ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

# Disentangling the Grammar of 3- to 6-Year-Old Dutch Children With a Developmental Language Disorder

**Anouk Scheffer,[a,b] Brigitta Keij,[a] Britt Hakvoort,[a] Esther Ottow,[a] Ellen Gerrits,[b,c] and Frank Wijnen[b]**

[a] Royal Dutch Auris Group, Rotterdam, the Netherlands [b] Utrecht Institute of Linguistics OTS, Utrecht University, the Netherlands [c] Research Center for Healthy and Sustainable Living, HU University of Applied Sciences Utrecht, the Netherlands

ARTICLE INFO

Article History:
Received November 10, 2021
Revision received January 19, 2022
Accepted July 30, 2022

Editor-in-Chief: Stephen M. Camarata
Editor: Mary Alt

https://doi.org/10.1044/2022_JSLHR-21-00598

ABSTRACT

**Purpose:** Children with a developmental language disorder (DLD) are often delayed in their grammatical development. This is suggested to be the most important characteristic and clinical marker of DLD. However, it is unknown if this assumption is valid for young children, in the earliest stages of grammatical development. For this reason, this study investigates the complexity, diversity, and accuracy of the grammatical repertoires of 3- to 6-year-old Dutch children with DLD, in comparison to that of typically developing (TD) children matched on grammatical level.

**Method:** Language samples of 59 children (29 children with DLD and 30 TD children) were analyzed using multiple measures of grammatical complexity, diversity, and accuracy. The TD children and children with DLD were language-matched on their grammatical development using the levels of the Dutch version of the Language Assessment, Remediation, and Screening Procedure, the Taal Analyse Remediëring en Screening Procedure (TARSP; Schlichting, 2017). Thus, the children with DLD were significantly older than the TD children (respectively DLD age range: 2;7–5;4 [years;months], $M_{\text{age}}$ = 4;1; and TD age range: 2;0–3;9, $M_{\text{age}}$ = 2;9).

**Results:** The results show that children with DLD are comparable to language-matched TD children in their grammatical accuracy and diversity, but that they produce less complex utterances.

**Conclusions:** The results indicate that children with DLD lag behind in their grammatical complexity as compared to language-matched TD children. The results also suggest that grammatical TARSP level is not sufficiently informative for selecting treatment goals. Instead, the results underline the importance of conducting language sample analyses, with special reference to the complexity of the utterances of a child with DLD.

Children with a developmental language disorder (DLD) have a language production and/or comprehension impairment that cannot be related to sensory, cognitive, or neurological deficits, an unfavorable psychological condition, or insufficient language input (Leonard, 2014). They form a heterogeneous population, both with regard to severity of their deficits and the linguistic domains affected. Despite this heterogeneity, virtually all children with DLD have grammatical difficulties, and these difficulties are, therefore, considered as a core feature of DLD (Leonard, 2014).

Grammatical performance of Dutch preschool-age children can not only be examined with standardized language tests but also by analyzing spontaneous language. In a language sample analysis (LSA), the language of a child is examined in a naturalistic context, and it is therefore often considered to have better ecological validity than standardized

Correspondence to Anouk Scheffer: a.scheffer@auris.nl. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

language tests (e.g., Heilmann et al., 2010; Pavelko et al., 2016). Moreover, it is a crucial first step for determining goals for grammatical interventions (Verbeek et al., 2007).

An LSA method, commonly used in clinical settings in the Netherlands, is Taal Analyse Remediëring en Screening Procedure (TARSP; Schlichting, 2017), a Dutch adaptation of the English Language Assessment, Remediation, and Screening Procedure (LARSP; Crystal et al., 1976). In TARSP, the grammatical structures produced by a child with DLD are compared to norm scores, which are based on the developmental trajectory of typically developing (TD) children (see http://www.liesbethschlichting.nl/fit/). TARSP partitions grammatical development into six stages, in each of which particular grammatical structures are expected to be produced by 1;6- (years;months) to 4;0-year-old TD children.[1] The grammatical structures are divided into structures on the level of main clauses (declarative, interrogative, and imperative clauses), phrases, and words. A child is assigned to a level corresponding to one of these stages based on the structures they produced. We use TARSP stage to refer to a stage on the TARSP profile chart containing specific structures, whereas we use level to refer to a child's grammatical developmental level as determined with the TARSP LSA procedure.

In this study, we compared the grammatical production patterns of 3- to 6-year-old children with DLD to the production patterns of TD children matched on TARSP level. The production patterns were compared on three grammatical dimensions; complexity, diversity, and accuracy. If the grammatical production patterns of children with DLD differ from the patterns of TARSP-level–matched TD children, this suggests that assigning a grammatical level to a child is not sufficient for selecting grammatical treatment goals, but that a detailed analysis of the grammatical structures a child does and does not produce is needed.

## Using TARSP for Selecting Treatment Goals

According to TARSP guidelines, at least 5% of the total number of utterances a child produces should be a sentence structure that the TARSP profile chart associates with a specific stage,[2] to assign the child to the level corresponding to that stage. The highest stage that meets this 5% criterion corresponds to the level of the child. For example, a child's grammatical level is 3 if the child produces grammatical structures associated with TARSP Stages I, II, III, and IV, and 5% of the sentence structures are Stage-III structures and 2% are Stage-IV structures. Because of this 5% rule, many differences between children's use of TARSP structures are possible, even if they are assigned to the same TARSP level.

Speech-language pathologists (SLPs) use the typical developmental trajectory and sequence in stages of TARSP to select goals for interventions. SLPs select TARSP structures that they judge to be important for reinforcing a child's functional communication (Klatte et al., 2022), and they select these goals in different ways. Some SLPs look closely at the structures a child does and does not produce, and select unused structures as treatment goals. Others select goals from the TARSP stage corresponding to a child's level or from one stage higher. Comparing children with DLD to norm scores based on the typical developmental sequence, however, is based on the idea that children with DLD are delayed in their grammatical development. However, thus far, it is not clear whether the grammatical production repertoire of Dutch preschool-age children with DLD is comparable to the production patterns of TD children at similar grammatical levels. School-age Dutch-speaking children with and without DLD matched on grammatical level do have different production patterns regarding, for instance, personal pronouns (Bol & Kasparian, 2009), subject–verb agreement (de Jong, 1999; Hammer et al., 2014; Spoelman & Bol, 2012), and argument structures (Bol & Kuiken, 1990; de Jong, 1999). In other words, school-age children with DLD and TD children that are assigned to the same grammatical level may actually differ in their grammatical repertoire. It is plausible that similar differences are already present at a younger age. We did not find studies that examined the spontaneous language of younger, preschool-age Dutch children with DLD thoroughly.

## Language-Matching Between Dutch Children With DLD and TD Children

In most studies examining differences between the grammatical production patterns of children with and without DLD, children with DLD were matched with TD children on mean length of utterance (MLU) in morphemes (e.g., Bol & Kasparian, 2009; de Jong, 1999; Spoelman & Bol, 2012; Zwitserlood et al., 2015). However, MLU does not provide information on the type of grammatical structures children produce.[3] Consequently, MLU cannot be used by SLPs in making predictions of a child's grammatical production patterns or in selecting more detailed goals for grammatical interventions.

[1] The 2017 edition of the Taal Analyse Remediëringen en Screening Procedure, Schlichting (TARSP) contains an additional part with grammatical structures for children until 6 years of age, but this is not used in this study.

[2] This calculation of TARSP level is only part of TARSP, not of the original, English Language Assessment, Remediation, and Screening Procedure procedure.

[3] For Dutch, there are no normative data that relate MLU with production of specific grammatical (TARSP) structures.

TARSP (Schlichting, 2017), on the other hand, indicates a general level of grammatical ability and provides insight into the grammatical structures a child produces. Like MLU, a child's level can be seen as a general indicator of his/her grammatical development. However, a TARSP level is more informative than MLU, as it indicates the specific structures a child should (be able to) produce; the structures in the TARSP stage corresponding to a child's level, as well as those associated with lower TARSP stages.

As the TARSP instrument is based on the developmental trajectory of TD children, it assumes that preschool-age children with DLD are delayed in their grammatical development. This is stated in the TARSP guidelines: "the expressive language of most children with language disorders can be seen as delayed and therefore comparable to the language of younger typically developing children" (Schlichting, 2017, p. 7). However, the studies summarized above indicate qualitative differences in the grammatical repertoire between older children with and without DLD at the same grammatical levels. As it is not clear whether the grammatical production patterns of Dutch preschool-age children with and without DLD at the same TARSP level are comparable, this is examined in this study. Comparisons between children with DLD and language-matched TD children can show whether children from these two groups at similar grammatical levels also have similar grammatical production patterns, or, alternatively, that grammatical levels do not provide sufficient information on what children actually produce.

## Grammatical Complexity

We compare the complexity, diversity, and accuracy of the grammatical production patterns of children with and without DLD at the same language level. We consider utterances as more complex when structures from higher TARSP stages (i.e., typically later-acquired structures) are produced. If a child can produce more complex structures, they can convey more (complex) messages, which likely results in better functional communication.

## Grammatical Diversity

Grammatical diversity can be defined in two ways: (a) the number of different structures a child is able to use and (b) how a child can "fill" a grammatical structure with different word combinations in different contexts, for example, using the structure "subject + verb" in combinations as "I eat," "Mommy read," and "Bear walks" instead of using the structure multiple times within the same combination of words. Grammatical diversity could provide valuable insight into the grammatical skills of children: The more different grammatical structures a child can produce, the more different messages they can convey, and the more likely it is that they can use grammatical structures in creative ways (Hadley et al., 2018).

