnea.oberg@lingfil.uu.se
* Correspondence: ute.bohnacker@lingfil.uu.se

Abstract: As in many other countries, baseline data concerning the linguistic development of bilingual children in Sweden are lacking, and suitable methods for identifying developmental language disorder (DLD) in bilinguals are lacking as well. This study presents reference data from 108 typically developing (TD) Turkish/Swedish-speaking children aged 4;0–8;1, for a range of language tasks developed specifically for the assessment of bilinguals (LITMUS test battery, COST Action IS0804). We report on different types of nonword repetition (NWR) tasks (language-specific and language-independent), receptive and expressive vocabulary (Cross-Linguistic Lexical Tasks, CLTs), and narrative macrostructure comprehension and production (Multilingual Assessment Instrument for Narratives, MAIN) in Turkish, the children’s home language, and in Swedish, the language of schooling and society. Performance was investigated in relation to age, language exposure, type of task, and (for NWR and narratives) vocabulary size. There was a positive development with age for all tasks, but effects of language exposure and vocabulary size differed between tasks. Six bilingual Turkish/Swedish children with DLD were individually compared to the TD children. TD/DLD performance overlapped substantially, particularly for NWR, and more so for the production than the comprehension tasks. Surprisingly, the discriminatory potential was poor for both language-specific and language-independent NWR. DLD case studies underscored the importance of interpreting language scores in relation to exposure history, and the need for an increased emphasis on functional language skills as reported by parents and teachers when assessing and diagnosing DLD in bilinguals.

Keywords: bilingual children; Developmental Language Disorder (DLD); language exposure; narrative macrostructure; nonword repetition; Swedish; Turkish; vocabulary size

1. Introduction

A large proportion of children in Sweden today are growing up multilingually, with a home language other than the language of society and schooling. A total of 28.3% of all children aged 7 to 15 in Sweden are entitled to heritage language education, which means that these children speak a home language other than, or in addition to, Swedish (National Agency for Education 2023 [1]), and in many urban areas, the proportion is more than twice as high. In such linguistically and culturally diverse settings, it is often a challenge to identify developmental language and learning disorders, as the children’s majority-language skills are still developing, and effective language assessment materials and methods for multilinguals are few and far between. Norm-referenced tests are mostly standardized on monolingual child populations, and in countries with a relatively small
population such as Sweden (10 million), proper norms may be lacking altogether; alternatively, existing tests frequently over-identify developmental language disorder (DLD)\(^1\) in multilinguals, which casts doubt on their clinical utility. For instance, Andersson et al. (2019 [2]) found that 80\% of the 224 bilingual 7-to-8-year-olds in their sample performed more than 1 SD below the normative monolingual mean on the Swedish version of a widely used clinical test, the Clinical Evaluation of Language Fundamentals (CELF-4).

In response to similar diagnostic test shortcomings also in other countries, and because of a general lack of suitable (and affordable) assessment materials, the research network COST Action IS0804 Language impairment in a multilingual society: Linguistic patterns and the road to assessment was founded (https://www.bi-sli.org/). From 2009 to 2013, a network of (mainly European) researchers in linguistics, speech-language therapy and developmental psychology collaborated in basic research and developed comparable language assessment materials and methods/tasks for many different languages, which expressly were not just translations of a test that had first been developed for one particular language (such as English). After the official end of the COST Action, many of the researchers continued to collaborate, and work with the Action materials has branched out into other parts of the world. Thus, a battery of tasks has been gathered under the umbrella of LITMUS (Language impairment testing in multilingual settings, Armon-Lotem et al. 2015 [3]). The Action and its spinoffs have been most productive, and new language versions of the LITMUS tasks are continually being developed. Since several tasks have been made freely available to other researchers and practitioners, their use has spiraled in many parts of the world.

The present paper investigates how bilingual Turkish/Swedish children with and without DLD fare on the LITMUS tasks in three domains: nonword repetition, cross-linguistic lexical tasks, and narrative macrostructure. They were developed with a view to provide less-biased methods of language assessment than the existing standardized language tests for bilingual children of a wide variety of language combinations. The rationale behind them can be captured as follows.

In nonword repetition (NWR), children are asked to repeat phonological forms that are not real words and that they have not heard before. NWR draws on phonological processing and cognitive-perceptual mechanisms.\(^2\) A deficit in verbal short-term memory is commonly associated with DLD (Gathercole & Baddeley 1990 [4]). Poor NWR performance is regarded as a clinical marker of DLD, as many studies have found that (monolingual) children with a DLD diagnosis show impaired NWR performance compared to TD peers (e.g., Chiat 2015 [5]) across languages, including for Turkish (Topbaş et al. 2014 [6]) and Swedish (Kalnak et al. 2014 [7]), the two languages of the present study. Impaired NWR performance has been documented for bilingual children with DLD as well, even though the evidence here is not as clear-cut as for monolinguals (Ortiz 2021 [8]; Schwob et al. 2021 [9]). There are no studies of NWR in a Swedish/Turkish context, and only one published study on NWR in bilinguals in Sweden (Öberg & Bohnacker 2022 [10], on Arabic/Swedish, see Section 2.1). Internationally, NWR has been promoted as a valid and diagnostically useful tool for the clinical assessment of bilingual children (Antonijevic-Elliott et al. 2020 [11]; Boerma et al. 2015 [12]; Boerma & Blom 2021 [13]; Chiat 2015 [5]; Paradis et al. 2021, p. 330 [14]; Sorenson Duncan & Paradis 2016 [15]; Thordardottir 2015, p. 348 [16]; Thordardottir & Brandeker 2013 [17]). Bilinguals are assumed to be able to transfer the phonological processing skills they have acquired in any language; thus, the performance on an NWR task should remain largely unaffected by limited exposure and/or language knowledge, especially when the nonword items are quasi-universal, i.e., (relatively) independent of language-specific traits. On such NWR tasks, TD bilinguals are predicted to be able to perform on a par with TD monolinguals, and better than DLD children. (For studies on this, see Section 2.1 below.) Moreover, since phonological processing skills are assumed to be transferrable between languages, it is often deemed sufficient to assess NWR in bilingual children in only one of their languages, especially when quasi-universal nonword items are used (Boerma & Blom 2017 [18], 2021 [13]; Dollaghan & Campbell 1998
The LITMUS battery includes several NWR tasks that are more or less language-(in)dependent, with items constructed according to certain principles (https://www.bi-sli.org/nonword-repetition, see also Sections 2.1 and 4.2.1).

Moving on from nonwords to real words (vocabulary), delayed and impaired lexical abilities are known to be one of the earliest indicators of DLD for monolingual children (e.g., Leonard 2014, pp. 53–57 [20]; Rice & Hoffman 2015 [21]; Trauner et al. 2000 [22]). This appears to hold for DLD in bilinguals as well; however, studies on the clinical utility of vocabulary assessments have come to divergent conclusions. Vocabulary development is heavily dependent on language input and exposure, and as the amount of exposure is necessarily divided between two languages in bilingual children, TD bilinguals may show lexical gaps and depressed vocabulary scores compared to monolingual peers (Oller & Eilers 2002 [23]; Thordardottir 2011 [24]). Bilinguals are frequently tested in one language only (typically in the majority or societal language, where vocabulary tests are available), even though this procedure does not fairly represent their full lexical competence. Comparable, parallel vocabulary tasks could be used to clinically assess the two languages of bilinguals, but for most language combinations such tasks do not (yet) exist. As part of the COST Action IS0804, cross-linguistic lexical tasks (CLTs) were developed using a multi-language parallel task-construction procedure (for details see Haman et al. 2015 [25], https://multilada.pl/en/projects/clt/), in order to provide fully comparable assessment tools for the comprehension and production of nouns and verbs in different languages and language combinations (see Section 2.2). The CLT developers assume that when lexical deficits are identified in both languages of a bilingual child, this is a warning sign for DLD (Haman et al. 2015, p. 198 [25]; Haman & Łuniewska 2013 [26]). There are, however, surprisingly few published studies that explore the diagnostic utility of CLT vocabulary assessment in bilingual children (see Section 2.2).

Narrative macrostructure is another domain that has been promoted as particularly suitable for the language assessment of bilingual children (Govindarajan & Paradis 2019 [27]; Paradis et al. 2013 [28], 2021, p. 330 [14]). Macrostructural skills include the ability to identify causal, hierarchical and thematic relations between events, resulting in an understanding of the overall structure of a story (so-called story grammar) and the ability to produce a coherent story (Stein & Glenn 1979 [29]). Children with DLD are frequently described as performing poorly on narrative tasks, across languages (Mäkinen 2014 [30]; Merritt & Liles 1987 [31]; Soodla & Kikas 2010 [32]). Of course, narrative skills involve many different aspects of language, some of which are highly language-specific (e.g., vocabulary, tense marking, morphosyntactic encoding of reference), and some may also be culture-specific. However, narrative macrostructure is generally regarded as being largely independent of the structural properties of a language and possibly shared across languages, due to its strong cognitive-general underpinnings (e.g., Berman 2001, pp. 420–426 [33]; Berman & Slobin 1994 [34]; Paradis et al. 2021, pp. 170–171 [14]; Pearson 2002 [35]). This would allow bilinguals to use the macrostructural skills they have acquired in any language and transfer them to another language when narrating or comprehending stories. TD bilinguals should thus be able to advance their macrostructural skills in a new language relatively fast, and tasks probing narrative macrostructure should be able to discriminate between bilingual TD and DLD. Only a few studies have evaluated the diagnostic validity of narrative tasks for children learning two or more languages (see Section 2.3). The Multilingual Assessment Instrument for Narratives (MAIN, Gagarina et al. 2019 [36], https://main.leibniz-zas.de/) assesses both the comprehension and production of macrostructure (see Section 2.3).

Whilst neither macrostructure nor NWR tasks are free of language, the rationale for promoting them as suitable assessment tools for bilinguals is that they have underlying cognitive-general components and tap into skills that are acquired in any language. This may reduce assessment bias against children with limited exposure and different language experiences.
The above LITMUS tasks were developed to support the identification of DLD in bilingual children. However, as bilingual language development is complex, varied, and influenced by a range of child-internal and child-external factors, it is not very likely that we will find a 'quick fix', such as the one language task that serves as a good discriminator of DLD; this may also be exacerbated by the heterogeneity of the disorder. We might need a combination (or battery) of tasks. However, it is also possible that the development and use of new language tasks as such will not help us to identify DLD in bilinguals. To interpret the performance on such tasks, we may need substantial background information on the child: reports on language exposure, language use at home and in (pre)school, early developmental milestones, family factors, parental and/or teacher concerns, and reports of functional language skills and communication difficulties (Öberg & Bohnacker 2022 [10]; Tuller 2015 [37]). Such information can be gathered via interviews with parents and teachers and/or via questionnaires (e.g., Paradis et al. 2010 [38] for the Canadian ALDEQ; Tuller 2015 [37] for the LITMUS PABIQ).

Since the official end of the COST Action in 2013, hundreds of studies have been carried out that employ one or several of the LITMUS tools, advancing our knowledge about monolingual and/or bilingual child language acquisition. Despite this flurry of papers, books and special issues (e.g., Applied Psycholinguistics 37(1), 2016; Armon-Lotem & Grohmann 2021 [39]; Bohnacker & Gagarina 2020b [40]; First Language 42(2), 2022; Linguistic Approaches to Bilingualism 12(4), 2022), it is noteworthy that only a small number of publications using the LITMUS language tasks provide reference data for a large enough group of TD bilingual children with the same language combination. Even fewer publications compare the performance of bilingual TD and bilingual DLD children growing up with the same language combination and in comparable settings. Very few studies investigate both of the children’s languages. Exposure and age are not always factored in. Studies that do include bilingual TD and DLD children largely confine themselves to reporting statistically significant differences between the TD and DLD groups. Whilst significant differences between groups may appear promising at first sight, not all studies expound on the finding that the performances of bilingual TD and DLD groups usually show a substantial amount of overlap. At the individual level then, particular language assessment tasks may not be that accurate, and accuracy at the individual level is what determines clinical utility.

The present study presents reference data for more than one hundred bilingual Turkish/Swedish TD 4-to-7-year-olds raised in Sweden for a range of language tasks from the LITMUS battery in both languages, investigating age, exposure factors and vocabulary size as predictors of performance, and comparing the different tasks to each other. The individual performances of a small sample of Turkish/Swedish DLD children are then compared to the TD group.

2. Previous Studies

This section provides an overview of existing studies that use the LITMUS NWR or CLT or MAIN with both bilingual TD and bilingual DLD children, and summarizes the main findings from these, as well as some other relevant studies in this domain.

2.1. NWR

Several North American studies on the clinical utility of NWR tasks with bilinguals have reported considerable TD/DLD overlap and mixed results regarding their diagnostic accuracy (e.g., Gutiérrez-Clellen & Simon-Cereijido 2010 [41]; Kohnert et al. 2006 [42]; Paradis et al. 2013 [28]; Sorenson Duncan & Paradis 2016 [15]; Windsor et al. 2010 [43]). They largely used language-specific NWR items.

In Europe, the COST Action IS0804 (e.g., Chiat 2015 [5]) put forward the idea that language-independent (quasi-universal) NWR tasks show greater potential to differentiate DLD from TD in bilingual children than language-specific NWR tasks. For instance, in
the Netherlands, Boerma et al. (2015 [12]) examined the NWR performance of 120 mono-
lingual and bilingual (early L2 Dutch, mixed L1) children aged 5–6 with and without a
diagnosis of DLD. Two different NWR tasks were used, a language-specific one with
Dutch-like nonwords, and a language-independent (quasi-universal) one, consisting of 2-
to-5-syllable items with simple CV syllable structure, no clusters and quasi-neutral pros-
ody (Chiat 2015 [5]; Chiat & Polišenská 2016 [44]). (This quasi-universal NWR task has
since then been renamed and is now called the cross-linguistic NWR task (https://www.bi-
sli.org/cl-nonword-repetition). Both NWR tasks distinguished the TD bilingual group (N =
30) from the DLD bilingual group (N = 30), but the quasi-universal NWR task did so
with higher sensitivity than the Dutch-specific NWR. There were significant differences
between the two groups, but Boerma et al.’s scores of individual TD and DLD children
showed a substantial overlap, even though their preselected DLD group was made up of
children in special education schools with relatively great educational needs who per-
formed at least 2 SD below the mean on a standardized Dutch test battery (e.g., CELF)
or at least 1.5 SD below the mean on two subscales. The L1s of the bilingual children were
not assessed; all testing was carried out in Dutch. Boerma and Blom (2017 [18]), investi-
gating roughly the same participant sample with a range of measures, again recom-
mended the language-independent NWR task (Chiat 2015 [5]) as a good diagnostic tool
for bilingual children. In a longitudinal follow-up, Boerma and Blom (2021 [13]) retested
the same children on the same tasks over the course of two years (from age 5–6 to 7–8).
Whilst the NWR results still distinguished the DLD children from the TD children at
group level when they were older, classification accuracy was lower than it had been at
age 5–6, at below 80%.

