



What Machine Learning Can Tell Us About the Role of Language Dominance in the Diagnostic Accuracy of German LITMUS Non-word and Sentence Repetition Tasks

Lina Abed Ibrahim^{1*} and István Fekete²

¹ Department of English, University of Oldenburg, Oldenburg, Germany, ² Department of Dutch, University of Oldenburg, Oldenburg, Germany

OPEN ACCESS

Edited by:

Esther Rinke,
Goethe-Universität Frankfurt am Main,
Germany

Reviewed by:

Angela Grimm,
Goethe-Universität Frankfurt am Main,
Germany

Sharon Armon-Lotem,
Bar-Ilan University, Israel

Elma Blom,
Utrecht University, Netherlands

*Correspondence:

Lina Abed Ibrahim
lina.abed.ibrahim@uni-oldenburg.de

Specialty section:

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

Received: 05 June 2018

Accepted: 20 December 2018

Published: 30 January 2019

Citation:

Abed Ibrahim L and Fekete I
(2019) What Machine Learning Can
Tell Us About the Role of Language
Dominance in the Diagnostic
Accuracy of German LITMUS
Non-word and Sentence Repetition
Tasks. *Front. Psychol.* 9:2757.
doi: 10.3389/fpsyg.2018.02757

The present study investigates the performance of 21 monolingual and 56 bilingual children aged 5;6–9;0 on German LITMUS-sentence-repetition (SRT; Hamann et al., 2013) and non-word-repetition-tasks (NVRT; Grimm et al., 2014), which were constructed in accordance with the **LITMUS**-principles (**L**anguage **I**mpairment **T**esting in **M**ultilingual **S**ettings; Armon-Lotem et al., 2015). Both tasks incorporate phonologically and syntactically complex structures shown to be cross-linguistically challenging for children with Specific Language Impairment (SLI) and aim at minimizing bias against bilingual children while still being indicative of the presence of language impairment across language combinations (see Marinis and Armon-Lotem, 2015; for sentence-repetition; Chiat, 2015 for non-word-repetition). Given the great variability in bilingual language exposure and the potential effect of language experience on language performance in bilingual children, we examined whether background variables related to bilingualism, particularly, the degree language dominance as measured by relative amount of use and exposure, could compromise the diagnostic accuracy of the German LITMUS-SRT and NVRT. We further investigated whether a combination of the two tasks provides better diagnostic accuracy and helps avoid cases of misdiagnosis. To address this, we used an unsupervised machine learning algorithm, the Partitioning-Around-Medoids (PAM, Kaufman and Rousseeuw, 2009), for deriving a clinical category for the children as \pm language-impaired based on their performance scores on SRT and NVRT (in isolation and combined) while withholding information about their clinical status based on standardized assessment in their first (home language, L1) and second language (societal language, L2). Subsequently, we calculated diagnostic accuracy and used regression analysis to investigate which background variables (age of onset, length of exposure, degree of language dominance, socio-economic-status, and risk factors for SLI) best explained clinical-group-membership yielded from the PAM-analysis based on the children's NVRT and SRT performance scores. Results show that although

language-dominance clearly influences the performance of bilingual typically developing children, especially in the SRT, the diagnostic accuracy of the tools is not compromised by language dominance: while risk factors for SLI were significant predictors for clinical group membership in all models, language dominance did not contribute at all to explaining clinical cluster membership as typically developing or SLI based on any of the combinations of the SRT and NWRT variables. Additionally, results confirm that a combination of SRT scored by correct target structure and the structurally more complex language-dependent part of the NWRT yields better diagnostic accuracy than single measures and is only sensitive to risk factors for SLI and not to dominance levels or SES.

Keywords: bilingualism, specific language impairment, sentence repetition, non-word repetition, language dominance, k-medoid clustering algorithm, unsupervised learning, conditional inference trees

INTRODUCTION

Recent research in language disorders has focused on problems of language assessment and the identification of what is currently referred to in the literature as Developmental Language Disorder (DLD, see Bishop et al., 2017) or Specific Language Impairment (SLI¹) in bilingual children. The latter term refers to a disorder in the development of language in the absence of auditory, cognitive, sensory-motor, neurological, or socio-emotional deficits (Leonard, 1998, 2014). A challenge constantly facing clinicians is to determine whether a bilingual child's poor performance on language tasks in the societal language (second language-L2) is due to an inborn language impairment (LI) or to insufficient exposure to the L2 (cf. Armon-Lotem et al., 2015; Marinis et al., 2017).