## Grammatical Accuracy

Grammatical accuracy is defined as how many grammatical structures children can produce correctly. TARSP does not include accuracy; structures are scored irrespective of whether they are produced correctly. Examining grammatical production patterns of children using these three dimensions and examining how they are related has not been done before. Consequently, it is not clear how young children with DLD and language-matched TD children differ on these dimensions and how the dimensions interact.

## This Study

In this study, the grammatical production patterns, as analyzed with TARSP, of 3- to 6-year-old children with DLD are compared to the production patterns of TD children matched on TARSP level. Our comparisons focus on the grammatical complexity and accuracy of their utterances, as well as the diversity of the structures in their grammatical repertoire. Taking this into account, these three dimensions provide a comprehensive overview of children's grammatical production patterns. Differences on one dimension would not necessarily lead to differences on the other dimensions. For instance, it is possible that the two groups of children produce utterances of similar complexity, but that children with DLD show less structural diversity (indicating a smaller grammatical repertoire), or that their use of certain structures is less accurate.

Comparing children with matching TARSP levels on their use of TARSP structure types may seem a circular approach. However, because of the 5% rule for assigning a level to a child's verbal output, many differences between children's use of TARSP structures are possible if they have matching levels.

If the grammatical production patterns are very similar for children with and without DLD, this result would validate TARSP as an instrument for monitoring the grammatical development of children with DLD. However, the literature above suggests that it is unlikely that the grammatical production patterns of children with DLD match those of TD children at the same grammatical level. Children with DLD might use different grammatical structures than TD children at the same level, or they might show less variation in the structures they produce. Another possibility is that children with DLD are still using structures from lower TARSP stages more often than TD children. If the grammatical production patterns of children with DLD differ from the patterns of TARSP-level–matched TD children, it would suggest that knowing a child's grammatical level is not sufficient for selecting grammatical treatment goals.

**Table 1.** Sample size, number of girls, number of multilingual children, mean age (standard deviation), and age range of children with developmental language disorder (DLD) and typically developing (TD) children per level.

| | DLD | | | | | TD | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Level | *n* | Number of girls | Number of multilingual children | $M_{age}$ (y;m) (*SD*) | Age range | *n* | Number of girls | Number of multilingual children | $M_{age}$ (y;m) (*SD*) | Age range |
| 3 | 9 | 0 | 1 | 3;5 (0;6) | 2;7–4;2 | 10 | 5 | 0 | 2;4 (0;4) | 2;0–2;11 |
| 4 | 10 | 4 | 0 | 4;3 (0;8) | 3;6–5;3 | 10 | 5 | 0 | 2;8 (0;4) | 2;3–3;4 |
| 5 | 10 | 3 | 1 | 4;6 (0;8) | 3;5–5;4 | 10 | 3 | 0 | 3;3 (0;4) | 2;11–3;9 |

*Note.* y = years; m = months.

The aim of this study is to contribute to a deeper understanding of the grammatical development of preschoolers with DLD and of differences and similarities in the grammatical production repertoires of children at the same grammatical level. We compared the grammatical production patterns of 3- to 6-year-old Dutch children with DLD and TD children matched on their grammatical developmental level as measured with TARSP.

# Method

We compared utterances from previously collected language samples of children with DLD and TD children with matching grammatical levels. We compared these utterances on the same grammatical dimensions: complexity, diversity, and accuracy.

## Participants

All samples of Dutch TD children were retrieved from the CHILDES database (MacWhinney, 2000). These samples were part of studies by Bol and Kuiken (1990), van Kampen (2009), Bol (1995; Groningen corpus), and Elbers and Wijnen (1992; Wijnen corpus). The samples of Dutch children with DLD were collected by Zwitserlood (2019), Bruinsma et al. (2020), Boerma et al. (2020), and Bol and Kuiken (1990), the latter available through CHILDES. The recording settings were play situations with at least one adult present. The samples of the TD children were recorded at home, whereas the samples of the children with DLD were recorded at their schools.

There were 59 children in total, 29 children with DLD and 30 TD children. The groups of children with and without DLD were group-matched on their TARSP level. We use TARSP stage to refer to a stage on the TARSP profile chart containing specific structures, whereas we use level to refer to a child's grammatical developmental TARSP level. The levels represented in our samples were 3, 4, and 5. There are nine children with DLD and a level of 3, the other five groups consist of 10 children each.

As shown in Table 1, children with DLD are older than the language-matched TD children in all groups. A one-way analysis of variance (ANOVA) showed that the main effect of group (Level 3, 4, or 5, and TD or DLD) on age was significant, $F(5, 53) = 31.39$, $p < .0001$, $\eta^2 = .75$. Planned comparisons with Holm–Bonferroni corrections showed that children with DLD were significantly older at Level 3, $F(1, 17) = 28.80$, $p < .001$; Level 4, $F(1, 18) = 51.66$, $p < .001$; and Level 5, $F(1, 18) = 35.31$, $p < .001$. Thus, the children with DLD are significantly older than the TD children at the same grammatical level. The mean ages of the TD children fall within the norm age ranges for the three levels (Schlichting, 2017), whereas the mean ages of the children with DLD are higher than the corresponding norm ages. This suggests that the children with DLD are delayed in their grammatical development.

The children were all classified either as having (presumed[4]) DLD or as TD in the studies the children participated in, following the criteria at time of diagnosis. Background information for all children, including language test scores of the children with DLD, can be found in the Appendix. It depended on the study in which a child participated which information was available.[5] Two children with DLD were bilingual. Because their general language scores (see Appendix) are comparable to the scores of other children, we did not exclude these children. For the TD children, very little information (i.e., only on sex, age, and language background) is available. For the 20 TD children from the Bol and Kuiken (1990) corpus, general information on their socioeconomic status (SES) as a group is available. They included equal amounts of children from three social classes (lower, middle, and upper). This information is not available for individual children. Because of the missing background information in the studies the children participated in, we were only able to match TD children and children with DLD on their TARSP level.

[4] Children were indicated to have presumed developmental language disorder (DLD) when a diagnosis of DLD was not yet possible due to their young age (below 4 years).

[5] In the Netherlands, it is not legally allowed to register special categories of personal data, including ethnicity, of participants.

## Procedure

We transcribed the spontaneous language samples according to the CHAT conventions (MacWhinney, 2000). Of each sample, we selected the first 40 analyzable utterances. According to the TARSP guidelines (Schlichting, 2017), 40 analyzable utterances is the minimum number needed to reliably determine a TARSP level. In the literature, samples of at least 50 utterances are recommended as reliable samples (e.g., Owens, 2014). However, Pavelko et al. (2020) showed that "when 25-utterance samples were compared to 50-utterance samples, the mean differences indicated that utterances were 0.037 morphemes longer, sentences were 0.093 words longer, and sentences included 0.019 more clauses in the 25-utterance condition" (p. 786). The authors conclude that such small differences are not clinically meaningful. Therefore, in our opinion, their results indicate that samples of less than 50 utterances are reliable as well. For this study, collecting more utterances per child would have led to a lower number of participants, because not all samples consisted of more than 40 analyzable utterances.

Unintelligible utterances, utterances only consisting of "yes/no," (self-)repetitions, and fixed expressions (as defined in the TARSP guidelines), were not included in the analyses. A level was assigned to a child when at least 5% of the total number of his/her analyzable utterances was a declarative, interrogative, or imperative structure in a specific TARSP stage. For the children from the studies by Zwitserlood (2019) and Bruinsma et al. (2020), the TARSP analyses had already been performed. For the other children, the analyses were done by the first author of this study. Six of these samples (240 analyzable utterances in total) were analyzed by a speech-language therapist as well for a reliability check on the TARSP analysis. The intraclass correlation coefficient was .86, which is indicative of good reliability (Koo & Li, 2016).

## Outcome Measures

### Grammatical Complexity

For grammatical complexity, we selected two outcome measures: (a) the mean number of structures within each TARSP stage and (b) verb complexity. Please note that in the literature, MLU is widely used as a global index of grammatical development (Parker & Brorson, 2005). It is also used as matching criterion in most studies describing grammatical development (e.g., Spoelman & Bol, 2012). We expected that MLU in words (MLUw)[6] and the mean number of structures within each TARSP stage would be strongly correlated, because the sentence structures in the TARSP stages are divided over the stages based on their number of constituents. This was what we found: $r = .96$, $p < .001$. We decided to analyze only the mean number of structures within each TARSP stage, so not MLUw, because we believe that this is more informative than MLUw, considering TARSP takes into account the specific grammatical structures a child produces.

For the first outcome measure, the mean number of structures within each TARSP stage, all occurrences of structures were counted; if a structure was produced multiple times, all occurrences (i.e., tokens) were counted. This measure provides insight into the extent to which a child uses structures that are associated with the level (s)he has been assigned to (on the basis of the TARSP guidelines). This could show, for instance, whether children at Level 4 produce mainly structures from TARSP Stage IV, or rather structures from lower, less advanced stages. If children with DLD produce more structures from lower stages than TD children at the same level, this would indicate that they produce less complex structures than TD children.