A number of other studies emanating from the COST Action have also recommended
NWR tasks for diagnostic use, documenting significant differences in the NWR perfor-
manoe of bilingual DLD and bilingual TD children at group level (e.g., Meir 2017 [45],
republished in Meir 2021 [46]) for Russian–Hebrew bilinguals (N = 108, age 5:5–6:9); Chilla
et al. (2021 [47]) for bilinguals aged 5:5–9:0, with L2 German (N = 56) or L2 French (N =
95); and dos Santos and Ferré (2018 [48]) for bilinguals aged 5:4–8:2 with L2 French (N =
43)). These studies did not use the same quasi-universal NWR task as Boerma and col-
leagues, but other NWR tasks from the LITMUS battery, and included both language-
specific and language-independent items. While there were significant differences be-
tween the TD and DLD groups in these studies, the NWR scores of individual DLD and
TD children again showed substantial overlaps. Bilingual DLD sample size was generally
small, and despite wide age ranges, the potential effect of age was not investigated.

A recent study of 110 Arabic/Swedish children aged 4–7 with and without a diagnosis
of DLD (Öberg & Bohnacker 2022 [10]) investigated the NWR performance on three dif-
ferent NWR tasks, and explored it in relation to age, language exposure, Swedish and
Arabic vocabulary size, and NWR item properties (length and complexity). The Ara-
bic/Swedish-speaking children performed a language-specific Swedish NWR task (with
word-like items, 2-to-5 syllables, also involving clusters) and two language-independent
LITMUS NWR tasks, one being the same as the one used by Boerma and colleagues (2015
[12], 2021 [13]), with simple CV syllable structure in all items (2-to-5 syllables, Chiat 2015
[5]), the other being another language-independent task (1-to-3 syllables, with clusters,
dos Santos and Ferré 2018 [48]). While repetition accuracy varied across these tasks and
was influenced by age, vocabulary size and NWR item properties, none of the three NWR
tasks, nor any particular type of nonword item, could identify DLD to a satisfactory de-
gree in the Arabic/Swedish bilingual children. Diagnostic utility was poor.

In an as yet unpublished large-scale study (Polišenská et al. 2022 [49]), which com-
bines the datasets of 1742 children from 18 teams spanning 15 countries using similar ver-
sions of the same language-independent NWR task (Chiat 2015 [5]), repetition accuracy
could not reliably differentiate children with a diagnosis of DLD from their TD peers.

Two recent reviews and meta-analyses of the discriminative function of NWR tasks
also report mixed results for bilinguals (Ortiz 2021 [8]; Schwob et al. 2021 [9]). Many of the
studies included in the meta-analyses did not reach a diagnostically informative value for language impairment. Both reviews also point to a possible publication bias, where studies reporting positive results concerning the discriminative function of NWR are being favored over studies with null results. The question of the diagnostic utility of NWR tasks with bilingual children is thus still unresolved.

The present paper is the first to explore NWR in a Turkish/Swedish context.

2.2. Vocabulary Assessment in Both Languages

When there is a suspicion of DLD, assessment often includes the child’s vocabulary skills (e.g., Peña et al. 2016 [50]). However, since bilinguals have their vocabulary distributed across both languages and since vocabulary size/growth is dependent on exposure, TD bilingual performance on standardized vocabulary tests may vary greatly. Some researchers therefore ascribe low clinical utility to vocabulary assessment tasks when trying to rule in or rule out DLD in bilinguals (e.g., Boerma & Blom 2017 [18]; Paradis et al. 2013 [28]). Other researchers, such as Smolander et al. (2021 [51]), find vocabulary tasks to be clinically useful for the assessment of bilingual children, provided that vocabulary scores are interpreted in relation to language exposure, or provided that the lexical skills in both languages are assessed (e.g., Peña et al. 2016 [50]).

As for the cross-linguistic vocabulary tasks (CLT, Haman et al. 2015 [25]) developed as part of the COST Action IS0804, we only know of two published studies on their diagnostic accuracy in bilinguals. Khoury Aouad Saliby et al. 2017 [52] compared the performance of 32 TD and 10 DLD bilingual children aged 5;7–7;10 in Lebanon who spoke Lebanese Arabic and French and/or English. Both groups scored near ceiling on Arabic noun comprehension, and also very highly on Arabic verb comprehension. The DLD group did score significantly lower than the TD group on both comprehension and production. Age and exposure were not factored in. Khoury et al. described the DLD children as having particular problems with the production of verbs. Despite a visible TD/DLD overlap in the plotted scores, Khoury et al. characterized the CLT as a diagnostically useful tool, stating that “the CLT-LB is reliable in the Lebanese context, as it differentiates Bi-TD from Bi-SLI children” (p. 16).

In another, larger, study also involving Arabic, but this time in a Swedish setting, Öberg and Bohnacker (2022 [10]) used CLTs with 110 Arabic–Swedish-speaking bilinguals aged 4;0–7;11 (99 TD and 11 DLD), both in their L1 Arabic and in their L2 Swedish. The children’s vocabulary scores were strongly influenced by age and exposure. Whilst some of the DLD children scored below their TD peers on either Arabic or Swedish vocabulary or both, others performed on a par with TD peers. In short, the CLT scores could not reliably distinguish between the DLD and TD children. Based on the results of these two studies, it is difficult to predict how other bilingual children will perform on the CLT vocabulary tasks.

In her doctoral thesis, Öztekin (2019 [53]) used a subset of the data of the present study to investigate the CLT scores of Turkish–Swedish-speaking children aged 4;0–8;1. Her comparisons of different age groups (4- vs. 5- vs. 6- vs. 7-year-olds) suggest an effect of age, particularly for the majority-language Swedish, and her longitudinal follow-up study of ten four-year-olds points in the same direction. Öztekin also argued that longer exposure time to Swedish, as well as an estimated daily exposure to Swedish at or above 80%, was associated with higher scores in Swedish. However, Öztekin (2019 [53]) did not employ any multivariate statistics to explore the potential/combined effects of different child-internal and environmental factors on the children’s vocabulary scores. The present study does just this, for both Turkish and Swedish.

2.3. Narrative Macrostructure

Prior to the COST Action, few studies investigated the narrative macrostructure of bilingual children in both their languages, or in bilinguals with and without DLD, in part due to a lack of comparable narrative assessment materials across languages. Some
(mainly North American) studies that only considered the majority language reported impaired macrostructural production in DLD bilingual children compared to TD bilingual children (e.g., Govindarajan & Paradis 2019 [27]; Paradis et al. 2013 [28]; Rezzonico et al. 2015 [54]). Other studies did not document any significant differences in macrostructure between these groups (e.g., Cleave et al. 2010 [55]; Iluz-Cohen & Walters 2012 [56]). These mixed results may be to do with the use of very different stimulus materials, elicitation modes (retelling vs. story generation) and macrostructural scoring methods. As for the comprehension of narrative macrostructure, which has been found to be impaired in DLD monolinguals (e.g., Bishop & Adams 1992 [57]; Dodwell & Bavin 2008 [58]; Merritt & Liles 1987 [31]), little is known about DLD bilinguals. The launch of the Multilingual Assessment Instrument for Narratives, MAIN, at the final conference of the COST Action (Gagarina et al. 2012 [59]), the subsequent launch of MAIN-Revised (Gagarina et al. 2019 [36]), and its availability in more than 80 languages (https://main.leibniz-zas.de/) has led to exponential growth in the field and may have the potential to unify bilingual narrative research methodologically. However, published studies that compare the narrative macrostructure skills of bilingual children with and without DLD are still very rare.

Tsimpli et al. (2016 [60]) let Greek-speaking children retell MAIN stories that they had listened to on a computer, including 15 TD and 15 DLD bilinguals (early L2 Greek, mixed L1s, aged 5−11). Results for MAIN story structure were not reported. For story complexity (calculated according to a formula of their own), Tsimpli et al. found no significant difference between the TD and DLD groups. It is difficult to draw any conclusions from this lack of a group difference, since neither exposure nor age (5–11 years) were factored in. The same group of researchers (Peristeri et al. 2020 [61]) also investigated the comprehension of macrostructure in 120 Greek-speaking children, including 30 TD and 30 DLD bilinguals aged 6;0−7;9 (L1 Albanian, and early L2 Greek); the sample overlapped in part with Tsimpli et al. The children answered the MAIN comprehension questions after listening to the story on a computer and retelling it to an experimenter. For this procedure and age range, narrative comprehension scores were high and the means did not significantly differ between the TD and DLD bilingual groups. However, this null result is hard to interpret, since the DLD children were more Greek-dominant than the TD group, according to Peristeri et al.’s measures of exposure and current language use. Only the children’s L2 Greek was assessed.

Boerma et al. (2016 [62]) investigated the comprehension and production of macrostructure with MAIN for 132 monolingual and bilingual children aged 5−6, in Dutch, including 33 TD and 33 DLD bilingual children (early L2 Dutch, mixed L1s); the sample greatly overlapped with the one in Boerma et al. (2015 [12]) described earlier for NWR. Only the children’s L2 Dutch was assessed. As the authors found significant differences between the TD and the DLD groups for macrostructure production, and even more so for the comprehension of macrostructure, they considered MAIN as a diagnostically valid tool for bilingual children. However, classification accuracy in Boerma et al. (2016 [62]) was only around 80%, which is not sufficient for clinical purposes, as the authors acknowledge themselves. In a study on the same participants, Boerma and Blom (2017 [18]) recommend that MAIN be used together with a language-independent/quasi-universal NWR task and a parental questionnaire to identify DLD in bilingual children. In a longitudinal follow-up that retested the same children on the same tasks, Boerma and Blom (2021 [13]) found that the magnitude of the difference in scores between the TD and DLD groups had shrunk over the course of two years. In fact, by age 7−8, the MAIN comprehension task (after listening to the story) no longer differentiated the TD and DLD groups, due to ceiling effects. Performance on the MAIN comprehension task without prior listening to the story could still differentiate the two groups, though less clearly than at lower ages, and this also held for the MAIN macrostructure production measure. Since the scores of individual DLD and TD children showed substantial overlaps, the question of the diagnostic validity of narrative macrostructure assessment in bilinguals remains open and needs to be researched further.
Öztekin’s (2019 [53]) doctoral thesis includes an investigation of the MAIN comprehension and production tasks for the Turkish/Swedish bilingual participants of the present study. Her comparisons of different age groups (4- vs. 5- vs. 6- vs. 7-year-olds) suggest an effect of age, particularly for the majority-language Swedish. However, Öztekin (2019 [53]) did not use any multivariate statistics to investigate background factors (such as age, length of L2 exposure, current daily L1/L2 exposure, or L1 and L2 vocabulary size) as predictors for narrative performance. The present paper adds this, for both Turkish and Swedish.

3. The Present Study: Aims and Research Questions

Although Turkish/Swedish-speaking children are a sizeable minority in Sweden, little is known about their language skills and DLD. There is a general lack of studies on this population, save for Öztekin’s doctoral thesis (2019 [53]). As in many other countries, research on reliable tools for DLD diagnosis in bilingual children is lacking in Sweden.

None of the (aforementioned) LITMUS language tasks are norm-referenced yet, therefore reference data from TD bilinguals are needed for a wide variety of language combinations, including information on task performance in relation to age, exposure and other language measures. Such reference data are a baseline to compare the performance of bilingual children with a DLD diagnosis against, not just at group level but also individually. Here, we are only at the beginning; our study of Turkish/Swedish 4-to-7-year-olds is one such contribution.

Our aim is to present reference data for four TD age groups for a range of tasks that can be utilized in future assessments for DLD. We investigate factors such as age, length of exposure, current daily language exposure and vocabulary size as predictors of performance, and compare the different tasks to each other, to show not only how well the children perform on the tasks themselves, but also to demonstrate how different linguistic skills in the different languages are sensitive to different background variables. Finally, the individual performances of a small sample of DLD children are compared to the TD group, to explore the discriminatory potential of the tasks. This mirrors the situation of speech-language pathologists who are confronted with individual children and not with groups. Our research questions are as follows:

**RQ1:** How do 4–7-year-old Turkish/Swedish-speaking bilinguals with typical language development perform on a language-independent and a language-specific Turkish NWR task, and how is performance affected by age, vocabulary size, and properties of the non-words (item length and type of task)?

**RQ2:** How do 4–7-year-old Turkish/Swedish-speaking bilinguals with typical language development perform on vocabulary comprehension and production in both languages, and how is that performance affected by age and language exposure?

**RQ3:** How do 4–7-year-old Turkish/Swedish-speaking bilinguals with typical language development perform on comprehension and production of narrative macrostructure in both languages, and how is that performance affected by age, language exposure and vocabulary size?

**RQ4:** By comparison, how do Turkish/Swedish-speaking bilinguals with a diagnosis of DLD perform on these NWR, vocabulary and narrative tasks, and does this allow for any conclusions to be drawn concerning the clinical utility of the LITMUS NWR, CLT and MAIN language tasks when assessing the language skills of bilingual children with suspected DLD?

4. Method

4.1. Participants

The participants were 114 Turkish–Swedish-speaking children aged 4;0–8;1, all growing up in Sweden. Of these, 108 had typical language development according to parental report (henceforth, the TD sample), and 6 had a DLD diagnosis from a certified speech-
language pathologist, SLP (the DLD sample). The children and their families were part of a larger child multilingualism project, BiLI-TAS (PI: Ute Bohnacker), at Uppsala University.

4.1.1. The TD Sample

The children in the TD sample were recruited by contacting 200 preschools and schools in urban areas of eastern central Sweden. Some children were also recruited via mother-tongue teachers of Turkish, personal contacts of Turkish-speaking members of the research team and associations arranging activities for Turkish-speaking children. One-hundred-and-ten children were recruited for the TD sample, two of whom dropped out of the study as parental consent was withdrawn. One-hundred-and-eight children remained.11 Inclusion criteria for participation were (a) being able to speak both Turkish and Swedish (at least to some extent) and (b) having no known hearing problems, language disorders or neuropsychiatric disorders according to parental report. The 108 children in the TD sample attended 50 different preschools and schools. As can be seen in Table 1, the number of participants was fairly even across age groups, with slightly more girls than boys overall.

Table 1. Participants in the TD sample: proportion girls/boys, mean age and age range.

<table>
<thead>
<tr>
<th>Age</th>
<th>Girls/boys</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Years</td>
<td>14/13</td>
<td>58/50</td>
</tr>
<tr>
<td>5 Years</td>
<td>15/10</td>
<td></td>
</tr>
<tr>
<td>6 Years</td>
<td>15/12</td>
<td></td>
</tr>
<tr>
<td>7 Years</td>
<td>14/15</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>58/50</td>
<td></td>
</tr>
</tbody>
</table>

Note. * Two children in the 7-year-old group had just turned 8 at the time of testing.

The vast majority (90%) of the participants were born in Sweden. The remaining 11 children were born in Turkey and had migrated with their families to Sweden. Most children were Turkish/Swedish-speaking bilinguals, but 13 children were exposed via their parents to an additional language other than Turkish and Swedish, and 6 were reported to also speak an additional language (mostly Kurdish).