The second measure was verb complexity. Within TARSP, children are expected to have acquired most structures from the TARSP stages lower than their grammatical level. Consequently, children at Level 3 are expected to produce less complex verb structures compared to children with higher levels. Therefore, we conducted different analyses per level. In Stages I and II, the only structures that have to do with verbs are on the level of sentence structures, for example, a single-verb utterance (*eten*, "to eat"), or an utterance consisting of an object and verb (*koekjes eten*, "cookie eat"). Most of the verbs in these structures are infinitives. Therefore, for Level 3, we chose the number of utterances containing at least one verb as the first verb complexity measure. Comparisons between children with DLD and TD children on this measure were conducted for each level. For Levels 4 and 5, we added the number of finite auxiliaries (independently used auxiliary and auxiliary + infinitive; Stage III) that children produce. For Level 5, we added an analysis of the total number of finite lexical verbs, which start to occur in Stage IV, produced by children with and without DLD.

Based on the literature (e.g., de Jong, 1999; Spoelman & Bol, 2012), the hypothesis is that children with DLD produce less complex verb constructions compared to language-matched TD children. For Level 3, this means that children with DLD produce fewer utterances containing a verb compared to TD children. Regarding the use of finite auxiliaries in Levels 4 and 5, two opposing predictions are possible: (a) children with DLD could produce fewer finite auxiliaries, because of their expected difficulty with finiteness (Hammer et al., 2014; Spoelman & Bol, 2012); or (b) children with DLD could produce more finite auxiliaries, in order to avoid finite lexical verbs (Wijnen & Verrips, 1998; Zwitserlood et al., 2015).

[6] Because MLUw is strongly correlated to MLU in morphemes (Parker & Brorson, 2005), but easier to calculate, we calculated MLUw.

Producing finite auxiliaries instead of finite lexical verbs would also result in producing fewer finite lexical verbs. Therefore, for Level 5, we hypothesize that children with DLD produce fewer finite lexical verbs than language-matched TD children.

### Grammatical Diversity

For grammatical diversity, we examined two outcome measures: (a) the number of different sentence structures (i.e., declarative, imperative, and interrogative TARSP structures) produced and (b) sentence diversity, which Hadley et al. (2018) defined as the number of different subject–verb combinations (i.e., the number of different tokens of the subject–verb type) produced by a child.

The number of different sentence structures indicates how varied a child's grammatical repertoire is; the more different sentence structures a child can use, the more varied their grammatical repertoire is, and the more different type(s) of messages they can convey. For this analysis, we counted all declarative, imperative, and interrogative structures, as well as one-word utterances (noun, verb, or adverb). First, we checked whether the total number of sentence structures differed per group. Although each sample consists of 40 analyzable utterances, the total number of sentence structures is not necessarily 40. An utterance could be, for instance, a complex sentence (i.e., two sentence structures within one utterance) or an isolated phrase (such as *mijn koekje*, "my cookie") that cannot be classified as a sentence within the TARSP instrument. If the total number of sentence structures does not differ between the groups of children, the number of different sentence structures can be seen as indicator of how varied a child's grammatical repertoire is.

The second diversity measure was sentence diversity, which indicates whether a child can use structures in a creative manner, as opposed to rote reproduction. Hadley et al. (2018) point out that the more different types of subject–verb combinations a child produces, the more likely it is that (s)he produces these structures by grammatical encoding rather than by memorizing chunks. We counted the number of different subject–verb combinations in each sample. Following Hadley et al., we did not take forms of the copula *zijn* ("to be"), for example, *dat is tijger* ("that is tiger"), into account.

### Grammatical Accuracy

As measures of grammatical accuracy, we analyzed (a) the percentage of correct utterances, (b) the percentage of verb-related errors, and (c) the percentage of nonverb-related errors. The first accuracy measure can be seen as a general measure of grammatical accuracy. We calculated the percentage of correct utterances of the total number of utterances containing at least two words. As grammatical correctness can only be determined for utterances that have a grammatical structure, one-word utterances were excluded. We scored an utterance as correct if a Dutch adult would consider the utterance as grammatically correct. So, for instance, subject drop that would be acceptable in adult Dutch was not scored as incorrect, but nonstandard forms that also occur in typical development were scored as incorrect (for instance subject–infinitival verb combinations ["root infinitives"], such as *papa eten*, "daddy eat"). We copied the classification procedure described in Zwitserlood et al. (2015) and applied it to utterances of both groups of children in the same way.

Follow-up analyses on verb-related and nonverb-related errors can provide insight into the specific grammatical difficulties preschool-age children with DLD might have. The division in verb-related and nonverb-related errors was based on the error analysis conducted by Zwitserlood et al. (2015). The percentage of verb-related errors on the total number of utterances containing at least one verb and consisting of at least two words, consisted of errors in subject–verb agreement, tense, omissions of auxiliaries, participles, root infinitives, and argument omissions.

Following Zwitserlood et al. (2015), we calculated the percentage of nonverb-related errors in the total set of utterances consisting of two words or more, of which at least one was a verb. This category consisted of all remaining errors (i.e., those related to determiners, prepositions, pronouns, adjectival inflection, congruency between an adjective and a noun, adverbs, or word order).

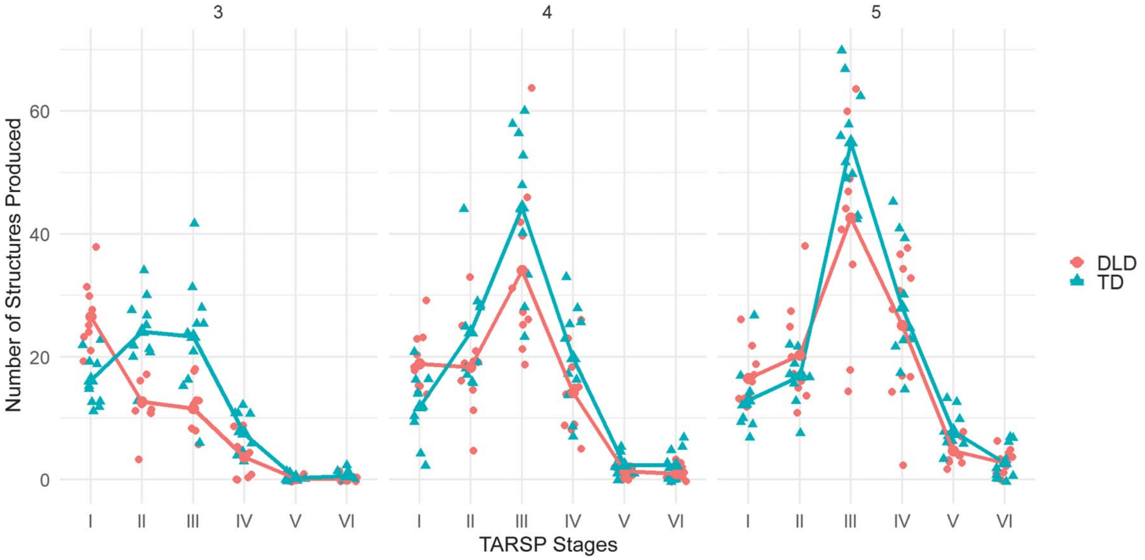
For school-age children, differences in grammatical accuracy scores and verb complexity between children with DLD and TD children matched on grammatical level were found (e.g., de Jong, 1999; Zwitserlood et al., 2015). However, whether these differences are also present in younger children is not yet clear. Therefore, no specific hypotheses can be formulated. The other complexity and diversity analyses are exploratory; we had no a priori predictions concerning differences between children with and without DLD at the same grammatical level.

## Statistical Analyses

All analyses were computed in R (R Core Team, 2020). Because we were only interested in specific comparisons (i.e., children with DLD and TD children at the same grammatical level), we examined the effect of a combined variable of level and having DLD or not for each analysis. In this way, we were able to conduct planned comparisons between children with and without DLD at the same grammatical level. This variable will be referred to as "Group" and consists of six levels: DLD3 (children with DLD of Level 3), TD3 (TD children of Level 3), DLD4, TD4, DLD5, and TD5.

For the mean number of structures within each TARSP stage, we used the R package lme4 (Bates et al.,

**Figure 1.** The mean number of structures produced per TARSP stage (Taal Analyse Remediëring en Screening Procedure; Schlichting, 2017) for each grammatical level. The numbers above each graph indicate grammatical level (of the children). Each dot or triangle indicates the number of structures associated with a TARSP stage (I-VI) produced by a child. The red dots represent the children with developmental language disorder (DLD) and the blue triangles represent the typically developing (TD) children.



2015) to perform a linear mixed-effects analysis of the relationships Group (six groups; nine children in group DLD3, the other five groups consist of 10 children each) and TARSP stage (I–VI). These two factors were the fixed effects of the model, with interaction term. An intercept for subject was added as random effect. The number of structures within each TARSP stage was the dependent variable. To test whether children with DLD and TD children at the same level produced different numbers of structures within each TARSP stage, we calculated the estimated marginal means, using the emmeans package (Lenth, 2020), and we performed pairwise comparisons with Holm–Bonferroni corrections between the six groups of children on each TARSP stage.

For the other measures, the assumptions (no outliers, normally distributed residuals, and homogeneity of variances) for a one-way ANOVA were checked first. When these assumptions were not met, a Kruskal–Wallis test was applied. As we were interested in potential differences between children at the same level, three planned contrasts, with Holm–Bonferroni corrections, were performed per analysis between children with and without DLD within the same grammatical level (i.e., Level 3, 4, or 5).

## Results

### Grammatical Complexity

The mixed-effects analysis returned a significant Group × TARSP stage interaction, $\chi^2(25) = 228.01$, $p < .001$. This interaction indicates that numbers of structures associated with different TARSP stages differ across the groups of children. Figure 1 shows the mean numbers of structures per TARSP stage produced by children assigned to one of the three levels, together with the scores of the individual children. Planned contrasts with Holm–Bonferroni corrections demonstrate differences between children with and without DLD at the same grammatical level. For Level 3, children with DLD produce significantly more Stage I structures (estimated difference[7] = 10.36, $p = .017$, confidence interval[8] [CI] [0.93, 19.78]), and significantly fewer Stage II (estimated difference = −11.43, $p < .01$, CI [−20.86, −2.01]) and Stage III structures (estimated difference = −11.74, $p < .01$, CI [−21.17, −2.32]) than TD children. DLD children at Level 4 produce significantly fewer Stage III structures (estimated difference = −10.20, $p < .01$, CI [−19.37, −1.03]) than TD children at the same grammatical level. This was also found for Level 5; children with DLD produce significantly fewer Stage III structures (estimated difference = −12.30, $p < .001$, CI [−21.47, −3.13]), but are similar with regard to Stage IV or V structures.

In Figure 1, the dispersion of the number of structures produced per TARSP stage is also shown. It demonstrates that all distributions of children with DLD and TD children overlap to a certain extent. Furthermore, the

[7] As compared to mean number of Stage-I structures produced by TD children with a grammatical level of 3.

[8] 95% CIs.

variation within the groups of children is relatively large for most TARSP stages, except for the numbers of structures in TARSP Stages V and VI.

For verb complexity, the first measure we examined was the number of utterances containing at least one verb. A Kruskal–Wallis test showed significant differences across the six groups of children, $\chi^2(5) = 30.0$, $p < .001$, $\eta^2 = .47$. The planned Mann–Whitney *U* tests with Holm–Bonferroni corrections showed that the difference between children with DLD and TD children at Level 3 was significant ($U = 6.5$, $p < .01$, $r = .73$). The number of utterances containing a verb did not differ between children with and without DLD for Levels 4 and 5.

Figure 2 shows the variation in the number of utterances containing a verb within each level. The figure shows much overlap for the groups of children with and without DLD at Levels 4 and 5, but to a lesser extent at Level 3. This also reflects the significant difference found for Level 3. The graph also shows that the variation within the group of children with DLD with Level 5 is larger than the variation within the group of TD children at the same level.

For Levels 4 and 5, we analyzed the number of auxiliaries (independently used auxiliary and auxiliary + infinite verb structures) produced. The analyses showed that the differences across the groups of children are not significant.

For Level 5, an additional analysis was conducted on the number of finite lexical verbs. Children with DLD produce significantly fewer finite lexical verbs than TD children, $F(1, 18) = 10.42$, $p < .01$, $\eta^2 = .37$. In Figure 3, the dispersion of the number of finite lexical verbs is shown per group. This figure demonstrates that the children with DLD overlap only with the lowest half of the TD children.

## Grammatical Diversity

Prior to analyzing the number of different TARSP sentence structures, we verified that the total number of sentence structures did not differ per group, $F(5, 53) = 1.66$, $p = .16$, $\eta^2 = .14$. The number of different TARSP sentence structures differed significantly across the six groups of children, $F(5, 53) = 7.85$, $p < .001$, $\eta^2 = .43$. However, planned contrasts with Holm–Bonferroni corrections showed that the differences between children with DLD and TD children at the same grammatical level were not significant for any of the levels. The mean scores per group can be found in Table 2.

Regarding the number of different SV combinations (i.e., sentence diversity), an overall effect was found for the Group variable, $F(5, 53) = 12.92$, $p < .001$, $\eta^2 = .55$, indicating that the six groups of children differ in the number of SV combinations produced. However, planned contrasts with Holm–Bonferroni corrections showed that there were no significant differences between children with DLD and TD children at Level 3, 4, or 5.

In Figure 4, it is shown that the variation within the groups of children with DLD is relatively large for Levels 4

**Figure 2.** The number of utterances containing at least one verb produced per child. The red dots represent children with developmental language disorder (DLD) and the blue triangles represent typically developing (TD) children. TARSP = Taal Analyse Remediëring en Screening Procedure.

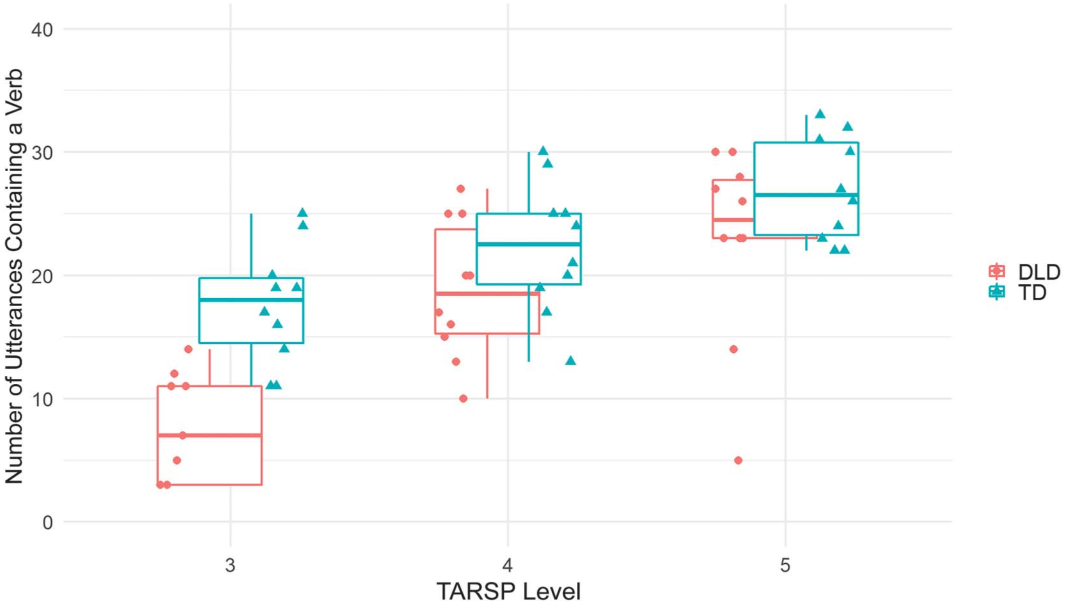
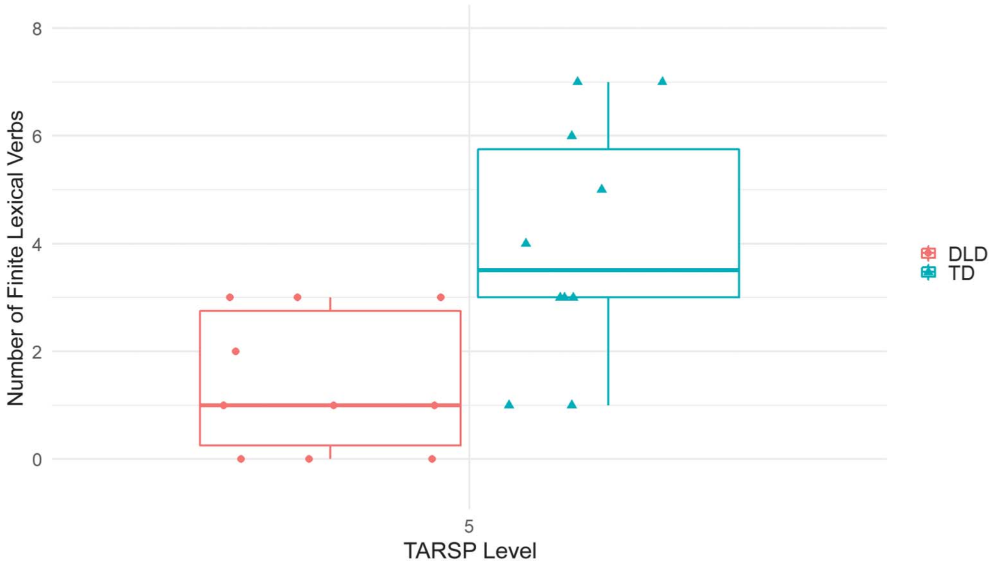


**Figure 3.** The numbers of finite lexical verbs produced by children at Level 5. The red dots represent the children with developmental language disorder (DLD) and the blue triangles represent the typically developing (TD) children. TARSP = Taal Analyse Remediëring en Screening Procedure.



and 5. Additionally, many children with DLD at Level 3 (five out of nine children) and one child with DLD at Level 5 did not produce an SV combination at all.

## Grammatical Accuracy

We compared the percentages of correct utterances, verb-related errors and nonverb-related errors between children with DLD and TD children (see Table 3 for the mean percentages and standard deviations). The percentages of correct utterances, $\chi^2(5) = 24.29$, $p < .001$, $\eta^2 = .36$, and verb-related errors, $\chi^2(5) = 25.23$, $p < .001$, $\eta^2 = .38$, differed significantly across the six groups of children. For the percentage of nonverb-related errors, no significant group effect was found, $\chi^2(5) = 10.45$, $p = .064$, $\eta^2 = .11$. We did not find differences between children with and without DLD at the same TARSP level for any of the comparisons. Therefore, these analyses indicate that children with DLD and TD children at the same level do not differ in their grammatical accuracy.