Age of onset (AoO) for Turkish was at birth for 99 children (the remainder were first exposed to a third language, most often Kurdish, via their parents and exposed to Turkish soon after), and before age 3 for all children (for three children this information was missing). AoO for Swedish was more varied, as only 29 children were exposed to Swedish from birth; however, almost all children started to receive regular exposure to Swedish before age 3 (N = 91). Four children had been exposed to Swedish for less than 24 months. We decided not to exclude children with short residence lengths a priori, as we wanted to explore the effect of length of exposure. Daily exposure (according to parental estimation) varied widely, but was overall a bit higher for Swedish (M = 54.4%, SD = 21.3, Median = 60) than for Turkish (M = 45.1%, SD = 21.1, Median = 40). Younger children (the four-year-olds) tended to have more exposure to Turkish though (M = 49.6%, SD = 20.36, Median = 60).

All children attended institutional childcare. All children at age 4, all but one at age 5, and six children at age 6 were in preschool. In total, 1 five-year-old, 21 of the children at age 6, and 14 of the children at age 7 were in förskoleklass (a preparatory year between preschool and primary school), and the remaining 15 7-year-olds were in first grade of primary school. (In Sweden, primary school starts at age 7, but attending state-funded preschools from age 1 or 2 is commonplace.) The children had varying socio-economic backgrounds, with a variety of parental occupations and education levels (all levels from less than six years of primary school to doctorate degrees were represented). The majority of parents had finished secondary school, and about a third had some form of tertiary
education. A majority of the children attended (pre)schools in low-SES urban areas with high proportions of multilingual peers and staff.

4.1.2. The DLD Sample

The children in the DLD sample were recruited by contacting SLP clinics and specialized (pre)school units for children with DLD in the same regions as the TD sample. All in all, fourteen children fulfilled the inclusion criteria (see below) and were invited to participate in the study, but only six of them decided to participate in the end. Inclusion criteria were the following: (i) age between 4;0 and 8;1, (ii) being regularly exposed to Turkish and Swedish, and (iii) having a DLD diagnosis. All children had been assessed and diagnosed by a licensed SLP. Diagnoses could include any of the following categories in the ICD-10-SE diagnostic framework, utilized in Swedish healthcare at the time of assessment: mixed receptive and expressive disorder (Swe: *generell språkstörning*), primarily receptive disorder (Swe: *impressiv språkstörning*), primarily expressive disorder (Swe: *expressiv språkstörning*), or grammatical disorder (Swe: *grammatisk språkstörning*), but not exclusively phonological or articulatory difficulties (Swe: *fonologisk språkstörning*). Exclusion criteria were the following: (i) having a biomedical condition associated with language difficulties (e.g., Down syndrome), (ii) a diagnosis within the autism spectrum, or (iii) intellectual disability at the time of assessment. Although not an exclusionary criterion for participating in the DLD study, none of the children had ADHD. Table 2 below and Table A1 in the Appendix provide more information on the participants’ language impairment.

Table 2. The participants in the DLD sample: age, age of onset of Turkish and Swedish, estimated daily exposure, (pre)school attendance and DLD diagnosis.

| BiTurLI-01 | 6;3 at birth | 1;0–1;11 | Swe 80%, Tur 20% | Språkförskola | Mixed rec. and expr. LD |
| BiTurLI-02 | 5;4 at birth | 2;0–2;11 | Swe 40%, Tur 60% | Preschool | Mixed rec. and expr. LD + pragm. |
| BiTurLI-03 | 5;11 at birth | at birth | Swe 50%, Tur 50% | Preschool | Expr. LD (previously mixed rec. and expr. LD) |
| BiTurLI-04 | 4;9 at birth | 1;0–1;11 | Swe 40%, Tur 60% | Preschool | Mixed rec. and expr. LD + pragm. |
| BiTurLI-05 | 6;10 1;0–1;11 | 2;0–2;11 | Swe 40%, Tur 40%, Kur 20% | Förskoleklass | Mixed rec. and expr. LD + pragm. |
| BiTurLI-06 | 8;1 at birth | 3;0–3;11 | Swe 50%, Tur 30%, Kur 20% | 1st grade | Unspec. LD (likely mixed rec. and expr. LD) |

Note. (D)LD = (Developmental) Language Disorder, rec. = receptive, expr. = expressive, pragm. = pragmatic difficulties, Språkförskola = a specialized preschool unit for children with severe DLD, Förskoleklass = a preparatory year between preschool and primary school. BiTurLI-05 was first exposed to Kurdish only (therefore, AoO was not at birth for either Tur or Swe).

There were as many girls (3) as boys (3), and the age range (4;9–8;1) matched that of the TD sample.13 As can be seen in Table 2, almost all children had mixed receptive and expressive difficulties (‘general language disorder’), but one child had recently had the diagnosis changed to expressive LD, and one child had an unspecified diagnosis at the time of testing since the SLP had not been able to complete the assessment in Turkish, despite several attempts. For another child (BiTurLI-05) it was revealed to the SLP and the research team three months after testing that he had also received an autism diagnosis. However, the parents did not think that this diagnosis was accurate. In the end, it was decided not to exclude this child a posteriori. We decided to do so since the children in the DLD sample are treated as individual case studies rather than as a group.
Five children had an age of onset of Turkish from birth, and as many had an AoO of Swedish before age 3. BiTurLI-05 was first exposed to Kurdish via his parents, and started to receive regular exposure to Turkish at age 1, and to Swedish at age 2. Three children were reported to have even daily exposure to both languages. One child, BiTurLI-01, had a much higher proportion of Swedish (80%) than Turkish. Two children were exposed to Kurdish in addition to Swedish and Turkish. All children attended institutional childcare. Three children were in preschool, one child attended förskoleklass, and one child was in first grade of primary school. The sixth child attended språkförskola, a specialized preschool unit for children with severe DLD.

4.2. Materials
4.2.1. Nonword Repetition Tasks (NWR)

Two types of NWR tasks were used with the TD and DLD children in the present study, a language-specific one and a quasi-universal/cross-linguistic one. The items in these tasks all adhere to Turkish lexical phonology and syllable structure. Turkish has a rich phoneme inventory, and word-final devoicing, as well as vowel harmony, where vowels agree in backness. Turkish is an inflectional, agglutinative language, which is syllable-timed (i.e., all syllables, stressed or unstressed, are of roughly equal duration). Turkish syllable structure is mainly CV and CVC, where CV predominates in word-initial syllables, CVC and CV word-medially, and CVC word-finally (Topbaş et al. 2014 [6]). Consonant clusters are uncommon and only occur word-finally. Accordingly, a Turkish language-specific task (LS-Tur) was developed by Seyhun Topbaş, Dilber Kaçar and Handan Kopkallı-Yavuz at the Department of Speech and Language Pathology (DİLKOM), Anadolu University, and kindly made available to us. The LS-Tur task contains sixteen test items, with four items of each syllable length (2–5), each item starting with CV and always ending with CVC. Eight of the items are more Turkish-like, as they are made up of high-frequency syllables and have been rated by native-speaker panels as sounding ‘very Turkish-like’ (wordlikeness). The other eight items are less Turkish-like, as they contain syllables that occur at lower frequencies in Turkish and where the item has received lower wordlikeness ratings. All 16 items conform to Turkish phonology (including backness vowel harmony), no phoneme occurs more than once per item, only early-acquired phonemes occur in initial and final position, and items are made up of open syllables apart from the coda in the final syllable (Topbaş 2015 [63]). An earlier, longer (30-item) version of the LS-task is described in Topbaş et al. (2014 [6]), including the rationale and procedure for item construction, selection and rating, as well as the repetition accuracies of word-like and less word-like items by monolingual Turkish children (age d 4–8 years). The phoneme inventory contains fifteen consonants (/p, t, d, k, ɡ, m, n, s, z, ʃ, dʒ, ʋ, j, l, ɫ/) and nine vowels (/i, y, e, ø, ɛ, a, o, ɯ, u/). The items are pronounced with Turkish prosody, with equal stress on all syllables in terms of duration and pitch, except for final-syllable pitch drop.

Secondly, a cross-linguistic NWR task was used (Chiat 2015 [5], https://www.bisl.org/cl-nonword-repetition). Note that this cross-linguistic NWR task was previously called ‘the quasi-universal nonword repetition test’ (Boerma et al. 2015 [12]; Chiat 2015 [5]; Chiat & Polišenská 2016 [44]). This task was used because the ability to repeat quasi-universal items is widely assumed to be transferrable across languages, not to disadvantage bilingual children, and because it has been put forward by several researchers as having the greatest potential to differentiate bilingual TD from bilingual DLD (Section 1). The Turkish version of the task (CL-Tur) was adapted by Topbaş and colleagues at Anadolu University as part of the COST Action IS0804. The CL-Tur task consists of sixteen test items, with four items of each syllable length (2–5) and simple syllabic structure (only open syllables and no clusters). The phoneme inventory is comprised of eleven consonants (/p, b, t, d, k, ɡ, s, z, m, n, l/) and three vowels (/a, i, u/) phonetically realized in Turkish.
The items are pronounced with quasi-neutral prosody (equal stress in terms of length and pitch), except for pitch drop of the final syllable, marking the end of an utterance.

Both NWR tasks were professionally recorded in a TV studio by a female native Turkish speaker, cleaned for background noise, and embedded in an audio-visual computer application featuring a parrot with headphones. The items were presented with a three-second pause between each item, starting with the shortest (2 syllables) and ending with the longest items (5 syllables).

In addition to these NWR tasks in Turkish, the DLD children were administered two NWR tasks in Swedish. The Swedish version of the cross-linguistic NWR task (Chiat 2015 [5]), the CL-Swe, is near-identical to the CL-Tur and was developed by the first author for the BiLI-TAS project. It consists of 16 open-syllable items without clusters, 4 nonwords per syllable length (2–5), pronounced with quasi-neutral prosody and pitch drop on the final syllable. The CL-Swe phoneme inventory contains three vowels (/a, i, u/) and ten consonants (/p, b, t, d, k, g, s, m, n, l/), phonetically realized in Swedish. (The CL-Swe inventory thus contains one consonant fewer than the CL-Tur, as /z/ does not exist in Swedish.) The Swedish language-specific NWR task (LS-Swe) is not part of the LITMUS battery, but had previously been developed by two Swedish SLP students (Barthelom & Åkesson 1995 [64]) and was subsequently published (Radeborg et al. 2006 [65]). The LS-Swe contains 24 2-to-5 syllable items that adhere to Swedish phonotactics, with Swedish phonemes, nineteen consonants (/p, b, t, d, k, g, m, n, ŋ, r, f, v, s, ð, ɾ, f, v, s,ɕ,ɧ,ʂ, h, j, l/) and fifteen vowels (/i, ɪ, y, ʏ, e, ɛ, œ, ɑ, a, o, ɔ, u, ʊ, ʉ, ɵ/). The phonological complexity of the LS-Swe items varies, as they contain open or closed syllables, with and without consonant clusters in onset and coda. Main stress and vowel duration in the nonwords also vary across the syllables, in typical Swedish stress pattern fashion. The LS-Swe and CL-Swe items were prerecorded by a female native speaker of a central Swedish variety.

A list of all the items in the NWR tasks is provided in Table A2 in the Appendix.

4.2.2. Cross-Linguistic Lexical Tasks (CLTs)

Vocabulary skills were assessed with the Turkish and the Swedish versions of the Cross-linguistic Lexical Tasks (CLT), a picture-based vocabulary assessment material targeting comprehension and production of nouns and verbs (Haman et al. 2015 [25]; Ringblom et al. 2014 [66]; Ünal-Loğacev et al. 2013 [67]; https://multilada.pl/en/projects/clt/). The CLT was developed specifically for assessing vocabulary in both languages of bilingual children as part of the COST Action IS0804. Each of the two comprehension parts (nouns, verbs) and each of the two production parts (nouns, verbs) consists of thirty test items preceded by two practice items. Thus, the maximum score is 60 for CLT comprehension and 60 for CLT production. Comprehension is assessed via picture selection. For each test item, the experimenter asks a prompt question (e.g., ‘who is pouring?’) and from an array of four, the child has to select the correct picture. Vocabulary production is assessed via picture naming, where individual pictures are shown to the child one by one. The experimenter asks the child to name a depicted object (e.g., ‘what is this?’) or action (e.g., ‘what is she doing?’).

4.2.3. Multilingual Assessment Instrument for Narratives (MAIN)

Narrative macrostructure comprehension and production was assessed with the Turkish and the Swedish versions of the Multilingual Assessment Instrument for Narratives (MAIN) (Gagarina et al. 2019 [36]) https://main.leibniz-zas.de/en/). The MAIN was developed for assessing narrative macrostructure in both languages of bilingual children as part of the COST Action IS0804. There are four picture sequences in MAIN (Cat, Dog, Baby Birds and Baby Goats), all containing six pictures each. These four stories were designed to be matching in terms of length and story grammar components. Furthermore, two stories each are parallel in terms of plotline and number of characters. Due to this parallelism, we refer to Cat/Dog as one task and Baby Birds/Baby Goats as the other.
4.2.4. Parental Questionnaire

All parents filled in a questionnaire on the social and linguistic background of the participating children and their families. The parental questionnaire was developed as part of a larger research project on child multilingualism (BiLI-TAS) at Uppsala University. The questionnaire could be answered in either Turkish or Swedish, whichever the parents preferred. The questions targeted language exposure, language use in the family, language activities in the home such as book reading and storytelling, language attitudes, parental language skills, and education and occupation, as well as family history of language and/or literacy difficulties, the child’s early language development and parental concerns. The parents of the children in the DLD sample filled in an almost identical questionnaire, with additional questions about SLP referral and therapy.

4.2.5. Interview Data

The interview questions for parents, teachers and SLPs of the DLD children were developed by the BiLI-TAS project research team. Parents were interviewed in connection with the Turkish data collection in the home. The interview targeted the same topics as the questionnaire, but elicited more detailed information about the child’s language development over time and whether the parents were concerned about their child’s language development, as well as the parents’ attitudes and beliefs about bilingualism. Teachers were interviewed in connection with the Swedish data collection at (pre)school. The questions targeted the child’s language and social skills, classroom behavior and learning outcomes. The SLPs were interviewed by telephone after data collection in both languages was done. SLPs were asked about assessment procedures (in which language(s), and which materials were used), language therapy, development over time and the child’s current language status. A condensed overview of the interview data can be found in Table A1 in the Appendix. For more details on the interviews with parents, teachers and SLPs, see Öztekin (2019 [53]).

4.3. Data Collection Procedure

The data were collected between February 2015 and October 2016 for the TD sample, and February through April 2018 for the DLD sample. Families received oral and written information about the project in both Turkish and Swedish. Informed consent was obtained from the parents in writing before data collection, and children gave oral assent. Participation could be discontinued at any time. All children were seen on two separate occasions, one in each language (interval between sessions: M = 13.6, SD = 7.9, range 4–36 days). The meetings took place in a quiet room at (pre)school or in the home. Each session was audio- and video-recorded for later transcription and scoring, and lasted for about 30–45 min. The experimenters were trained native speakers of Turkish and Swedish who spoke to the child only in the language of testing, in order to be able to assess the children’s abilities in each language separately. The order of the languages (Swedish or Turkish first), as well as the order of the tasks, was counterbalanced. The children were assessed with the NWR tasks, the CLT and the MAIN, as part of a test battery that also included a second narrative task (for the TD sample) and two Swedish NWR tasks (for the DLD sample). The child was given stickers and praise at the end of each task.