**Table 2.** The mean scores (standard deviations) for the number of different sentence structures and the sentence diversity scores per subgroup.

| Level | Subgroup | Number of different sentence structures | Sentence diversity |
|---|---|---|---|
| 3 | DLD (*n* = 9) | 8.89 (1.76) | 1.22 (1.79) |
|  | TD (*n* = 10) | 11.0 (2.0) | 5.0 (3.56) |
| 4 | DLD (*n* = 10) | 13.1 (2.38) | 7.7 (4.14) |
|  | TD (*n* = 10) | 13.1 (2.28) | 8.5 (2.59) |
| 5 | DLD (*n* = 10) | 13.6 (2.55) | 12.1 (5.82) |
|  | TD (*n* = 10) | 14.2 (1.99) | 13.8 (4.34) |

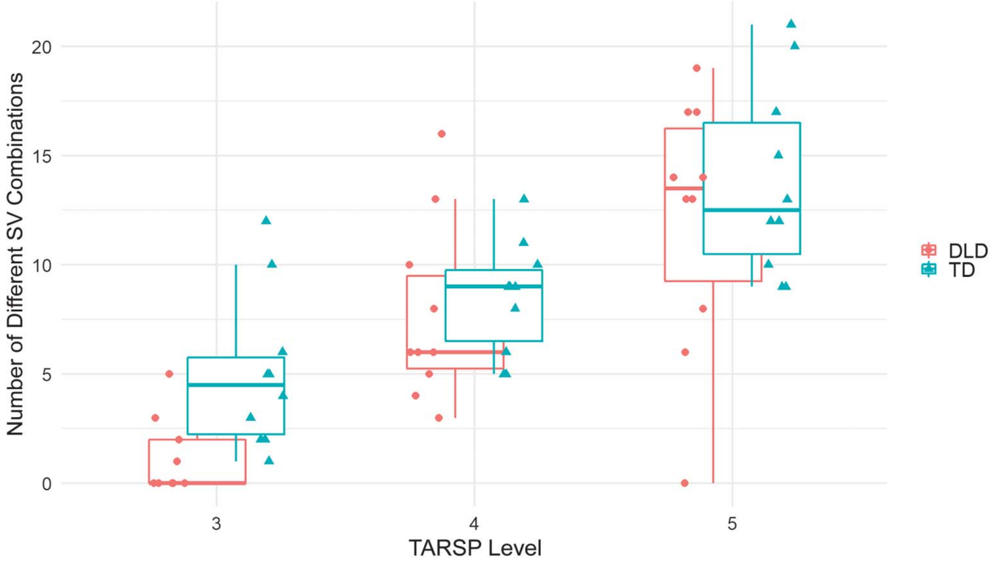
*Note.* DLD = developmental language disorder; TD = typically developing.

## Correlations Between Dimensions

Finally, we investigated if the scores on the three dimensions were correlated. To do this, we transformed the scores on the measures within each dimension into *z* scores. The mean of the *z* scores of the measures within each dimension was used as total score on that dimension. For diversity, all measures were taken into account. For accuracy, we selected only the percentage of correct utterances, as the verb- and nonverb-related error analyses were secondary. For complexity, the mean *z* score was calculated differently for each level, due to the different verb complexity measures that were administered in each level.

Spearman correlation analyses showed that the correlation between complexity and diversity for all children together was .32 ($p = .013$), between complexity and accuracy .33 ($p = .011$), and between diversity and accuracy .59 ($p < .001$). These correlations are also reflected in Figure 5. When we examine these correlations more closely, we can see an asymmetric pattern: children who have low complexity and diversity scores are also likely to have low accuracy scores (most scores on the lower left side of the graph are light blue or red); however, children who have high complexity scores do not necessarily have high diversity or

**Figure 4.** The numbers of different subject–verb (SV) combinations produced per child. The red dots represent children with developmental language disorder (DLD) and the blue triangles represent typically developing (TD) children. TARSP = Taal Analyse Remediëring en Screening Procedure.



accuracy scores. The highest diversity scores are related to high accuracy scores (i.e., darker blue dots and triangles).

For the TD children only, there were significant correlations between complexity and diversity ($r$ = .42, $p$ = .02), between complexity and accuracy ($r$ = .40, $p$ = .03), and between diversity and accuracy ($r$ = .55, $p$ < .01). For the children with DLD, the correlations including complexity were not significant. For these children, the correlation between complexity and diversity was .24 ($p$ = .21), between complexity and accuracy .25 ($p$ = .18), and between diversity and accuracy .65 ($p$ < .001).

## Discussion

This study examined the grammatical repertoires of 3- to 6-year-old children with DLD, and compared these to those of TD children matched on grammatical level as indicated by TARSP (Schlichting, 2017). The aim was to determine if comparing children with DLD to norm scores is an appropriate method for selecting goals for grammatical interventions. As the children were matched on grammatical level, the TD children were younger than the children with DLD. The analyses showed that children with DLD and TD children at the same grammatical level are comparable in the diversity of their grammatical repertoire and the accuracy of their utterances, but not in the complexity of the grammatical structures produced. Children with DLD produce less complex structures than would be expected based on their assigned grammatical level. These less complex structures are corresponding to lower TARSP stages.

### Grammatical Complexity

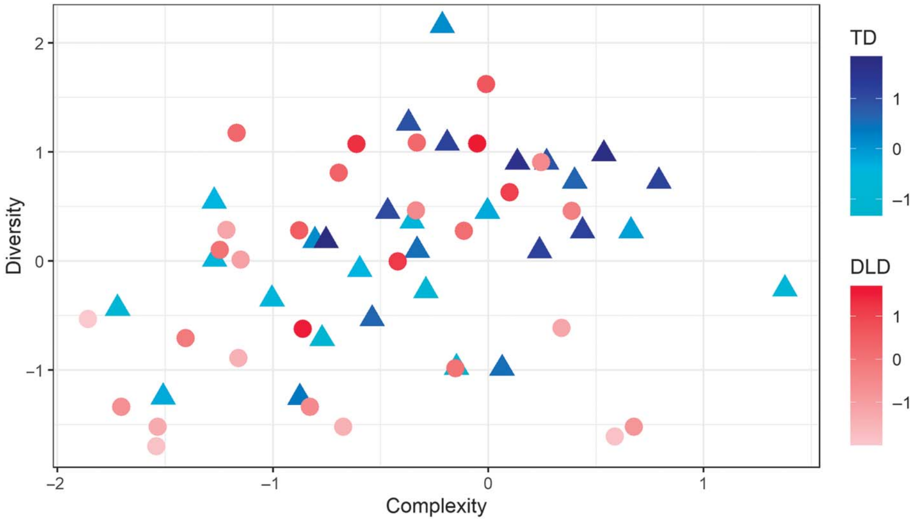
The structures produced by children with DLD were less complex on average than those produced by TD

**Table 3.** The mean percentages (and standard deviations) of correct utterances, verb-related errors, and nonverb-related errors per grammatical level and diagnosis.

| Level | Subgroup | Percentage correct utterances | Percentage verb-related errors | Percentage nonverb-related errors |
|---|---|---|---|---|
| 3 | DLD (*n* = 9) | 40% (18%) | 53% (25%) | 33% (33%) |
|  | TD (*n* = 10) | 46% (12%) | 53% (13%) | 20% (16%) |
| 4 | DLD (*n* = 10) | 54% (13%) | 30% (14%) | 31% (12%) |
|  | TD (*n* = 10) | 65% (20%) | 26% (20%) | 19% (13%) |
| 5 | DLD (*n* = 10) | 68% (8%) | 23% (9%) | 15% (9%) |
|  | TD (*n* = 10) | 72% (11%) | 21% (12%) | 13% (7%) |

*Note.* DLD = developmental language disorder; TD = typically developing.

**Figure 5.** Mean diversity and complexity *z* scores for all children. The darker the blue or red, the higher the mean accuracy *z* score. The red dots represent the children with developmental language disorder (DLD) and the blue triangles represent the typically developing (TD) children.



children. More specifically, children with DLD at Level 3 produce more one-word-utterances, and fewer structures linked to TARSP Stages II and III (i.e., the stages in which combinations of words and more complex phrase and word structures start to emerge). At Levels 4 and 5, children with DLD produce fewer structures from Stage III than TD children do, whereas they do not produce more complex structures from higher, more advanced stages compared to the language-matched TD children.

Differences in performance on grammatical complexity were also found by de Jong (1999). He found that children with DLD, aged between 4;8 and 8;2, produce simpler argument structures (i.e., more intransitive verbs, and fewer verbs with more than one internal argument) compared to TD children matched on MLU. Because sentence structures containing a subject, a verb, and an object start to occur in TARSP Stage III, the result of this study could be seen as, indirectly, similar to de Jong's finding. Zwitserlood et al. (2015), on the other hand, did not find differences in grammatical complexity in children with DLD and language age-matched TD children, as reflected by MLUw and percentages of compound sentences. A potential explanation for these contrasting results might be that the children with DLD in this study were younger than the children in the study by Zwitserlood et al. (mean age 6;5 on the first measurement). Consequently, the language level of the children in this study is lower as well, and differences in grammatical complexity between children with DLD and language-matched TD children might occur in an earlier stage of language acquisition. This explanation might also be supported by the variation found between the groups of children (as illustrated by the dispersion graphs). On a group level, the differences between children with DLD and TD children appear to decrease from Level 3 to Level 5. Additionally, the variation within the group of children with DLD increases per level. This suggests that, especially for the higher levels, some children with DLD perform TD-like, whereas others do not. An explanation for this might be that some children with DLD benefit more from the language therapy they receive than others, resulting in TD-like scores at higher grammatical levels. Additionally, it might be that there is a trade-off between complexity and accuracy; children with a higher grammatical level produce utterances of similar complexity to the utterances of language-matched TD children, but they make more grammatical errors than TD children. We will explain this in more detail under the Grammatical Accuracy section below.

For verb complexity, the results showed that children with DLD at Level 3 produce fewer utterances containing a verb compared to TD children at the same grammatical level, and that children with DLD at Level 5 produce fewer finite lexical verbs. These results indicate that children with DLD have more difficulties with verb constructions than could be expected based on their grammatical level, both on a basic (Level 3) and on a more complex level (5). This is in line with research that showed that school-age children with DLD have difficulties with verb structures (e.g., de Jong, 1999; Spoelman & Bol, 2012; Zwitserlood et al., 2015). No differences were found for the number of finite auxiliaries produced. This could be explained by earlier findings that children with DLD use "dummy verbs" (which are counted as finite auxiliaries in TARSP) instead of finite lexical verbs for a longer period of time than TD children (Zwitserlood et al., 2015).

## Grammatical Diversity

No differences in numbers of different sentence structures were found between children with and without DLD for any of the levels. Additionally, no differences were found in sentence diversity (i.e., the number of different subject–verb combinations). Note, however, that children still could have produced a subject with a form of the copula *be*. These structures were not taken into account in this analysis, following Hadley et al. (2018). The absence of significant differences between the groups of children might be caused by the low number of children per group. Larger sample sizes might show differences that we were not able to find. Especially because, despite the lack of significant group differences, the scatter plot of the number of SV combinations produced suggests that the visible differences between children with and without DLD do decrease between Level 3 and Level 5. In other words, it seems that children with DLD score more TD-like on grammatical diversity when their grammatical level is higher.

## Grammatical Accuracy

No significant difference was found between the children with DLD and the TD children on the percentage of correct utterances, nor on the percentages of verb-related and nonverb-related errors. The current finding that children with DLD do not differ in their overall grammatical accuracy scores is not in line with an earlier study that found differences on this measure for Dutch school-age children with and without DLD by Zwitserlood et al. (2015), on which the procedure of the accuracy analysis of this study was based. An explanation for these contrasting findings could be that the children in this study are younger than the children in the Zwitserlood study, and consequently, their language level is lower as well. It is possible that different error patterns arise in more advanced grammatical structures than in the structures produced by preschoolers. In relation to the opposite finding for complexity (i.e., fewer differences for children with higher grammatical levels), this could mean that the nature of differences between children with and without DLD changes with increasing proficiency; at lower grammatical levels, children with DLD do not produce structures as complex as those of TD children, and consequently, are not making errors in these structures. However, at higher grammatical levels, children with DLD have begun to acquire these complex structures, but have not mastered them in full (yet), or have difficulty with producing them, resulting in errors. A second explanation might be found in the different tasks; whereas the children in this study were recorded during free-play situations, the children in the study by Zwitserlood et al. were recorded during a narrative task. Sealey and Gilmore (2008) found that children are more accurate on their finite-verb production during free-play situations. This could also explain why this study did not find differences in verb-related accuracy scores, in contrast to the studies by Zwitserlood et al. (2015) and de Jong (1999).

## Variation Between and Within Children

There is much variation between children with DLD on almost all measures. Some children with DLD score TD-like on some measures, while others score lower than their language-matched peers. This is in line with the general observation that children with DLD form a very heterogeneous population.

Additionally, the correlations between the three dimensions (i.e., complexity, diversity, and accuracy) showed that children who score relatively high on grammatical diversity are likely to score relatively high on grammatical accuracy as well. In other words, children who have a larger grammatical repertoire, are likely to make fewer grammatical errors. However, for the children with DLD, the correlations with complexity are not significant, whereas they are for the TD children. This suggests that there is more variation between children with DLD on grammatical complexity than on accuracy and diversity. Therefore, the correlation analyses reflect the results on the separate measures; children with DLD seem to lag behind on grammatical complexity compared to TD children matched on grammatical level, and not that much on grammatical diversity and accuracy, indicating an uneven profile (Leonard, 2014).

## Limitations and Further Research

In this study, the language samples were taken from previous studies, and consequently, the recording settings differed: recordings were made at home or at school. It might be that these different, although familiar, settings influenced the grammatical productions of the children. According to Bornstein et al. (2000), different settings (i.e., a familiar home vs. an unfamiliar laboratory) do not influence the total number of utterances, word roots, and MLU in 2-year-olds. We found no study that compared recordings elicited at home and at school. In addition, there were differences between interactants. The TD children were recorded with a parent and examiner(s) present, whereas the children with DLD were recorded with only an examiner present. Eisenberg et al. (2018) suggest that interactant (i.e., parent vs. examiner) affects the grammatical productions of children. Indeed, Bornstein et al. (2000) showed that children produce more utterances and more word roots when speaking with their mother compared to a stranger, but no differences were found for MLU. As such, the difference in conversation partners between the two groups of children in this study might have influenced the grammatical structures produced. Other factors, such as the toys with which the children played might have affected the results as well. For instance, in the recordings of the children with DLD, Playmobil toys were used, whereas in the CHILDES recordings of TD children, children played with different types of toys. Playing with Playmobil, for example, might result in using more personal pronouns, whereas playing with toy blocks or toy cars might result in using more prepositions. We did not include the type of toys children played with as variable in our analyses. Because of the differences in elicitation and recording settings between the children with DLD and the TD children, our results should be interpreted with some caution. Further research should make sure that if children are recorded during free-play, the play settings are as similar as possible, including the type of toys with which the children are playing.

Another limitation of this study is that it was only possible to group-match the TD children and children with DLD on their TARSP level. Individual matching was not possible because little background information on the children was available, especially for the TD children. We could, therefore, not match the groups of children on, for instance, their SES or language test scores. Additionally, it is not entirely clear how the TD children were classified as TD in the studies the children participated in. Consequently, our results should be interpreted with some caution; the differences between the groups of children might not be explained by DLD status alone. It is recommended that future studies match children with and without DLD on other characteristics than TARSP level. Moreover, some of the children with DLD have phonological difficulties as well (see the Appendix). Because there were only five children known to have a phonological disorder, and we did not have specific phonological scores of the children, we could not look into the possible influence this may have had on grammatical development. In further research it would be interesting to see whether phonological difficulties influence grammatical development.

We compared children with DLD to TD children matched on their TARSP level, which is derived from the structures a child produces and predicts which other structures the child is expected to produce. Most previous studies used MLU as matching criterion. As TARSP only has a few stages and assigns a grammatical level based on the 5% rule, it is less fine-grained than MLU, although MLU does not contain information on the use of grammatical structures for Dutch children. Although MLU and the mean number of structures within each TARSP stage were highly correlated, it is still possible that children with the same TARSP level had different MLUs. Matching on MLU might, therefore, have led to different results. However, because information of MLU and corresponding grammatical structures is lacking for Dutch, MLU cannot be used for selecting goals for interventions. TARSP, on the other hand, is used for selecting grammatical treatment goals in clinical practice in the Netherlands. Therefore, we believe that matching on TARSP level, although it is less fine-grained, is more informative for clinical practice.

A consequence of using an LSA method is that the measures are by definition indirect measures of grammatical knowledge.[9] For children with DLD there may be an influence of suboptimal processing skills. Therefore, children could have had more knowledge of structures than they showed in their language use during the recordings. In other words, children with DLD might have difficulty in applying their knowledge of certain grammatical structures during online (i.e., immediate) language production, which is also reflected in literature on speech disruptions (e.g., Guo et al., 2008). Indeed, differences between knowledge and production of grammatical aspects by children with DLD have been found, for instance in the acquisition of grammatical gender and the definite article in Dutch (Keij et al., 2012). Additionally, children could have avoided structures they thought were difficult, resulting in fewer grammatical errors. To test whether this is caused by grammatical knowledge deficiencies or production difficulties, future research should test grammatical comprehension as well.

Another suggestion for further research is related to the fact that we did not find differences in grammatical accuracy between children with and without DLD. An explanation for this might be that we defined errors as structures that Dutch adults would not consider as grammatically correct. It would be interesting to examine whether young children with DLD produce atypical errors or immature forms of structures compared to typical development. In this study, it was not possible to examine specific (atypical) error types, because many of the error types were associated with limited numbers of tokens. Potentially, this could be examined in longer language samples than samples consisting of forty analyzable utterances as was done in this study.

This study did not collect longitudinal language samples of each child. Consequently, it was not possible to see how children develop and whether the developmental trajectories of children with DLD are similar to the trajectories of TD children in the first stages of their grammatical development. It would be interesting to examine whether grammatical complexity, accuracy, and variation in children's grammatical repertoire develop simultaneously.

[9] This holds for TARSP as well as for MLU.

## Clinical Implications

Differences between children with DLD and language-matched TD children, and the substantial interindividual variation observed indicate that comparing children with DLD to developmental stages in typical development is not sufficient for determining a child's grammatical level or for selecting goals for interventions. Instead, a thorough analysis is needed of the structures a child does and does not produce within their assigned TARSP level. The results underline the importance of conducting language sample analyses, in which it is especially important to include the complexity of the utterances a child with DLD produces.

The uneven profile (Leonard, 2014) that we found for children with DLD suggests that SLPs using TARSP may want to consider less complex grammatical structures of lower TARSP-stages than the level of a child as goals for interventions. The results also suggest that utterances of children with DLD are grammatically less complex than utterances of language-matched TD children, but that differences are smaller for grammatical diversity and accuracy. This implies that SLPs may want to prioritize treating grammatical complexity.