The NWR tasks were administered via audio-visual computer applications. The tasks were presented to the children as an imitation game, featuring a parrot that the child was instructed to imitate. The task was presented on a laptop and the audio was played to the child via headphones.15

The CLT was administered via colored picture booklets and followed the standard procedure described by Haman et al. (2015 [25]). During the session, responses were noted on paper forms. The experimenter gave only neutral feedback (e.g., aha, mhm, okay) irrespective of whether the child had provided a correct answer or not. All responses were scored after each session.
The MAIN was administered according to the ‘telling’ mode described by Gagarina et al. (2019 [36]). All children told two stories per language, Cat or Dog and Baby Birds or Baby Goats. First, the child told the story viewing the pictures, but they were not visible to the experimenter. After the child had finished telling, the experimenter laid down the picture sequence on the table for shared visual attention and asked ten standardized questions, all probing inferential comprehension of the macrostructural components of the story.

As a few children did not complete every task in the Turkish and/or the Swedish session, Table 3 provides an overview of how many children per age group, as well as in the sample as a whole, performed which task in the TD sample. In the DLD sample, all children completed all tasks in both languages, apart from BiAraLI-03, who was unable to carry out MAIN BB/BG in Turkish.

Table 3. Number of children who completed each task (TD sample).

<table>
<thead>
<tr>
<th></th>
<th>4 Years</th>
<th>5 Years</th>
<th>6 Years</th>
<th>7 Years</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NWR LS-Tur and CL-Tur</td>
<td>N = 26</td>
<td>N = 25</td>
<td>N = 26</td>
<td>N = 26</td>
<td>N = 103</td>
</tr>
<tr>
<td>Turkish vocabulary comp and prod</td>
<td>N = 25</td>
<td>N = 23</td>
<td>N = 26</td>
<td>N = 28</td>
<td>N = 102</td>
</tr>
<tr>
<td>Swedish vocabulary comp and prod</td>
<td>N = 25</td>
<td>N = 23</td>
<td>N = 26</td>
<td>N = 28</td>
<td>N = 102</td>
</tr>
<tr>
<td>Turkish Cat/Dog narrative comp/prod</td>
<td>N = 24/24</td>
<td>N = 22/22</td>
<td>N = 26/26</td>
<td>N = 28/28</td>
<td>N = 100/100</td>
</tr>
<tr>
<td>Turkish BB/BG narrative comp/prod</td>
<td>N = 24/24</td>
<td>N = 21/22</td>
<td>N = 26/26</td>
<td>N = 28/28</td>
<td>N = 99/100</td>
</tr>
<tr>
<td>Swedish BB/BG narrative comp/prod</td>
<td>N = 23/24</td>
<td>N = 22/22</td>
<td>N = 26/26</td>
<td>N = 28/28</td>
<td>N = 99/100</td>
</tr>
</tbody>
</table>

Note. One five-year-old only completed the Turkish session. One seven-year-old only completed the Swedish session.

4.4. Data Treatment

4.4.1. Scoring

All NWR responses from the Turkish session were transcribed and scored by a trained native-Turkish research assistant. The total number of responses was 3 488 (1 744 for each NWR task: 16 items × 109 participants, 103 TD children and 6 DLD children). A second rater, a native-Turkish PhD student of linguistics and SLP who had also been the experimenter in the Turkish data collection for most of the children, independently scored the NWR responses of a random sample of 14% of the data (12% of the TD children (N = 12), and 50% of the DLD children (N = 3)). Any disagreements were resolved by consensus after discussions with the authors. The authors also checked the scoring of items for internal consistency. Scores (item correct vs. incorrect) matched for 88% (211/240 items) of the checked items for the LS-task, and for 90% (215/240 items) of the items for the CL-task. Inter-rater reliability for item vs. incorrect item (unweighted Cohen’s Kappa) was 0.75 (p < 0.001) for the LS-task and 0.74 (p < 0.001) for the CL-task. Scoring was carried out as follows. One point was awarded for each correctly repeated item, and 0 points for responses containing an error. Any consistent phonological substitution processes in the child’s speech were disregarded, and allowances were made for indistinct pronunciation of /s/. Errors of voicing (e.g., /p/ vs. /b/) and minor vowel deviations (e.g., /a/ vs. /o/) were also disregarded, but major vowel substitutions (e.g., substituting /a/ for /i/) were not allowed. Following Dollaghan and Campbell (1998 [19]), additions of syllables or phonemes before or after an otherwise correctly repeated item were also disregarded (i.e., children were not penalized for hesitation noises). The responses on the Swedish NWR tasks (administered only to the 6 DLD children), a total of 240 responses ((16 × 24 items) × 6) were transcribed and scored by a trained native-Swedish research assistant and SLP, and checked by the authors. Scoring followed the same principles as the scoring of the Turkish NWR responses.

The CLT forms filled in by the experimenters were checked against the audio and video recordings. There were 25,920 responses (108 participants × 2 languages × 120 test
items, 60 for comprehension and 60 for production) in total. Scoring was carried out by trained native or near-native speakers of Turkish and Swedish, respectively. Responses were carefully checked for accuracy and consistency, following project-internal scoring guidelines based on data from 220 monolingual and bilingual children. As there is no standardized published procedure for scoring the CLT, scoring was carried out as follows. One point was awarded for each correct response in the target language, and zero points were awarded for incorrect responses. For the comprehension tasks, only target picture identifications were scored as correct. For the production tasks, a point was awarded for each correctly produced target word. In addition, adult-like synonyms, responses more specific than the target word but still corresponding to the picture, and responses pronounced slightly off-target were also scored as correct. All other responses were scored as incorrect. For more detailed information about transcription and scoring, see Öztekin (2019, pp. 91–92 [53]).

All MAIN narratives, as well as the responses to the comprehension questions, were transcribed orthographically and carefully checked by trained native or near-native speakers of Turkish and Swedish, respectively. Next, all narratives and the responses to the comprehension questions were scored for story structure, according to the standardized procedure described by Gagarina et al. (2019 [36]). The maximum score for story structure production was 17 points, one point for each of the story components that the child mentioned in their narrative: time setting and place setting, plus one point each for internal state as initiating event, goal, attempt, outcome, and internal state as reaction (5 components × 3 episodes). For story comprehension, the maximum score was 10 points. The questions probed the children’s comprehension of the characters’ goals and internal states, as well as inferences about the general plotline. One point was awarded for each of the comprehension questions that were answered correctly. See Gagarina et al. (2019 [36]) and Bohnacker and Gagarina (2020a [68]), for more details.

4.4.2. Statistical Analyses

All statistical analyses were conducted in R (R Core Team 2021 [69]). The level of significance was set at $p < 0.05$ (two-tailed) for all analyses. Correlations were calculated with Pearson’s correlation coefficient (Pearson’s $r$).

Three variables from the questionnaire were investigated with respect to performance on NWR, vocabulary and narrative tasks: chronological age (in months), Length of Exposure to Swedish (LoE Swe) and current daily exposure (daily exp) to Turkish and Swedish. LoE Swe was calculated by subtracting Age of Onset (AoO) for Swedish (in months) from the child’s current age (in months). (As AoO for Turkish was at birth for 92% (99/108) of the children in the TD sample, LoE to Turkish could not be investigated as a separate variable.) The child’s current daily exposure to each language was estimated by the parents on a scale with seven levels ranging from almost only Swedish (95% Swe, 5% Tur) to almost only Turkish (5% Swe, 95% Tur). Parents could also note a different distribution; for example, if the child was exposed to a third language in addition to Turkish and Swedish. Daily exposure was then split into one separate variable for each language: daily exposure to Turkish (daily exp Tur) thus specifies the percentage of daily exposure to Turkish, and daily exposure to Swedish (daily exp Swe) specifies the percentage of daily exposure to Swedish.

Accuracy was investigated for the Turkish NWR tasks with a logistic mixed-effects regression model, using the function `glmer` from the `lme4` package (Bates et al. 2015 [70]). The mixed-effects model was evaluated with pseudo-$R^2$ and concordance index ($c$-index). Pseudo-$R^2$ was obtained with the `r.squaredGLMM` function from the MuMIn package (Bartoń 2020 [71]). See Section 5.1 for more information about the model.

For vocabulary and narrative macrostructure, multivariate linear regression models were fitted with the vocabulary or narrative macrostructure score as the dependent variable. Comprehension and production was investigated separately for each language, thus
there were eight separate models. For more information about specific models, see Sections 5.1–5.3.

Finally, age-adjusted z-scores were calculated, in order to compare individual performances of the DLD children on the NWR, vocabulary and narrative macrostructure tasks to those of the children in the TD sample. All individual z-scores were calculated based on the mean and SD for the children in the corresponding age group in the TD sample. Thus, all z-scores indicate how each individual performed on a specific task compared to age-group peers in the TD sample.

5. Results: The TD Sample

5.1. NWR

First, scores are reported by age group for the two NWR tasks. As evident in Table 4, overall performance was slightly lower in the LS-Tur task than in the CL-Tur task for all age groups, and mean scores increased with age among all age groups. In the following, age will be treated as a continuous variable. There were positive correlations between linear age and scores on both NWR tasks (LS-Tur: \( df = 101, r = 0.28, p = 0.004 \); CL-Tur: \( df = 101, r = 0.29, p = 0.003 \)). However, scores did not increase substantially with age, as evident from the small effect sizes of the correlations, and only small differences in mean scores between the age groups (see Table 4). Note that this was not due to a ceiling effect, since even the mean scores for the seven-year-olds were several points below the maximum score. Only for the LS-Tur task was there a (positive) correlation with Turkish vocabulary (comprehension: \( df = 96, r = 0.21, p = 0.04 \), production: \( df = 96, r = 0.22, p = 0.03 \) but not for the CL-Tur task.16

Table 4. Mean scores, standard deviations (SD) and ranges for the NWR tasks by age groups. Max score = 16 for both tasks.

<table>
<thead>
<tr>
<th>Age Group</th>
<th>LS-Tur Mean (SD)</th>
<th>LS-Tur Range</th>
<th>CL-Tur Mean (SD)</th>
<th>CL-Tur Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Years</td>
<td>9.6 (3.1)</td>
<td>4–14</td>
<td>10.9 (2.3)</td>
<td>7–15</td>
</tr>
<tr>
<td>5 Years</td>
<td>10.6 (2.5)</td>
<td>6–15</td>
<td>11.2 (2.2)</td>
<td>7–15</td>
</tr>
<tr>
<td>6 Years</td>
<td>11.1 (2.6)</td>
<td>4–15</td>
<td>11.7 (2.8)</td>
<td>4–16</td>
</tr>
<tr>
<td>7 Years</td>
<td>11.6 (2.4)</td>
<td>5–16</td>
<td>12.6 (1.8)</td>
<td>8–15</td>
</tr>
<tr>
<td>Total</td>
<td>10.7 (2.7)</td>
<td>4–16</td>
<td>11.6 (2.4)</td>
<td>4–16</td>
</tr>
</tbody>
</table>

Next, accuracy (percent correctly repeated items) was investigated in relation to item length (number of syllables). As shown in Figure 1, accuracy is the highest for the shortest items (2 syllables) in both tasks, and gradually decreases for each added syllable. For three- and four-syllable items, accuracy is somewhat higher on the CL-Tur task. For the shortest as well as the longest items (2 and 5 syllables, respectively), accuracy is similar in both tasks.

A logistic mixed-effects regression model (Model 1, see Table A3 in the Appendix for a summary) was fitted in order to investigate the effect of item-related factors and participant-related factors on repetition accuracy. The model investigated the effect of age, Turkish vocabulary comprehension,17 task (CL-Tur vs. LS-Tur), item length (number of syllables), and the interaction between task and Turkish vocabulary on repetition accuracy in both NWR tasks (fixed effects). By-participant and by-item intercepts were included in the model as random effects. While chronological age (\( B = 0.20, SE = 0.09, p = 0.02 \) had a positive effect on repetition accuracy, an increasing number of syllables had a negative impact (\( B = -0.82, SE = 0.16, p < 0.001 \)). While there were no main effects of Turkish vocabulary scores (\( B = -0.05, SE = 0.10, p = 0.64 \)) or task (\( B = -0.42, SE = 0.31, p = 0.17 \)), there was a significant interaction between task and vocabulary (\( B = 0.21, SE = 0.09, p = 0.02 \), demonstrating that vocabulary size positively affected repetition accuracy for LS-Tur items but not for CL-Tur items. The explanatory power of the full model was considerable (conditional \( R^2 = 0.37 \) and better than that of the fixed effects alone (marginal \( R^2 = 0.15 \)). The C-index of 0.82 indicates a good model fit.
5.2. Vocabulary

First, we present CLT scores in the two languages separately for each age group. As Table 5 shows, mean scores increased with age for comprehension and production in both languages. Scores are presented here for age groups so that they are able to be used as reference data, but in the following, age will be treated as a continuous variable. Linear age correlated positively with Turkish comprehension ($df = 100$, $r = 0.28$, $p = 0.005$), Swedish comprehension ($df = 100$, $r = 0.62$, $p < 0.001$) and Swedish production ($df = 100$, $r = 0.61$, $p < 0.001$). For Turkish production, the trend was similar, but failed to reach significance ($df = 100$, $r = 0.19$, $p = 0.06$). (This was likely due to substantial individual variation depending on factors other than age.)

Table 5. Mean scores, standard deviations (SDs) and ranges for Turkish and Swedish CLT vocabulary comprehension and production by age groups. Max score = 60 for all tasks.

<table>
<thead>
<tr>
<th></th>
<th>4 Years</th>
<th>5 Years</th>
<th>6 Years</th>
<th>7 Years</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Turkish</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>comprehension</td>
<td>Mean (SD)</td>
<td>51.4 (5.7)</td>
<td>53.8 (8.3)</td>
<td>54.4 (7.1)</td>
<td>56.9 (4.1)</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>35–58</td>
<td>26–60</td>
<td>27–60</td>
<td>40–60</td>
</tr>
<tr>
<td>Vocabulary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>production</td>
<td>Mean (SD)</td>
<td>36.4 (10.3)</td>
<td>37.6 (13.2)</td>
<td>39.5 (12.2)</td>
<td>42.4 (10.9)</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>5–52</td>
<td>3–56</td>
<td>12–53</td>
<td>13–60</td>
</tr>
<tr>
<td><strong>Swedish</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>comprehension</td>
<td>Mean (SD)</td>
<td>41.2 (9.3)</td>
<td>48.6 (8.1)</td>
<td>55.3 (4.5)</td>
<td>54.9 (5.0)</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>18–60</td>
<td>31–59</td>
<td>45–60</td>
<td>42–60</td>
</tr>
<tr>
<td>Vocabulary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>production</td>
<td>Mean (SD)</td>
<td>25.4 (9.6)</td>
<td>31.1 (9.2)</td>
<td>39.8 (7.7)</td>
<td>42.1 (9.0)</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>8–47</td>
<td>16–46</td>
<td>25–51</td>
<td>19–54</td>
</tr>
</tbody>
</table>

Four multivariate linear regression models were run to investigate the joint effect of age, estimated daily exposure and length of exposure (for Swedish) on Swedish and Turkish vocabulary scores. Summaries of the regression models can be found in Tables A4 and A5 in the Appendix.

For Turkish comprehension scores, age, length of exposure to Swedish and daily exposure to Turkish explained 20% of the variance (Model 2). Standardized estimates showed that age ($\beta = 0.50$, $p < 0.001$) was a stronger predictor than daily exposure to Turkish ($\beta = 0.25$, $p = 0.009$) for Turkish comprehension scores. There was also a tendency for length of exposure to Swedish to be a negative predictor of Turkish comprehension scores ($\beta = -0.22$, $p = 0.057$). For Turkish production, age, length of exposure to Swedish and daily exposure to Turkish explained 19% of the variance in scores (Model 3). Just like for comprehension, age was the strongest predictor ($\beta = 0.44$, $p < 0.001$), followed by a negative effect of length of exposure to Swedish ($\beta = -0.33$, $p = 0.005$) and finally a positive effect of daily exposure to Turkish ($\beta = 0.26$, $p = 0.005$).
As for Swedish vocabulary, age, length of exposure to Swedish and daily exposure explained 49% (Model 4) of the variance in comprehension scores, with age (β = 0.38, p < 0.001) being a slightly stronger predictor than length of exposure (β = 0.34, p < 0.001) and daily exposure (β = 0.16, p = 0.03). For Swedish production, age, length of exposure and daily exposure explained 53% of the variance (Model 5). Age (β = 0.36, p < 0.001) and length of exposure (β = 0.34, p < 0.001) were stronger predictors than daily exposure (β = 0.26, p < 0.001). Thus, the models investigating Swedish vocabulary explained more than twice as much of the variance in scores compared to the models investigating Turkish vocabulary.

5.3. Narrative Macrostructure

First, we present the MAIN scores separately for each age group (Table 6). There were four scores per child, one for macrostructure comprehension and one for macrostructure production in Turkish and Swedish, respectively. For simplicity, the story structure scores for MAIN Cat/Dog and Baby Birds/Baby Goats have been combined here. (Descriptive statistics for the two tasks separately (Cat/Dog, Baby Birds/Baby Goats) are found in Tables A8 and A9 in the Appendix.) Table 6 shows that the mean scores for comprehension and production increase with age in both languages. Scores for age groups are provided here as reference data, but in the following, age will be treated as a continuous variable. Linear age correlated positively with Turkish comprehension (df = 100, r = 0.43, p < 0.001) and production (df = 100, r = 0.55, p < 0.001), as well as with Swedish comprehension (df = 100, r = 0.68, p < 0.001) and production (df = 100, r = 0.66, p < 0.001).

Table 6. Mean scores, standard deviations (SDs) and ranges for Turkish and Swedish narrative macrostructure comprehension (max score = 20) and production (max score = 34) by age groups. Cat/Dog and Baby Birds/Baby Goats are combined.

<table>
<thead>
<tr>
<th></th>
<th>4 Years</th>
<th>5 Years</th>
<th>6 Years</th>
<th>7 Years</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Turkish narrative comprehension</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>9.8 (4.5)</td>
<td>12.4 (5.5)</td>
<td>14.1 (4.3)</td>
<td>15.1 (3.9)</td>
<td>12.9 (4.9)</td>
</tr>
<tr>
<td>Range</td>
<td>0–17</td>
<td>0–19</td>
<td>1–20</td>
<td>1–20</td>
<td>0–20</td>
</tr>
<tr>
<td><strong>Turkish narrative production</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>6.2 (3.9)</td>
<td>8.7 (4.5)</td>
<td>10.2 (3.5)</td>
<td>12.5 (3.7)</td>
<td>9.5 (4.5)</td>
</tr>
<tr>
<td>Range</td>
<td>0–11</td>
<td>0–15</td>
<td>3–16</td>
<td>1–18</td>
<td>0–18</td>
</tr>
<tr>
<td><strong>Swedish narrative comprehension</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>8.8 (5.0)</td>
<td>11.9 (4.9)</td>
<td>15.7 (1.8)</td>
<td>17.2 (2.1)</td>
<td>13.6 (4.9)</td>
</tr>
<tr>
<td>Range</td>
<td>0–16</td>
<td>0–19</td>
<td>12–19</td>
<td>10–20</td>
<td>0–20</td>
</tr>
<tr>
<td><strong>Swedish narrative production</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>6.0 (4.2)</td>
<td>8.9 (3.8)</td>
<td>11.5 (3.0)</td>
<td>13.7 (3.4)</td>
<td>10.1 (4.6)</td>
</tr>
<tr>
<td>Range</td>
<td>0–15</td>
<td>3–17</td>
<td>7–17</td>
<td>5–20</td>
<td>0–20</td>
</tr>
</tbody>
</table>

Four multivariate regression models were run to investigate the joint effect of age and vocabulary on narrative scores, controlling for language exposure in terms of length of exposure to Swedish and percent daily exposure to Turkish/Swedish. The regression models are summarized in Tables A6 and A7 in the Appendix.

Model 6 explained 53% of the variance in Turkish macrostructure comprehension scores. Standardized estimates showed that Turkish vocabulary production was the strongest predictor (β = 0.62, p < 0.001), followed by age (β = 0.27, p = 0.004). Length of exposure to Swedish (β = 0.06, p = 0.55) and daily exposure to Turkish (β = 0.02, p = 0.82) were not significant predictors.16 For Turkish macrostructure production, Model 7 explained 57% of the variance. Again, Turkish vocabulary production (β = 0.56, p < 0.001) was a stronger predictor than age (β = 0.36, p < 0.001). Just like for comprehension, neither length of exposure to Swedish (β = 0.12, p = 0.17) nor daily exposure to Turkish (β = 0.04, p = 0.61) were significant predictors of Turkish narrative production scores.

For Swedish macrostructure comprehension scores, Model 8 explained 64% of the variance. Swedish vocabulary production was the strongest predictor (β = 0.60, p < 0.001), followed by age (β = 0.36, p < 0.001). Neither of the exposure measures were significant (length of exposure to Swedish: β = −0.06, p = 0.51, daily exposure to Swedish: β = −0.06, p
Model 9 explained 61% of Swedish macrostructure production scores, with Swedish vocabulary production being a stronger predictor ($\beta = 0.55$, $p < 0.001$) than age ($\beta = 0.32$, $p < 0.001$). Again, neither length of exposure to Swedish ($\beta = 0.00$, $p = 0.99$) nor daily exposure to Swedish ($\beta = 0.00$, $p = 0.96$) were significant predictors.

5.4. Summary: NWR, Vocabulary, and Narrative Macrostructure in the TD Sample

Summarizing the results for the TD sample, for NWR, there was only a little development with age. Performance was slightly higher on the cross-linguistic task (CL-Tur) than on the language-specific task (LS-Tur). As item length increased, repetition accuracy decreased with each added syllable for both tasks. When age was accounted for in the multivariate model, there was a positive effect of Turkish vocabulary scores only for the LS-Tur task.

For vocabulary, development with age was found in both languages and both modalities, though it was more pronounced for Swedish than for Turkish. Age, length of exposure to Swedish and percentage daily exposure (to the specific language) explained more of the variation in Swedish vocabulary scores than for Turkish. Age was the most important predictor of vocabulary scores in both languages and modalities, overshadowing exposure measures. Percent daily exposure to Turkish and Swedish in the respective language was a positive predictor of vocabulary scores in both Turkish and Swedish. Length of exposure to Swedish was a positive predictor of Swedish vocabulary comprehension and production scores, and a negative predictor of Turkish production scores.

For narrative macrostructure, we found a development with age in both languages and both modalities. Vocabulary production scores in the respective language were the strongest predictors of macrostructure scores for comprehension and production in both languages, followed by chronological age. Interestingly, language exposure measures (length of exposure to Swedish and daily exposure to the corresponding language) were not significant predictors of narrative macrostructure in either language or modality.

6. Results: The DLD Sample

6.1. Profiles of the Individual Children

This section compares the performance of the DLD children to their TD peers. The comparison is made by transforming each child’s raw scores into an age-adjusted z-score (based on the mean and SD for that task for the corresponding age group in the TD sample). We opted here for a z-score below $-1.25$ as a cut-off, which identifies the lowest-scoring 10.6% of an age group on a specific task (Tomblin et al. 1997 [72]).

Individual scores for all tasks are presented in Table 7 (raw scores and z-scores) and Figure 2 (z-scores). Performance was mostly below the mean, as evident by the predominantly negative z-scores. Looking at individual profiles, some children had a clear discrepancy in performance between the two languages; BiTurLI-01, for instance, had very low scores in Turkish and within the typical range in Swedish, while BiTurLI-02 had low scores in Swedish but average scores in Turkish. Interestingly, vocabulary z-scores of the DLD children were generally lower for comprehension than for production, particularly in the minority language (Turkish). A similar pattern emerged for narrative macrostructure, with overall lower z-scores for comprehension in Turkish. Furthermore, the five children who had a z-score below $-1.25$ in Turkish vocabulary comprehension all scored below $-1.25$ in Turkish narrative macrostructure comprehension. One child (BiArALI-05) had very low scores across the board, particularly in Turkish and Swedish vocabulary and narrative comprehension.
Table 7. NWR, vocabulary and narrative macrostructure scores (raw scores and age-adjusted z-scores) for the children in the DLD sample.

<table>
<thead>
<tr>
<th>Age</th>
<th>Score</th>
<th>NWR Comp</th>
<th>NWR Prod</th>
<th>Vocabulary Comp</th>
<th>Vocabulary Prod</th>
<th>Narrative Macrostructure Comp</th>
<th>Narrative Macrostructure Prod</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiTur</td>
<td>6;3</td>
<td>raw 4</td>
<td>2</td>
<td>30</td>
<td>4</td>
<td>57</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>z -2.76*</td>
<td>-2.71*</td>
<td>-3.43*</td>
<td>-2.92*</td>
<td>-3.02*</td>
<td>-2.90*</td>
</tr>
<tr>
<td>BiTur</td>
<td>5;4</td>
<td>raw  3</td>
<td>5</td>
<td>59</td>
<td>38</td>
<td>35</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>z -2.98*</td>
<td>-2.81*</td>
<td>0.63</td>
<td>0.03</td>
<td>-1.67*</td>
<td>-1.75*</td>
</tr>
<tr>
<td>BiTur</td>
<td>5;11</td>
<td>raw  7</td>
<td>14</td>
<td>24</td>
<td>5</td>
<td>46</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>z -1.40*</td>
<td>1.24</td>
<td>-3.59*</td>
<td>-2.46*</td>
<td>-0.32</td>
<td>0.10</td>
</tr>
<tr>
<td>BiTur</td>
<td>4;9</td>
<td>raw  10</td>
<td>14</td>
<td>40</td>
<td>25</td>
<td>24</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>z 0.12</td>
<td>1.37</td>
<td>-2.00*</td>
<td>-1.11</td>
<td>-1.85*</td>
<td>-1.08</td>
</tr>
<tr>
<td>BiTur</td>
<td>6;10</td>
<td>raw  3</td>
<td>9</td>
<td>45</td>
<td>5</td>
<td>46</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>z -3.14*</td>
<td>-0.94</td>
<td>-1.33*</td>
<td>-1.52*</td>
<td>-6.75*</td>
<td>-3.61*</td>
</tr>
<tr>
<td>BiTur</td>
<td>8;1</td>
<td>raw 11</td>
<td>11</td>
<td>49</td>
<td>38</td>
<td>47</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>z -0.25</td>
<td>-0.89</td>
<td>-1.90*</td>
<td>-0.41</td>
<td>-1.58*</td>
<td>-1.00</td>
</tr>
</tbody>
</table>

Note. Z-scores < -1.25 are in bold and marked with *. BiTurLI-06 is compared to the group closest in age (the 7-year-olds).

Figure 2. Age-adjusted z-scores for all tasks for each participant in the clinical sample. Black solid line at 0; red dashed line at -1.25.
6.2. NWR

Figure 3 shows age-adjusted z-scores plotted for the two NWR tasks for the children in the TD and the DLD samples. Four children in the DLD sample scored below the −1.25 cut-off on the LS-Tur, but only two children scored below the cut-off on the CL-Tur. Notably, the DLD children were scattered all over the plot, with two children scoring well below the cut-off in both tasks, and one child (BiTurLI-04) scored remarkably well in both tasks. Overall, there was a substantial overlap in performance between the TD and the DLD groups.

Figure 3. Age-adjusted z-scores of the LS-Tur and the CL-Tur NWR task for the children in the DLD sample (diamonds and labels) compared to the children in the TD sample (circles). Dashed lines at −1.25.

There were three children in the TD sample who performed below −1.25 in both tasks. Two of them had normal language development according to parental report and the parents did not express concern about their children’s language. When inspecting the video recordings from the Turkish session it became evident that these children had trouble understanding the task and therefore performed poorly. The third child was reported to have late language development, and the parents voiced concern about the language abilities of their child.

In addition to the two NWR tasks in the Turkish session, the DLD children also did two NWR tasks in the Swedish session: the Swedish version of the cross-linguistic task (CL-Swe) and a language-specific Swedish task (LS-Swe). A table with individual raw scores and proportional scores (%) is available in Table A10 in the Appendix. As evident from Table A10, performance on the two cross-linguistic tasks was similar across language versions (three children’s scores differed by one point, two by two points, and one by three points). The similar performance on both tasks was confirmed by a correlation analysis (df = 4, r = 0.95, p = 0.003). Compared to the cross-linguistic NWR tasks, all children scored proportionally lower (or, at best, the same) on the two language-specific NWR tasks.

6.3. Vocabulary

In Figure 4, age-adjusted z-scores are plotted for Turkish and Swedish vocabulary comprehension (a) and production (b) for the TD and the DLD children. All DLD children scored below the −1.25 cut-off for comprehension in at least one of the languages, but only three scored below in both languages. For production, four children scored below the cut-off in one language, but only one in both languages. Two children in the DLD sample also scored above the cut-off in both languages. The DLD and TD samples showed some overlap, and this overlap was more pronounced for production than for comprehension. Most TD children are in the upper right quadrant with scores above the −1.25 cut-off in both
languages. The ones performing below the cut-off in one language tend to perform above or just below the mean in the other language, i.e., they have uneven proficiency levels.

Next, we wanted to explore whether these uneven vocabulary scores were related to language exposure. LoE Swe (in months) was transformed into age-adjusted z-scores in order to be comparable across the age groups. The TD children that fell below the cut-off in Turkish (Ncomp = 7, Nprod = 10) had generally been exposed to Swedish longer (age-adjusted z-scores: Mcomp = 0.4, SD = 0.9, Mprod = 0.4, SD = 0.8) and had less daily exposure to Turkish (percent exposure: Mcomp = 30%, SD = 10%, Mprod = 30%, SD = 10%) compared to the whole-group average. By contrast, the TD children that fell below the cut-off in Swedish (Ncomp = 13, Nprod = 17) had generally been exposed to Swedish for a shorter time (age-adjusted z-scores: Mcomp = −1.0, SD = 1.2, Mprod = −1.0, SD = 1.0) and had less daily exposure to Swedish (percent exposure: Mcomp = 50%, SD = 20%, Mprod = 40%, SD = 20%) compared to the whole-group average. This finding aligns with the results of Models 2–5, demonstrating an association between language exposure and vocabulary scores.