## Conclusions

In this study, the grammatical production patterns of 3- to 6-year-old children with DLD were compared to younger TD children with the same TARSP levels on grammatical complexity, diversity, and accuracy. The results showed that children with DLD are comparable to language-matched TD children in their grammatical accuracy and variation, but they lag behind in their grammatical complexity. This could be seen as an uneven grammatical profile (Leonard, 2014). Additionally, the results suggest that overall grammatical level, as indicated by TARSP level, is not sufficient for selecting grammatical treatment goals: children with the same TARSP outcome do not necessarily have the same inventory of grammatical structures. This also demonstrates the difficulty of validly measuring grammatical development. The results indicate that in addition to general measures of grammatical development, it is important to include the complexity of the utterances a child with DLD produces for determining a child's grammatical level.

## Data Availability Statement

The language samples analyzed in this study were collected in previous studies. The samples of Dutch TD children were retrieved from the CHILDES database (MacWhinney, 2000). These samples were part of studies by Bol and Kuiken (1990), van Kampen (2009), Bol (1995), and Elbers and Wijnen (1992). The samples of Dutch children with DLD were collected by Zwitserlood (2019), Bruinsma et al. (2020), Boerma et al. (2020), and Bol and Kuiken (1990), the latter available through CHILDES. The dataset analyzed during this study is available from the corresponding author on reasonable request.

## Acknowledgments

This research was funded by the Royal Dutch Auris Group. The authors would like to thank Gerda Bruinsma, Rob Zwitserlood (HU University of Applied Sciences Utrecht), and Tessel Boerma, Emma Everaert, and Iris Selten (Utrecht University, Utrecht Institute of Linguistics OTS) for sharing their data with them. Additionally, they would like to thank Iris Duinmeijer (Dutch Foundation of the Deaf and Hard of Hearing Child) for her valuable feedback, Kirsten Schutter (Utrecht University, Utrecht Institute of Linguistics OTS) for her statistical advice, and Inge van Dijke (Royal Dutch Auris Group) for the interrater reliability check.

## References

**Bates, D., Mächler, M., Bolker, B., & Walker, S.** (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

**Boerma, T., Selten, I., & Everaert, E.** (2020). *Het taalprofiel van kinderen met het 22q11.2 deletiesyndroom in vergelijking met kinderen met TOS* [Conference poster presentation]. TaalStaal 2020, Nieuwegein, the Netherlands. https://www.taalstaal.nl/wp-content/uploads/2020/04/VCF-Syndroom-Everaert-e.a.pdf

**Bol, G. W.** (1995). Implicational scaling in child language acquisition: The order of production of Dutch verb constructions. In M. Verrips & F. Wijnen (Eds.), *Papers from the Dutch-German colloquium on language acquisition, Amsterdam series in child language development, 3*. Institute for General Linguistics.

**Bol, G. W., & Kasparian, K.** (2009). The production of pronouns in Dutch children with developmental language disorders: A comparison between children with SLI, hearing impairment, and Down's syndrome. *Clinical Linguistics & Phonetics, 23*(9), 631–646. https://doi.org/10.1080/02699200902995677

**Bol, G. W., & Kuiken, F.** (1990). Grammatical analysis of developmental language disorders: A study of the morphosyntax of children with specific language disorders, with hearing impairment and with Down's syndrome. *Clinical Linguistics & Phonetics, 4*(1), 77–86. https://doi.org/10.3109/02699209008985472

**Boomers, A., & Mugge, A.** (1982). Reynell Taalontwikkelingstest: Nederlandse formulieren en Nederlandse instructie [Reynell Language development test: Dutch form and instruction].

**Bornstein, M. H., Haynes, O., Painter, K. M., & Genevro, J. L.** (2000). Child language with mother and with stranger at home and in the laboratory: A methodological study. *Journal of Child Language, 27*(2), 407–420. https://doi.org/10.1017/S0305000900004165

**Bruinsma, G., Wijnen, F., & Gerrits, E.** (2020). Focused stimulation intervention in 4- and 5-year-old children with developmental language disorder: Exploring implementation in clinical practice. *Language, Speech, and Hearing Services in Schools, 51*(2), 247–269. https://doi.org/10.1044/2020_LSHSS-19-00069

**Crystal, D., Fletcher, P., & Garman, M.** (1976). *The Grammatical Analysis of Language Disability*. Edward Arnold.

**de Jong, J**. (1999). *Specific language impairment in Dutch: Inflectional morphology and argument structure* [Doctoral dissertation, Rijksuniversiteit Groningen].

**Elbers, L., & Wijnen, F.** (1992). Effort, production skill and language learning. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications*. York Press.

**Eisenberg, S. L., Guo, L.-Y., & Mucchetti, E.** (2018). Eliciting the language sample for developmental sentence scoring: A comparison of play with toys and elicited picture description. *American Journal of Speech-Language Pathology, 27*(2), 633–646. https://doi.org/10.1044/2017_AJSLP-16-0161

**Guo, L. Y., Tomblin, J. B., & Samelson, V.** (2008). Speech disruptions in the narratives of English-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 51*(3), 722–738. https://doi.org/10.1044/1092-4388(2008/051)

**Hadley, P. A., McKenna, M. M., & Rispoli, M.** (2018). Sentence diversity in early language development: Recommendations for target selection and progress monitoring. *American Journal of Speech-Language Pathology, 27*(2), 553–565. https://doi.org/10.1044/2017_AJSLP-17-0098

**Hammer, A., Coene, M., Rooryck, J., & Govaerts, P. J.** (2014). The production of Dutch finite verb morphology: A comparison between hearing-impaired CI children and specific language impaired children. *Lingua, 139*(1), 68–79. https://doi.org/10.1016/j.lingua.2013.11.010

**Heilmann, J. J., Miller, J. F., & Nockerts, A.** (2010). Using language sample databases. *Language, Speech, and Hearing Services in Schools, 41*(1), 84–95. https://doi.org/10.1044/0161-1461(2009/08-0075)

**Keij, B., Cornips, L. M. E. A., van Hout, L., Hulk, A., & van Emmerik, J.** (2012). Knowing versus producing. *Linguistic Approaches to Bilingualism, 2*(4), 379–403. https://doi.org/10.1075/lab.2.4.02kei

**Klatte, I. S., van Heugten, V., Zwitserlood, R., & Gerrits, E.** (2022). Language sample analysis in clinical practice: Speech-language pathologists' barriers, facilitators, and needs. *Language, Speech, and Hearing Services in Schools, 53*(1), 1–16. https://doi.org/10.1044/2021_LSHSS-21-00026

**Koo, T. K., & Li, M. Y.** (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

**Lenth, R.** (2020). *Emmeans: Estimated marginal means, aka least-squares means.* R package Version 1.5.0. https://CRAN.R-project.org/package-emmeans

**Leonard, L. B.** (2014). *Children with specific language impairment* (2nd ed.). The MIT Press. https://doi.org/10.7551/mitpress/9152.001.0001

**MacWhinney, B.** (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Erlbaum.

**Owens, R. E.** (2014). *Language disorders: A functional approach to assessment and intervention* (6th ed.). Pearson.

**Parker, M. D., & Brorson, K.** (2005). A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language, 25*(3), 365–376. https://doi.org/10.1177/0142723705059114

**Pavelko, S. L., Owens, R. E., Ireland, M., & Hahs-Vaughn, D. L.** (2016). Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools, 47*(3), 246–258. https://doi.org/10.1044/2016_LSHSS-15-0044

**Pavelko, S. L., Price, L. R., & Owens, R. E., Jr.** (2020). Revisiting reliability: Using Sampling Utterances and Grammatical Analysis Revised (SUGAR) to compare 25- and 50-utterance language samples. *Language, Speech, and Hearing Services in Schools, 51*(3), 778–794. https://doi.org/10.1044/2020_LSHSS-19-00026

**R Core Team.** (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

**Schlichting, L.** (2017). *TARSP: Taalontwikkelingsschaal van Nederlandse kinderen van 1–4 jaar met aanvullende structuren tot 6 jaar* [Language development scale of Dutch children 1–4 years with additional structures up to 6 years] (8th ed.). Pearson Benelux.

**Sealey, L. R., & Gilmore, S. E.** (2008). Effects of sampling context on the finite verb production of children with and without delayed language development. *Journal of Communication Disorders, 41*(3), 223–258. https://doi.org/10.1016/j.jcomdis.2007.10.002

**Spoelman, M., & Bol, G. W.** (2012). The use of subject–verb agreement and verb argument structure in monolingual and bilingual children with specific language impairment. *Clinical Linguistics & Phonetics, 26*(4), 357–379. https://doi.org/10.3109/02699206.2011.637658

**van Kampen, J.** (2009). The non-biological evolution of grammar: Wh-question formation in Germanic. *Biolinguistics, 3*(2–3), 154–185. https://doi.org/10.5964/bioling.8719

**Verbeek, J., van den Dungen, L., & Baker, A.** (2007). *Verantwoording van het STAP-instrument*. Universiteit van Amsterdam. https://www.hetwap.nl/wp-content/uploads/2018/04/2007-STAP-VERANTWOORDING.pdf

**Wijnen, F., & Verrips, M.** (1998). The acquisition of Dutch syntax. In S. Gillis & A. De Houwer (Eds.), *The acquisition of Dutch (pragmatics and beyond, NS 52)* (pp. 223–299). John Benjamins. https://doi.org/10.1075/pbns.52.07wij

**Zwitserlood, R.** (2019). Effectiviteit en bruikbaarheid van FIT-Digitaal in TOS behandelgroepen van Auris: Een pilotstudie [Effectiveness and usability of FIT-Digitaal in language-focused treatment groups for children with DLD of Auris: A pilot study.]. https://vhz-online.nl/effectiviteit-en-bruikbaarheid-van-fit-digitaal-in-tos-behandelgroepen-van-auris-een-pilotstudie

**Zwitserlood, R., Weerdenburg, M., Verhoeven, L., & Wijnen, F.** (2015). Development of morphosyntactic accuracy and grammatical complexity in Dutch school-age children with SLI. *Journal of Speech, Language, and Hearing Research, 58*(3), 891–905. https://doi.org/10.1044/2015_JSLHR-L-14-0015

**Appendix** (p. 1 of 3)

Available Background Information of the Children per Study

This appendix contains the background information of the typically developing (TD) children and children with developmental language disorder (DLD) that was available in the studies by Boerma et al. (2020), Bol (1995), Bol and Kuiken (1990), Bruinsma et al. (2020), Elbers and Wijnen (1992), van Kampen (2009), and Zwitserlood (2019).