Figure 4. Age-adjusted z-scores of (a) Turkish and Swedish vocabulary comprehension and (b) Turkish and Swedish vocabulary production for the children in the DLD sample (diamonds and labels) compared to the children in the TD sample (circles). Dashed lines at −1.25.

We then reviewed the exposure measures for the children in the DLD sample with uneven vocabulary scores in order to explore whether exposure patterns could provide an explanation for their results. BiTurLI-01 scored well below the cut-off in both Turkish comprehension and production, but within 1 SD from the mean in Swedish comprehension and production. BiTurLI-01 had 80% daily input in Swedish, which could provide an explanation for the proportionally higher Swedish scores. BiTurLI-02 was relatively stronger in Turkish than Swedish. He had a higher proportion of Turkish (60%) than Swedish (40%) in his daily input, but his Swedish was unexpectedly low considering that onset of exposure to Swedish was at age two. Similarly to BiTurLI-01, BiTurLI-03 scored close to the mean in Swedish, but his Turkish scores fell well below average, despite an age of onset for Turkish at birth and receiving continuous daily exposure to Turkish (50%). BiTurLI-05 scored below average in both Turkish and Swedish, but his Swedish vocabulary scores were much lower than his Turkish vocabulary scores, despite equal daily exposure to Turkish (40%) and Swedish (40%) (the remaining 20% exposure was to Kurdish, recall Table 2). In summary, exposure measures seemingly provide an explanation for uneven vocabulary scores in one of the DLD children, but, otherwise, vocabulary scores were not generally concurrent with exposure measures.

6.4. Narrative Macrostructure

Figure 5 plots age-adjusted z-scores for Turkish and Swedish narrative macrostructure comprehension (a) and production (b) for the TD and the DLD children. In comprehension, five of the DLD children scored below the −1.25 cut-off in one language (Turkish),
but only two did so in both languages. In production, three children performed below the cut-off in one language (again, Turkish), but only two children scored below in both languages. Just like for vocabulary, there was more overlap between the TD and DLD samples in production than in comprehension. Most TD children scored above the cut-off in both languages. There were also a few TD children who fell below the cut-off in one of the languages; most often, they performed above or close to the mean in the other language.

**Figure 5.** Age-adjusted z-scores of (a) Turkish and Swedish narrative macrostructure comprehension and (b) Turkish and Swedish narrative macrostructure production (MAIN story structure) for the children in the DLD sample (diamonds and labels) compared to the children in the TD sample (circles). Dashed lines at −1.25.

In order to explore the effect of vocabulary production on narrative macrostructure for these low-scoring TD children, vocabulary scores were transformed into age-adjusted z-scores in order to be comparable across the age groups. The TD children that fell below the cut-off in Turkish narrative macrostructure (N<sub>comp</sub> = 10, N<sub>prod</sub> = 10) generally had lower Turkish vocabulary production scores (age-adjusted z-scores: M<sub>comp</sub> = −1.5, SD = 1.3, M<sub>prod</sub> = −1.3, SD = 1.5) compared to the whole-group average. Likewise, the TD children that fell below the cut-off in Swedish (N<sub>comp</sub> = 12, N<sub>prod</sub> = 10) generally had lower Swedish vocabulary production scores (age-adjusted z-scores: M<sub>comp</sub> = −1.1, SD = 1.1, M<sub>prod</sub> = −1.3, SD = 1.0) compared to the whole-group average. Thus, scoring below the −1.25 cut-off in narrative macrostructure was associated with low expressive vocabulary skills in that language. This finding aligns with the results of Models 6–9, demonstrating an association between vocabulary production scores and narrative macrostructure scores.

Also, for the DLD children, low vocabulary scores were typically associated with low narrative macrostructure scores in the corresponding language (Turkish or Swedish) and modality (comprehension or production).

### 6.5. NWR, Vocabulary, and Narrative Macrostructure in the DLD Sample: Summary

In summary, the children in the DLD sample scored lower than the children in the TD sample on all tasks. There were, however, individual differences which manifested in different profiles, with individual strengths or weaknesses in one of the languages. For NWR, there was a considerable overlap in performance between the TD and the DLD group, which is notable as NWR is frequently put forward as a clinical marker for DLD and one that also works well for bilingual children. Although the TD and the DLD groups also showed some overlap in vocabulary and narrative macrostructure, there was a clearer difference in performance for Turkish vocabulary comprehension and narrative comprehension than for production, with the DLD children generally performing below the −1.25 z-score cut-off on both comprehension tasks.
7. Discussion

7.1. Factors Affecting Performance on NWR, Vocabulary, and Narrative Macrostructure in the TD Sample

We found that performance increased with age for both NWR tasks, the LS-Tur (Topbaş et al. 2014 [6]) and the CL-Tur (Chiat 2015 [5]), and that these age effects held also when controlling for vocabulary scores. This is consistent with several previous studies showing development with age for NWR tasks in other populations (Kalnak et al. 2014 [7]; Öberg & Bohnacker 2022 [10]; Topbaş et al. 2014 [6]). For both the LS-Tur and the CL-Tur, accuracy decreased as item length (i.e., number of syllables) increased, in line with previous studies using other language versions of the cross-linguistic NWR task, as well as studies using different stimulus items (Boerma et al. 2015 [12]; Öberg & Bohnacker 2022 [10]; Schwob et al. 2021 [9]; Topbaş et al. 2014 [6]). Higher vocabulary scores were associated with better NWR performance, but this effect held only for the LS-Tur task when controlling for chronological age. This finding supports evidence from previous studies demonstrating an effect of vocabulary on language-specific NWR tasks (Öberg & Bohnacker 2022 [10]; Sorenson Duncan & Paradis 2016 [15]; Thordardottir & Brandeker 2013 [17]).

As regards vocabulary, chronological age, length of exposure to Swedish and percent daily exposure (to Turkish/Swedish) explained more than twice as much of the variance in the children’s Swedish comprehension and production scores compared to their Turkish scores. For Turkish comprehension, this result might be explained by a tendency for a ceiling effect, as scores were already high overall among the four-year-olds (M = 51.4, i.e., 86% correct), and individual variation was smaller compared to Turkish production and Swedish comprehension and production scores (recall Table 5). By contrast, for Turkish production, there was no ceiling effect and large individual variations in scores. Thus, for vocabulary in the minority language, particularly expressive vocabulary, crude measures of exposure length or percent daily exposure could not sufficiently explain vocabulary growth. LoE to Swedish added unique explanatory value to both Swedish comprehension and production scores, with an effect size as large as that of chronological age and larger than that of percent daily exposure to Swedish. Interestingly, LoE to the majority language (Swedish) was also a significant negative predictor for minority language (Turkish) production scores, with a somewhat larger effect size than percent daily exposure to Turkish.

For both narrative macrostructure comprehension and story structure production in Turkish and Swedish, we found a clear development with age. Interestingly, exposure measures were not significant predictors for macrostructure in either language or modality; rather, vocabulary production scores emerged as the most important predictor of narrative macrostructure scores, overshadowing chronological age. Thus, expressive narrative macrostructure skills in both Turkish and Swedish seem to be affected by the same factors (namely, vocabulary size and age, but not quantitative measures of language exposure). Additionally, we found that these patterns were also alike for comprehension of narrative macrostructure in the majority language, Swedish (the language of schooling and society) and in the minority language, Turkish (the language of the home). We interpret the finding that both age and vocabulary were strong, uniquely contributing predictors for all macrostructure measures, as follows. Older children are cognitively more mature and likely to have more experience with stories, which makes it easier for them to interpret causal, hierarchical and thematic relations between depicted story events, and relate them to a listener. Yet, for children with only limited lexical skills in the language in question, it is difficult to verbally communicate the plot and point of the story and its macrostructural components to a listener, both when telling the story and when answering comprehension questions (see also, Bohnacker et al. 2020 [73]; 2022 [74]). The proposal that narrative macrostructure skills are easily transferred from the L1 to the L2 and quickly put to use there (Iluz-Cohen & Walters 2012 [56]; Paradis et al. 2021, pp. 170–171 [14];
Pearson 2002 ([35]) would thus have to be qualified and made contingent on sufficient expressive vocabulary in the L2.

The results of this study indicate that depending on the task at hand, different aspects of a bilingual child’s linguistic background and their proficiency in other language domains need to be taken into account when interpreting their performance on language tests. For instance, TD children with low vocabulary scores may be disadvantaged compared to their peers who have high vocabulary scores on language-specific NWR tasks and narrative macrostructure comprehension and production tasks. Vocabulary scores, unlike narrative macrostructure scores, are strongly affected by quantitative measures of exposure (LoE to the majority language and proportion of daily exposure to the language of testing), and in different ways, depending on language status (minority vs. majority) and modality (comprehension vs. production). For example, TD children with longer LoE to the majority language may be clearly disadvantaged (compared to their peers with shorter LoE on minority-language vocabulary production tasks; however, they may not be at a clear disadvantage for minority-language vocabulary comprehension. These insights have implications for the linguistic assessment of bilinguals with suspected DLD.

7.2. NWR, Vocabulary, and Narrative Macrostructure in the DLD Sample

In this section, we discuss the performance of the DLD children on the NWR, vocabulary and narrative tasks.

Concerning *nonword repetition*, there was substantial individual variation, with a lot of overlap between the two groups (TD and DLD), for both the language-independent and the Turkish language-specific NWR task. Only two children in the DLD sample scored below the −1.25 z-score cut-off, and one DLD child scored well below the cut-off in one task, but just above in the other. An additional two DLD children had a large discrepancy (high scores on one NWR task, but low scores on the other task), and one child scored high in both tasks. Neither the CL-Tur (Chiat 2015 [5]) nor the LS-Tur (Topbaş et al. 2014 [6]) were useful for identifying DLD in our sample of Turkish–Swedish-speaking bilinguals. A similar overlap in NWR performance of TD and DLD children has been found for 109 Arabic–Swedish-speaking bilinguals of the same age as the present sample (Öberg & Bohnacker 2022 [10]).

Unlike Boerma et al. (2015 [12]) we did not find the CL-NWR to be a better diagnostic tool than the LS-task. There could be multiple reasons for these divergent findings. One explanation might lie in differences in the lexical phonology of Turkish vs. Dutch, with Turkish syllable structure being less phonologically complex and thus easier to repeat for both DLD and TD children. Another explanation could be that the bilinguals in our study were presented with a language-specific NWR task in their L1 Turkish whereas Boerma et al. (2015 [12]) used a language-specific NWR task in the children’s L2 Dutch. (However, remember that our DLD children were tested on language-specific NWR tasks in both Turkish and Swedish in addition to the Turkish and Swedish versions of the cross-linguistic tasks. Performance was generally lower on both language-specific tasks, even though the Turkish language-specific task is very dissimilar from the LS-Swe task (which includes many clusters and a variety of open and closed syllables) and more similar to the cross-linguistic tasks with regard to phonological setup (due to its lack of consonant clusters and predominance of open syllables.) A third explanation that we find plausible is that of differences in sampling. Boerma et al. (2015 [12]) only assessed the children in their L2 Dutch, and their two groups were heavily preselected. An inclusion criterion for their DLD group was scoring below −2 SD (composite score), or −1.5 SD on two subscales on a standardized test battery. As Boerma et al. acknowledge, their research design may have boosted diagnostic accuracy, as NWR is known to classify TD vs. DLD less well in population-based samples compared to preselected groups (Ellis Weismer et al. 2000 [75]). Our findings are consistent with those of Polišenská et al. (2022 [49]), who reported that the CL-NWR could not reliably discriminate bilingual TD children from their peers with DLD, in a large-scale study combining datasets from 18 research teams in 15 countries.
Our results also align with multiple reports of a higher degree of TD/DLD overlap and poorer NWR diagnostic accuracy for identifying DLD in bilinguals (dos Santos & Ferré 2018 [48]; Schwob et al. 2021 [9]; Thordardottir & Brandeker 2013 [17]). Taken together, our discouraging results regarding the use of NWR to identify DLD in bilinguals can be understood in light of two recent meta-analyses that report evidence for a publication bias for studies finding positive and strong effects of NWR as a clinical marker for DLD in bilinguals (Ortiz 2021 [8]; Schwob et al. 2021 [9]).

Moving on to the vocabulary tasks, all DLD children scored below the cut-off for vocabulary comprehension in at least one of their languages, most often in Turkish, but only three did so in both languages. For vocabulary production, four children scored below the cut-off in one language, but only one in both languages. There was some overlap in individual scores between the TD and the DLD groups, but more so for production than for comprehension. These results do not match those of Khoury Aouad Saliby et al. (2017 [52]) who found less overlap in production scores between the TD and the DLD group in their sample of Arabic/French/English-speaking bilinguals (although they used conceptual scoring and did not factor in age in their analyses). The scores of the TD and the DLD children in the present study overlapped less than those of the TD and DLD Arabic–Swedish-speaking bilinguals of the same age in Öberg & Bohnacker (2022 [10]). Note, however, that the DLD sample in our study was only half as large (N = 6) as that of Öberg and Bohnacker (N = 11). We did not find any evidence for Haman et al.’s (2015, p. 198 [25]; Haman & Łuniewska 2013 [26]) contention that lexical deficits in both languages are a warning sign for DLD.

For narrative macrostructure, five of our six DLD children scored below the cut-off in comprehension in the minority language, Turkish. As mentioned in Section 2.3, there are very few studies comparing MAIN narrative macrostructure in bilingual TD vs. DLD children, and they have only investigated comprehension and production in the L2. Similarly to Boerma et al. (2016 [62]) and Boerma and Blom (2017 [18], 2021 [13]), we found that narrative comprehension differentiated the TD and DLD children better than narrative production. Gagarina et al. (2015, p. 264 [76]) suggested that MAIN can be used for screening and identification of bilinguals at risk of DLD. In our study, which included both languages, narrative comprehension in the minority language (Turkish) was more differentiating than minority language production or majority-language narrative measures.

Interestingly, nonword repetition and narrative skills, the two types of tasks that have widely been proposed as suitable to use in bilingual assessment due to their more general-cognitive and less exposure-dependent nature (e.g., Boerma et al. 2015 [12]; Boerma & Blom 2021 [13]; Paradis et al. 2013 [28], 2021, pp. 170–171 [14]; Thordardottir & Brandeker 2013 [17]), were not the most differentiating tasks in our study. Rather, out of all measures included, CLT vocabulary comprehension scores (particularly in the minority language) seemed to best identify the bilingual children with DLD. The finding that the language comprehension scores better identified the DLD children in our sample does not mean that their problems with comprehension are greater than with production. Rather, the TD children cluster together more tightly for comprehension (and are scattered more for production), which makes low comprehension scores stand out more for the DLD children. Of course, it is difficult to make generalizations from these findings, since our sample of DLD children is small and DLD is a heterogeneous disorder. Still, our findings align with those of Smolander et al. (2021 [51]), who found (although they only investigated the majority language) that vocabulary tests were informative in the assessment and diagnosis of DLD in their sample of Finnish bilinguals, when exposure history was taken into account.

It is noteworthy that the DLD children in the present study, despite extensive and sustained exposure to the minority language, Turkish, had very low vocabulary-comprehension scores in that language, and a majority of them also had very low vocabulary comprehension in the majority language, Swedish, despite extensive exposure to Swedish. By contrast, the TD children scored high on vocabulary comprehension when extensively
exposed to the language in question. This could suggest that bilingual DLD children make less efficient use of the language input they receive (Govindarajan & Paradis 2019 [27]).

7.3. Individual Children in the DLD Sample: Language Scores in Relation to Exposure and Functional Communication

As seen in Section 6.1, there was great variation in the language profiles in the DLD group. Some variation with regards to language proficiency and communication difficulties in the DLD group can be attributed to the heterogeneity of the DLD diagnosis, which means that a child may have individual strengths and weaknesses in one modality (i.e., comprehension or production) and/or in a particular linguistic domain. Another source of variation is language exposure history. In this section, the language scores of the DLD children will be discussed in light of the information obtained from the parent, teacher and SLP interviews, as well as the experimenters’ observations during assessment (summarized in Table A1 in the Appendix). While exposure seemingly provided a partial explanation for the performance of one child, it did not provide an explanation for the poor performance in all children.

BiTurLI-01 (age 6;3) was one of only two DLD children whose scores fell well below the cut-off in both NWR tasks. Her proficiency in Turkish and Swedish was uneven, as reflected in much higher z-scores in Swedish vocabulary and narrative macrostructure, compared to the scores in Turkish. Regular exposure to Swedish started early (between the age of one and two), and daily exposure was mostly in Swedish (80%). This child’s Swedish vocabulary and narrative scores were within 1 SD from the TD mean for her age, and for narrative production even above 1 SD from the mean, but several SDs below the mean in Turkish. If we were to consider only these language scores and exposure measures, we could be led to believe that this was simply a TD Swedish-dominant child as a result of uneven exposure. However, BiTurLI-01 attended a Swedish-medium språkförskola due to severe DLD, where she received a lot of tailored support to develop her Swedish language and communication skills on a daily basis. According to parental report, neither her Turkish nor Swedish were strong, and communicative problems were frequent. The teacher described frequent misunderstandings and peer conflicts in preschool, and the SLP pointed out that progress in therapy was very slow. Thus, the child’s relatively high Swedish vocabulary and narrative scores on the LITMUS tasks did not match the described and observed language and overall communication skills. Here, problems were reported, despite the fact that BiTurLI-01 had an early onset and a large proportion of daily exposure to Swedish and received a lot of language and communication support. The child’s middling performance on the Swedish vocabulary and narrative tasks may thus be partly due to an effect of intensive therapy in Swedish.

BiTurLI-02 (age 5;4) was the only DLD child who performed relatively well in Turkish, as all Turkish vocabulary and narrative tasks were close to the age-group mean score. His Swedish was comparatively weaker, as both vocabulary and narrative macrostructure scores were well below the age-group mean score. Exposure to Swedish had started between age two and three, and daily exposure to Turkish was relatively high (60%). In this case, it was unsurprising that his Turkish scores were high, but his Swedish scores were unexpectedly low, considering that he had regular exposure to Swedish for about three years and attended a Swedish-medium preschool. Interviews with parents, teachers, and the child’s SLP were consistent with his language scores, as BiTurLI-02 was consistently described to have very poor Swedish, and to speak Turkish to interlocutors who did not understand any Turkish. At preschool he constantly sought help from other Turkish-speaking peers, and the preschool teacher reported that the child had great difficulties remembering words in Swedish. The SLP reported that his Swedish was virtually ‘non-existent’ and that he had great difficulties learning new words in Swedish. Despite having age-appropriate scores in Turkish, BiTurLI-02 had great difficulties with functional com-
munication, and unexpectedly large difficulties acquiring Swedish, despite receiving ample input in Swedish. Thus, scoring well in minority-language Turkish tasks disguises the severe problems this child has in his majority language, Swedish.

BiTurLI-03 (age 5;11) had a similar profile to BiTurLI-01, as his Turkish vocabulary and narrative scores were all well below the −1.25 cut-off, but his Swedish scores (save for narrative production) were close to the age-group mean. He had received regular exposure to both Turkish and Swedish since birth, and had a relatively high proportion of daily exposure to Turkish, even at almost six years of age. Despite such ample Turkish input, both the parents and the bilingual Turkish/Swedish-speaking schoolteacher emphasized that BiLI-03’s Turkish was very limited, and that he always answered in Swedish when spoken to in Turkish. Also, the experimenter during the Turkish session reported that the child basically spoke no Turkish during assessment. Similarly to BiTurLI-01, the child’s scoring close to the mean in one language (majority language Swedish) ‘hides’ the severe problems the child has in his other language (minority language Turkish).

BiTurLI-04 (age 4;9) did not have a stronger language, as her vocabulary and narrative scores were all similarly low in both Turkish and Swedish. Exposure to Turkish had started at birth, and exposure to Swedish within the second year of life. Current daily exposure was slightly higher in Turkish (60%) than in Swedish (40%). Additionally, the parents, the preschool teachers, and the SLP reported that the child preferred to speak English. According to the parents, BiTurLI-04 spent a lot of time in front of screens, mostly watching YouTube videos in English. They were also of the opinion that their child did not receive enough ‘proper’ Swedish input at school. The teacher reported that BiTurLI-04 had very limited Swedish and had difficulties with social interaction, and that she was frequently involved in peer conflicts at school.

BiTurLI-05 (age 6;10) scored low in both languages, but particularly low in Swedish comprehension of narratives and vocabulary. This child had Kurdish as his first language, but was also exposed to Turkish and Swedish from an early age (Turkish from age 1, Swedish from age 2). By the time of testing, he had more exposure to Turkish and Swedish (estimated daily exposure was 40% for both languages) than Kurdish (estimated daily exposure was 20%). Thus, despite having similar proportions of exposure to Turkish and Swedish, scores were considerably lower in Swedish than in Turkish. The parents reported that he had a late language development, and that he rarely initiated communication outside the family, due to shyness. The teacher also described him as a quiet child who showed a slow development at school and often needed one-on-one instruction. Similarly to the parents and the teacher, BiTurLI-05’s SLP described him as a quiet boy with pragmatic difficulties and limited conceptual knowledge, who made slow progress in therapy.

Finally, BiTurLI-06 (age 8;1) had a similar profile to BiTurLI-04, who did not score clearly better in one language than the other. BiTurLI-06 was exposed to both Kurdish and Turkish from birth and began to receive regular exposure to Swedish at age three. By the time of testing (at eight years of age), she had proportionally more daily exposure to Swedish (50%) than to Turkish (30%) and Kurdish (20%). The parents originally stated that there were no problems in either language, but later gave contradictory information in the interview, saying that she had poor Swedish comprehension. The parents, the teacher, and the experimenter all described BiTurLI-06 as a social child. However, the teacher reported that she often had difficulties understanding instructions. She had a few friends at school but there were sometimes misunderstandings, which made BiTurLI-06 frustrated when she could not get her message across. According to the SLP, she had difficulties in all areas of language, and often misunderstood simple instructions.

These examples illustrate how individual DLD children can have similar language scores/profiles, despite very different exposure patterns, and they underscore the need for interpreting results from formal language assessment in relation to language exposure and in relation to descriptions and observations of functional language and communication skills in naturalistic contexts.
8. Conclusions and Future Directions

In line with Paradis et al. (2013 [28]) and Govindarajan and Paradis (2019 [27]), we conclude that language scores do not really tell so much on their own, and propose that results from language assessment tasks, even those developed specifically for bilinguals, such as the LITMUS tools, need to be interpreted in light of information about language exposure. This information about language exposure and use can be gathered from questionnaires and interviews with parents and teachers (Paradis et al. 2010 [38]; Tuller 2015 [37]). In addition, conducting interviews with parents and teachers in conjunction with formal assessment makes it possible to obtain a picture of the children’s functional language skills and can identify communication difficulties that are hard to catch with standardized tests (Öberg & Bohnacker 2022 [10]). We see that an increased use of mixed-methods research designs that combine quantitative and qualitative methods may be fruitful in the advancement of investigating language assessment in bilingual children, and in developing assessment methods that are valid for identifying DLD in bilingual populations (e.g., Glogowska 2011 [77]).

The DLD children in the present study generally had low vocabulary and narrative macrostructure comprehension scores in the home/minority language, despite extensive and sustained exposure to that language. Many also had comparably low vocabulary and narrative comprehension scores in the majority language, despite extensive exposure to the majority language. Whilst one should be cautious when making generalizations based on a DLD sample as limited in size as ours, low comprehension scores in a bilingual child, despite continuous, extensive exposure to that language from an early age, might be considered a warning sign for DLD. Recent advances suggest that the language skills of bilingual TD and DLD populations may be differently affected by language exposure and cognitive abilities, such as analytical reasoning and short-term memory (Paradis 2023 [78]), see also, Smolander et al. (2021 [51]) on input and vocabulary development in the L2, and Govindarajan and Paradis (2019 [27]) on input and narrative production in the L2. However, more research needs to be undertaken before the association between language exposure, cognition and language abilities in bilingual TD and DLD populations is well understood. Future work is therefore required to establish whether language scores of children with DLD are affected by background factors such as input, cognitive abilities and other language measures in the same way as for their TD peers, and whether these (non-)differences are dependent on the measure, i.e., the language domain, the modality (comprehension or production) and the minority or majority language.

Author Contributions: Conceptualization, U.B. and L.Ö.; methodology, U.B. and L.Ö.; validation, U.B. and L.Ö.; formal analysis, U.B. & L.Ö.; investigation, U.B. and L.Ö.; resources, U.B.; data curation, U.B. and L.Ö.; writing—original draft preparation, U.B. and L.Ö.; writing—review and editing, visualisation, L.Ö.; supervision, U.B.; project administration, U.B.; funding acquisition, U.B. All authors have read and agreed to the published version of the manuscript.

Funding: We gratefully acknowledge research grants to the second author from the Swedish Research Council [VR 421-2013-1309] and the Bank of Sweden Tercentenary Foundation [P19-0644:1].

Institutional Review Board Statement: The study has been carried out in accordance with the Declaration of Helsinki and Swedish regulations concerning data protection and adheres to the local university ethical code of conduct. At the time when this project started, institutional review structures were not yet in place at the Faculty of Languages. However, rigorous procedures were put in place in the project to preserve the confidentiality and anonymity of participating subjects and to ensure that the research was carried out in an ethical way.

Informed Consent Statement: Families received oral and written information about the project in both Turkish and Swedish. Informed written consent was obtained from the parents of all children involved in the study before data collection. Children gave oral assent. Participation could be discontinued at any time.

Data availability statement: The present study is part of a larger, ongoing research project on the multilingual development of Turkish/Swedish and Arabic/Swedish. The original raw dataset cannot
be shared in full, as this would breach participant confidentiality and GDPR guidance. For data supporting the reported results, please consult the authors.

Acknowledgments: The authors would like to thank the members of the BiLI-TAS team at Uppsala University who were involved in data collection, transcription and scoring: Buket Öztekin, Josefín Lindgren, Sibylle Dillström, Karin Koltay, Ercan Aras and Berrak Uluer. A special thank you goes to our former colleague, Buket Öztekin, who collected a large number of the Turkish data, interviewed the parents and carried out many of the initial data analyses, which we then could build on. We would also like to thank all the Turkish/Swedish-speaking children and their families for participating in this study; without them, there would be no data and no results. Teşekkürler!

Conflicts of Interest: Neither author declares a conflict of interest.

Appendix A

Table A1. Language abilities and behavior of the DLD children as described by parents, teachers, SLPs and experimenters.

<table>
<thead>
<tr>
<th>Parents</th>
<th>Teacher</th>
<th>SLP</th>
<th>Experimenters’ Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BiTur LI-01</strong></td>
<td><strong>behavior/comments</strong></td>
<td><strong>language</strong></td>
<td>Tur: −comp, −prod Swe: so-so</td>
</tr>
<tr>
<td>Tur is very limited; answers in Swe when spoken to in Tur</td>
<td></td>
<td></td>
<td>engages in communication, but frequent misunderstandings/conflicts; uses a lot of Eng</td>
</tr>
<tr>
<td><strong>BiTur LI-02</strong></td>
<td><strong>behavior/comments</strong></td>
<td><strong>language</strong></td>
<td>Tur: +comp, +prod Swe: so-so comp, −prod</td>
</tr>
<tr>
<td>LLD; Tur is very limited; had hearing problems in the past</td>
<td></td>
<td></td>
<td>seeks help from Tur-speaking peers when he does not know a word in Swe; speaks Tur with non-Tur-speakers; has trouble ‘remembering’ info in Swe</td>
</tr>
<tr>
<td><strong>BiTur LI-03</strong></td>
<td><strong>behavior/comments</strong></td>
<td><strong>language</strong></td>
<td>Tur: +comp, −prod Swe: +comp, +prod</td>
</tr>
<tr>
<td>LLD in Tur; had hearing problems until age 3; always answers in Swe when spoken to in Tur</td>
<td></td>
<td></td>
<td>polite and ‘nice’ child; language skills have developed a lot; functions well in the group; can follow instructions; Tur is limited acc to Tur-speaking teacher</td>
</tr>
<tr>
<td><strong>BiTur LI-04</strong></td>
<td><strong>behavior/comments</strong></td>
<td><strong>language</strong></td>
<td>Tur: +comp, +prod Sve: +comp, −prod Eng better than Sve</td>
</tr>
<tr>
<td>not enough ‘proper’ Swe input at school; prefers to speak Eng, spends a lot of time with digital devices in Eng</td>
<td></td>
<td></td>
<td>prefers Eng; very limited Swe; difficulties with social interaction; has jealousy issues and obsessions, frequent peer conflicts; needs one-on-one instructions</td>
</tr>
<tr>
<td>Language</td>
<td>Tur: ~comp, ~prod</td>
<td>Swe: +comp, +prod</td>
<td>Swe: ~comp, ~prod</td>
</tr>
<tr>
<td>----------</td>
<td>-----------------</td>
<td>------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>BiTur LI-05</td>
<td>LLD; rarely initiates communication outside the family, is shy</td>
<td>often quiet; rarely shows facial expressions; needs one-on-one instructions; slow development, teacher thinks child would be better off in a smaller group/class with more tailored support</td>
<td>pragmatic difficulties; often quiet, slow progress in therapy; limited conceptual knowledge</td>
</tr>
<tr>
<td>BiTur LI-06</td>
<td>LLD; social, likes to play at school, no problems at home</td>
<td>has trouble understanding instructions; becomes angry when she cannot make herself understood; has a few friends but there are sometimes misunderstandings</td>
<td>has difficulties in all areas of language, misunderstands simple questions; SLP has not been able to conduct an assessment in Tur, since the parents think that the child only has difficulties in Swe</td>
</tr>
</tbody>
</table>

**Note.** LLD = Late language development; ‘comp’ = language comprehension; ‘prod’ = language production; ‘~’ indicates relative weaknesses, ‘~’ indicates relative strengths in language ability.