### TD Children

#### Bol and Kuiken (1990)

| ID | Sex | Age (y;m) | Home language | TARSP level |
|---|---|---|---|---|
| TD22 | M | 2;11 | Dutch | 5 |
| TD23 | M | 3;0 | Dutch | 5 |
| TD24 | F | 3;5 | Dutch | 5 |
| TD25 | M | 3;0 | Dutch | 5 |
| TD26 | M | 2;7 | Dutch | 3 |
| TD27 | M | 2;0 | Dutch | 3 |
| TD28 | M | 2;3 | Dutch | 3 |

*Note.* TD = typically developing; M = male; F = female; y = years; m = months; TARSP = Taal Analyse Remediëring en Screening Procedure.

#### Bol (1995)

| ID | Sex | Age (y;m) | Home language | TARSP level |
|---|---|---|---|---|
| TD01 | F | 2;2 | Dutch | 3 |
| TD02 | M | 2;3 | Dutch | 4 |
| TD03 | M | 2;5 | Dutch | 4 |
| TD04 | F | 2;4 | Dutch | 4 |
| TD05 | F | 2;6 | Dutch | 4 |
| TD06 | F | 2;6 | Dutch | 3 |
| TD07 | F | 2;7 | Dutch | 4 |
| TD08 | F | 2;0 | Dutch | 3 |
| TD09 | M | 2;11 | Dutch | 3 |
| TD10 | M | 3;0 | Dutch | 5 |
| TD11 | M | 3;5 | Dutch | 5 |
| TD12 | M | 2;1 | Dutch | 3 |
| TD13 | M | 2;11 | Dutch | 4 |
| TD14 | M | 2;10 | Dutch | 4 |
| TD16 | F | 3;5 | Dutch | 5 |
| TD17 | M | 3;4 | Dutch | 4 |
| TD18 | M | 3;0 | Dutch | 5 |
| TD19 | F | 2;8 | Dutch | 4 |
| TD20 | F | 3;3 | Dutch | 5 |
| TD21 | F | 2;5 | Dutch | 3 |

*Note.* F = female; M = male; y = years; m = months; TARSP = Taal Analyse Remediëring en Screening Procedure.

#### van Kampen (2009)

| ID | Sex | Age (y;m) | Home language | TARSP level |
|---|---|---|---|---|
| TD15 | F | 2;2 | Dutch | 3 |
| TD29 | F | 2;11 | Dutch | 4 |

*Note.* TD = typically developing; F = female; y = years; m = months; TARSP = Taal Analyse Remediëring en Screening Procedure.

**Appendix** (p. 2 of 3)

Available Background Information of the Children per Study

**Elbers and Wijnen (1992)**

| ID | Sex | Age (y;m) | Home language | TARSP level |
|---|---|---|---|---|
| TD30 | M | 3;9 | Dutch | 5 |

*Note.* TD = typically developing; M = male; y = years; m = months; TARSP = Taal Analyse Remediëring en Screening Procedure.

### Children With DLD

**Zwitserlood (2019)**
Only transcripts of the language samples were available.

| ID | Sex | Age (y;m) | Home language | TARSP level | IQ[1] | Schlichting test for language comprehension | Schlichting test for sentence production | PPVT-III-NL[2] |
|---|---|---|---|---|---|---|---|---|
| DLD01 | M | 3;3 | Polish | 3 | 90 | 65 | 77 | 95 |
| DLD02 | M | 3;10 | Dutch | 4 | 95 | 91 | 72 | 88 |
| DLD04 | M | 3;6 | Dutch | 3 | — | 83 | 73 | 71 |
| DLD05 | M | 3;4 | Dutch | 3 | 92 | 80 | 69 | 91 |
| DLD06 | M | 4;1 | Dutch | 3 | 90 | 61 | 69 | 67 |
| DLD08 | M | 3;7 | Dutch | 3 | 86 | 98 | 65 | 90 |
| DLD09 | M | 3;8 | Dutch | 4 | 114 | 88 | 82 | 93 |
| DLD10 | M | 3;5 | Dutch | 5 | — | 94 | 97 | 107 |
| DLD12 | M | 3;6 | Dutch | 4 | 91 | 89 | 91 | 90 |
| DLD13 | M | 2;7 | Dutch | 3 | 86 | 81 | 74 | 67 |
| DLD15 | M | 4;1 | Dutch | 4 | 87 | 72 | 75 | 84 |
| DLD17 | M | 4;2 | Dutch | 3 | 125 | 73 | 81 | — |
| DLD18 | F | 3;9 | Dutch | 4 | 106 | 90 | 85 | 94 |
| DLD19 | F | 3;8 | Dutch | 5 | 116 | 103 | 68 | 100 |
| DLD21 | M | 3;9 | Dutch | 4 | 133 | 103 | 77 | 98 |

*Note.* Em dashes indicate data not available. DLD = developmental language disorder; F = female; M = male; y = years; m = months; TARSP = Taal Analyse Remediëring en Screening Procedure; PPVT-III-NL = Peabody Picture Vocabulary Test, Dutch version.

[1] Nonverbal IQ measured with the SON-R for children aged between 2;6 and 7 years. [2] The Peabody Picture Vocabulary Test, Dutch version by Schlichting (2005).

**Bruinsma et al. (2020)**
Recordings and transcripts of the language samples were shared.

| ID | Sex | Age (y;m) | Home language | TARSP level | IQ[1] | Schlichting test for language comprehension | Schlichting test for sentence production | Schlichting test for lexical comprehension |
|---|---|---|---|---|---|---|---|---|
| DLD23 | F | 4;10 | Dutch | 5 | 74–90 | 77 | 79 | 99 |
| DLD24 | M | 4;6 | Dutch | 4 | 96 | 74 | 75 | 92 |
| DLD26 | F | 5;3 | Dutch | 5 | 96 | — | — | 104 |
| DLD27 | M | 4;5 | Dutch | 5 | 86 | 80 | 72 | 86 |
| DLD28 | M | 5;1 | Dutch as dominant language and English | 5 | 101 | 77 | 74 | 80 |
| DLD29 | M | 3;8 | Dutch | 5 | 102 | 63 | 69 | 105 |

*Note.* Em dashes indicate data not available. DLD = developmental language disorder; F = female; M = male; y = years; m = months; TARSP = Taal Analyse Remediëring en Screening Procedure.

[1] Nonverbal IQ measured with the SON-R for children aged between 2;6 and 7 years.

**Appendix** (p. 3 of 3)

Available Background Information of the Children per Study

**Boerma et al. (2020)**
Recordings and transcripts of the language samples were shared.

| ID | Sex | Age (y;m) | Home language | TARSP level | IQ[1] | Schlichting test for language comprehension | Schlichting test for sentence production | PVTT-III-NL |
|---|---|---|---|---|---|---|---|---|
| DLD22 | M | 3;2 | Dutch | 3 | 94 | 90 | 79 | 101 |
| DLD25 | M | 3;3 | Dutch | 3 | 70 | 84 | 82 | 83 |

*Note.* DLD = developmental language disorder; M = male; y = years; m = months; TARSP = Taal Analyse Remediëring en Screening Procedure; PPVT-III-NL = Peabody Picture Vocabulary Test, Dutch version.

[1] Nonverbal IQ measured with the SON-R for children aged between 2;0 and 8;0.

**Bol and Kuiken (1990)**
No information on the IQ of the children is available, except for the notification that the IQ of the children has been tested and fell within the normal ranges.

| Child | Sex | Age (y;m) | Home language | TARSP level | Verbal Comprehension delay[1] | Other disorders |
|---|---|---|---|---|---|---|
| DLD03 | F | 5;1 | Dutch | 4 | — | Phonological problems |
| DLD07 | F | 5;3 | Dutch | 4 | Verbal comprehension delay of 1;9 | |
| DLD11 | F | 4;9 | Dutch | 4 | — | Dyspraxia |
| DLD14 | M | 5;4 | Dutch | 5 | Verbal comprehension delay of 1;9 | Weak auditory memory |
| DLD16 | M | 4;7 | Dutch | 5 | — | Phonological problems |
| DLD20 | M | 4;8 | Dutch | 5 | — | Dyspraxia |

*Note.* Em dashes indicate data not available. DLD = developmental language disorder; F = female; M = male; y = years; m = months; TARSP = Taal Analyse Remediëring en Screening Procedure.

[1] Verbal comprehension delay was tested with the Dutch adaptation of the Reynell Developmental Language Scales by Boomers and Mugge (1982). The delay is indicated in years and months.