**Table A2.** All items in the NWR tasks: LS-Tur (language-specific Turkish), CL-Tur (cross-linguistic Turkish), LS-Swe (language-specific Swedish) and CL-Tur (cross-linguistic Swedish).

**Table A3.** Summary of logistic mixed-effects regression model (Model 1): accuracy of CL-Tur vs. LS-Tur task.
### Table A4. Summary of linear regression models: Turkish vocabulary comprehension and production.

<table>
<thead>
<tr>
<th>Model 2 (Turkish Vocabulary Comprehension)</th>
<th>Model 3 (Turkish Vocabulary Production)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>SE (B)</td>
</tr>
<tr>
<td>----</td>
<td>--------</td>
</tr>
<tr>
<td>Intercept</td>
<td>54.49***</td>
</tr>
<tr>
<td>Age</td>
<td>0.21***</td>
</tr>
<tr>
<td>LoE Swe</td>
<td>–0.07</td>
</tr>
<tr>
<td>Daily exp Tur</td>
<td>0.07**</td>
</tr>
</tbody>
</table>

R² (adj.) = 0.20, F(3,97) = 9.13, p < 0.001

#### Note.
VIF values were 1–2 for all predictors, indicating low levels of collinearity. *** p < 0.001, ** p < 0.01.

### Table A5. Summary of linear regression models: Swedish vocabulary comprehension and production.

<table>
<thead>
<tr>
<th>Model 4 (Swedish Vocabulary Comprehension)</th>
<th>Model 5 (Swedish Vocabulary Production)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>SE (B)</td>
</tr>
<tr>
<td>----</td>
<td>--------</td>
</tr>
<tr>
<td>Intercept</td>
<td>50.17***</td>
</tr>
<tr>
<td>Age</td>
<td>0.24***</td>
</tr>
<tr>
<td>LoE Swe</td>
<td>0.16***</td>
</tr>
<tr>
<td>Daily exp Swe</td>
<td>0.07*</td>
</tr>
</tbody>
</table>

R² (adj.) = 0.49, F(3,97) = 32.40, p < 0.001

#### Note.
VIF values were 1–2 for all predictors, indicating low levels of collinearity. *** p < 0.001, * p < 0.05.

### Table A6. Summary of linear regression models: Turkish narrative macrostructure comprehension and production.

<table>
<thead>
<tr>
<th>Model 6 (Turkish Narrative Comprehension)</th>
<th>Model 7 (Turkish Narrative Production)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>SE (B)</td>
</tr>
<tr>
<td>----</td>
<td>--------</td>
</tr>
<tr>
<td>Intercept</td>
<td>12.97***</td>
</tr>
<tr>
<td>Age</td>
<td>0.09**</td>
</tr>
<tr>
<td>LoE Swe</td>
<td>0.01</td>
</tr>
<tr>
<td>Daily exp Tur</td>
<td>0.00</td>
</tr>
<tr>
<td>Tur vocabulary</td>
<td>0.27***</td>
</tr>
</tbody>
</table>

R² (adj.) = 0.53, F(4,96) = 29.05, p < 0.001

#### Note.
‘Tur vocabulary’ denotes vocabulary production scores. VIF values were 1–2 for all predictors, indicating low levels of collinearity. *** p < 0.001, ** p < 0.01.

### Table A7. Summary of linear regression models: Swedish narrative macrostructure comprehension and production.

<table>
<thead>
<tr>
<th>Model 8 (Swedish Narrative Comprehension)</th>
<th>Model 9 (Swedish Narrative Production)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>SE (B)</td>
</tr>
<tr>
<td>----</td>
<td>--------</td>
</tr>
<tr>
<td>Intercept</td>
<td>13.56***</td>
</tr>
<tr>
<td>Age</td>
<td>0.13***</td>
</tr>
<tr>
<td>LoE Swe</td>
<td>–0.01</td>
</tr>
<tr>
<td>Daily exp Swe</td>
<td>–0.01</td>
</tr>
<tr>
<td>Swe vocabulary</td>
<td>0.27***</td>
</tr>
</tbody>
</table>

R² (adj.) = 0.64, F(4,96) = 45.79, p < 0.001

#### Note.
‘Swe vocabulary’ denotes vocabulary production scores. VIF values were 1–2 for all predictors, indicating low levels of collinearity. *** p < 0.001.
Table A8. Mean scores, standard deviations (SDs) and ranges for Turkish and Swedish narrative macrostructure comprehension (max score = 10) and story structure production (max score = 17) by age group (MAIN Cat/Dog).

<table>
<thead>
<tr>
<th></th>
<th>4 Years</th>
<th>5 Years</th>
<th>6 Years</th>
<th>7 Years</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Turkish narrative comprehension</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>5.8 (2.2)</td>
<td>7.0 (2.7)</td>
<td>7.8 (2.2)</td>
<td>7.9 (2.0)</td>
<td>7.2 (2.4)</td>
</tr>
<tr>
<td>Range</td>
<td>2–10</td>
<td>0–10</td>
<td>1–10</td>
<td>1–10</td>
<td>0–10</td>
</tr>
<tr>
<td><strong>Turkish narrative production</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>3.2 (1.8)</td>
<td>4.3 (2.4)</td>
<td>4.6 (1.7)</td>
<td>6.2 (2.1)</td>
<td>4.6 (2.3)</td>
</tr>
<tr>
<td>Range</td>
<td>0–6</td>
<td>1–10</td>
<td>1–7</td>
<td>0–9</td>
<td>0–10</td>
</tr>
<tr>
<td><strong>Swedish narrative comprehension</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>6.0 (2.6)</td>
<td>7.1 (2.3)</td>
<td>8.6 (1.1)</td>
<td>8.8 (1.3)</td>
<td>7.7 (2.2)</td>
</tr>
<tr>
<td>Range</td>
<td>1–9</td>
<td>1–10</td>
<td>6–10</td>
<td>6–10</td>
<td>1–10</td>
</tr>
<tr>
<td><strong>Swedish narrative production</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>2.8 (1.7)</td>
<td>4.7 (1.9)</td>
<td>5.9 (1.7)</td>
<td>7.1 (2.1)</td>
<td>5.2 (2.5)</td>
</tr>
<tr>
<td>Range</td>
<td>0–6</td>
<td>1–8</td>
<td>2–9</td>
<td>2–12</td>
<td>0–12</td>
</tr>
</tbody>
</table>

Table A9. Mean scores, standard deviations (SDs) and ranges for Turkish and Swedish narrative macrostructure comprehension (max score = 10) and story structure production (max score = 17) by age group (MAIN Baby Birds/Baby Goats).

<table>
<thead>
<tr>
<th></th>
<th>4 Years</th>
<th>5 Years</th>
<th>6 Years</th>
<th>7 Years</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Turkish narrative comprehension</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>4.4 (2.4)</td>
<td>6.2 (2.4)</td>
<td>6.3 (2.5)</td>
<td>7.2 (2.5)</td>
<td>6.1 (2.6)</td>
</tr>
<tr>
<td>Range</td>
<td>0–9</td>
<td>0–9</td>
<td>1–10</td>
<td>0–10</td>
<td>0–10</td>
</tr>
<tr>
<td><strong>Turkish narrative production</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>3.3 (2.6)</td>
<td>4.7 (2.7)</td>
<td>5.6 (2.4)</td>
<td>6.3 (2.4)</td>
<td>5.1 (2.7)</td>
</tr>
<tr>
<td>Range</td>
<td>0–8</td>
<td>0–9</td>
<td>1–9</td>
<td>1–11</td>
<td>0–11</td>
</tr>
<tr>
<td><strong>Swedish narrative comprehension</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>3.6 (2.3)</td>
<td>5.3 (2.3)</td>
<td>7.1 (1.7)</td>
<td>8.4 (1.4)</td>
<td>6.2 (2.6)</td>
</tr>
<tr>
<td>Range</td>
<td>0–8</td>
<td>0–10</td>
<td>3–10</td>
<td>4–10</td>
<td>0–10</td>
</tr>
<tr>
<td><strong>Swedish narrative production</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>3.3 (2.8)</td>
<td>4.4 (2.7)</td>
<td>5.6 (2.1)</td>
<td>6.5 (2.1)</td>
<td>5.0 (2.7)</td>
</tr>
<tr>
<td>Range</td>
<td>0–9</td>
<td>0–9</td>
<td>1–10</td>
<td>1–11</td>
<td>0–11</td>
</tr>
</tbody>
</table>

Table A10. Individual NWR raw scores and percentages for the DLD children.

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>LS-Tur</th>
<th>CL-Tur</th>
<th>CL-Swe</th>
<th>LS-Swe</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiTurLI-01</td>
<td>6;3</td>
<td>4 (25%)</td>
<td>4 (25%)</td>
<td>6 (38%)</td>
<td>6 (25%)</td>
</tr>
<tr>
<td>BiTurLI-02</td>
<td>5;4</td>
<td>3 (19%)</td>
<td>5 (31%)</td>
<td>7 (44%)</td>
<td>6 (25%)</td>
</tr>
<tr>
<td>BiTurLI-03</td>
<td>5;11</td>
<td>7 (44%)</td>
<td>14 (88%)</td>
<td>13 (81%)</td>
<td>12 (50%)</td>
</tr>
<tr>
<td>BiTurLI-04</td>
<td>4;9</td>
<td>10 (63%)</td>
<td>14 (88%)</td>
<td>11 (69%)</td>
<td>14 (58%)</td>
</tr>
<tr>
<td>BiTurLI-05</td>
<td>6;10</td>
<td>3 (19%)</td>
<td>9 (56%)</td>
<td>8 (50%)</td>
<td>9 (38%)</td>
</tr>
<tr>
<td>BiTurLI-06</td>
<td>8;1</td>
<td>11 (69%)</td>
<td>11 (69%)</td>
<td>10 (63%)</td>
<td>7 (29%)</td>
</tr>
</tbody>
</table>

Note. LS-Tur = Language-Specific Turkish task (max score = 16), CL-Tur = Cross-Linguistic Turkish task (max score = 16), CL-Swe = Cross-Linguistic Swedish task (max score = 16), LS-Swe = Language-Specific Swedish task (max score = 24).

Notes

1. DLD (Developmental language disorder), also known as (Specific) language impairment (SLI) or Primary language impairment, manifests as a pronounced deficit in the development of language skills, which cannot be attributed to intellectual disability, hearing impairment, medical syndromes or neurological disorders (Bishop 1997, pp. 21–23 [79]; Leonard 2014, p. 3 [20]). It negatively affects children’s oral communication, literacy and advancement in school (Norbury et al. 2016 [80]). DLD is a heterogeneous disorder where deficits in a particular language domain may be more pronounced for some individuals than for others.

2. Nonword repetition and learning new words both require the ability to temporarily store phonological representations. The phonological loop enables individuals to quietly rehearse linguistic material that is otherwise subject to rapid decay in short-
term memory, and is therefore thought to play an important role particularly in the early stages of language acquisition (Gathercole 2006 [81]).

3. When suspecting DLD, it is otherwise recommended that both languages of a bilingual child be assessed (Kohnert 2010 [82]; Peña, Bedore & Kester 2016 [50]).

4. In addition, several papers (e.g. Abed Ibrahim and Hamann 2017 [83]; Hamann and Abed Ibrahim 2017 [84]; de Almeida et al. 2017 [85]; and Tuller et al. 2018 [86]) report on NWR in subsamples of the same participant groups as Chilla et al. (2021 [47]).

5. Moreover, Öberg (2020 [87]) compared these children’s performance on near-identical versions of the language-independent NWR task (2-to-5 syllables), once administered in Swedish and once administered in Arabic (with slightly different phoneme realizations). Performance was similar but slightly depressed for the Arabic version, mainly due to some item effects.

6. Khoury et al. (2017 [52]) chose to implement a conceptual vocabulary scoring procedure for CLT production (expressive vocabulary) because of widespread code-mixing in Lebanese society and in the children’s immediate environment. Thus, Lebanese Arabic, French, and English words corresponding to the CLT target items were scored as correct on the CLT-LB (i.e. Lebanese Arabic) picture naming task.

7. Khoury et al. (2017 [52]) also mentioned that certain types of incorrect CLT responses were more frequent in the DLD than the TD group, such as I don’t know, no response, onomatopoeic answers, and gestures.

8. MAIN includes four picture-based fictional stories carefully designed to be comparable in length and macrostructural components (see Sections 4.2.3 and 4.3), standardized prompts, comprehension questions and scoring protocols for macrostructure (for details see Gagarina et al. 2012 [59], 2019 [36]; Bohnacker & Gagarina 2020a [68]). MAIN can be used to assess story generation (‘telling’), story retelling (after listening), and story comprehension (where the child is probed to make inferences from the pictures and verbalize this understanding).

9. Recall that Boerma et al.’s DLD sample consisted of children in special education who performed at least 2 SD below the mean on a standardized Dutch language test or at least 1.5 SD below the mean on two subscales. This preselection might have enlarged the difference between the DLD and TD samples.

10. As mentioned in Sections 2.2 and 2.3, for 108 of these participants, some vocabulary and macrostructure results have been reported previously in a doctoral thesis (Öztekin 2019 [53]). For the present study, the dataset has been considerably expanded and it is explored from different angles and with multivariate statistics.

11. One child was accidentally seen twice, once at age 4 (when testing was terminated early) and again one year later at age 5 (testing sessions completed). Only the data at age 5 were included in the study.

12. One child, BiTurLI-03, was 5;11 at the first testing (in Turkish), but had turned 6 by the second testing. His scores are compared to the five-year-olds in the TD sample.

13. Another vowel harmony rule only applies to high vowels, making them agree in rounding with the preceding vowels. In many lexical borrowings, however, vowel harmony does not apply.

14. According to a large-scale norming study, monolingual TD children generally master Turkish phonology early, i.e., before age 3;6 (Topbaş 1997 [88], 2006 [89]). The norming study also provides reference data for the typical age of acquisition of individual phonemes.

15. In a few cases, headphones were not used, and the audio was played from the laptop speakers.

16. For neither NWR task was there any correlation with language exposure or SES (parental education.

17. A model with Turkish vocabulary production yielded equivalent results.

18. A model with Turkish vocabulary comprehension yielded equivalent results.

19. A model with Swedish vocabulary comprehension yielded equivalent results.
Note that this particular child was described by both parents, the teacher, and the SLP as having severe communication difficulties (see Table A1 in the Appendix). The experimenter in the Swedish session also reported that the child often repeated the experimenter’s utterances, seemingly without understanding their meaning. This ‘echoing’ behavior could provide an explanation for BiTurLI-04’s above-average performance on both NWR tasks.

The same pattern was found for the Turkish comprehension scores.

This also held for the Swedish comprehension scores.

References


63. Topbaş, S. Developing assessment tools for the identification of language impairments in monolingual and multilingual children speaking Turkish. Presented at the Turkish and Multilingualism Workshop, Uppsala University, Uppsala, Sweden, 10 June 2015.


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.