A major contributor to the diagnostic difficulties of SLI is the heterogeneity of children with SLI, who constitute a group with diverse linguistic profiles and deficits of varying severity across language components (Crutchley et al., 1997; Conti-Ramsden et al., 2001; Friedmann and Novogrodsky, 2011; Leonard, 2014 among others). For many children with SLI, deficits in the area of morphosyntax (grammatical morphology and syntactic structure) stand out (Leonard, 2007; Marinis and van der Lely, 2007; Marinis, 2011). On the one hand, certain complex syntactic structures with linguistic operations involving dependencies such as syntactic movement (e.g., Wh-questions) and embedding (e.g., relative clauses), have been shown to be cross-linguistically problematic for children with SLI (Jakubowicz et al., 1998; van der Lely, 1998; Friedmann and Novogrodsky, 2011; Jakubowicz, 2011; Hamann and Tuller, 2014; Hamann et al., 2017). On the other hand, SLI may manifest itself differently depending on the language being acquired so that clinical markers vary across languages.

Problems of language-impaired children are not restricted to the morphosyntactic domain, albeit being most deficient. Various studies have shown that children with SLI also evince

deficits in the area of phonology. These children lag behind their age matched peers in the acquisition of consonants and are particularly sensitive to phonological complexity such as consonant clusters (Gallon et al., 2007; Ferré et al., 2015; dos Santos and Ferré, 2018; Grimm and Hübner, in press), coda position (Tamburelli and Jones, 2013) and syllabic position in the foot (Bortolini and Leonard, 2000). As a coping strategy, consonant clusters are often reduced or even avoided (Bortolini and Leonard, 2000; Orsolini et al., 2001; Marshall et al., 2003). Although morphosyntactic and phonological deficits are more commonly reported in the literature (Leonard, 2014), children with SLI also have deficient lexical retrieval abilities, which are not only delayed but also qualitatively different from those of children with typical language development (Novogrodsky and Kreiser, 2015). A number of studies have further shown that children with SLI exhibit deficits in the interface between syntax-semantics and pragmatics, e.g., universal quantification, telicity and exhaustivity in Wh-questions (Roeper, 2004; Schulz and Roeper, 2011). Even though children with SLI often present different combinations of the deficits, Friedmann and Novogrodsky (2008, p. 214) point to the existence of “selective impairments in one module of language, and not in others.” Accordingly, it is possible “to identify subgroups within SLI with selective deficits in various language modules: syntax [grammatical/syntactic-SLI], lexicon [lexical-SLI], phonology [phonological-SLI] and pragmatics [pragmatic-SLI]” (ibid., p. 214).

Aside from the aforementioned language deficits, a large body of research has identified deficits in phonological short-term memory, as indicated by poor performance on repeating non-words with a length of two to four syllables as a special weakness in children with SLI (Gathercole and Baddeley, 1990; Archibald and Gathercole, 2006; Gathercole, 2006; for a meta-analysis see Graf Estes et al., 2007). Although deficits in phonological short-term memory and certain aspects of grammar involving grammatical computational aspects² such as verbal morphology and syntactic comprehension often co-occur in children with SLI, evidence from a twin study by Bishop et al.

¹We are aware of the recent consensus on using the term “Developmental Language Disorder-DLD” for unexplained language impairment in the absence of primary deficits. Nevertheless, we chose to refer to this disorder as SLI in this paper for the sake of continuity with much of the existing literature on bilingual SLI and our own collaborative research within the Franco-German project “BiLaD.”

²Linguistic operations such as recursion and hierarchical non-local dependencies between grammatical elements (van der Lely, 2005, p. 53).

(2006) has shown that despite being significantly heritable, the two vulnerable areas were separable. While some children displayed deficits in both areas, other children displayed deficits in one but not the other, suggesting that they are not “different manifestations of the same underlying deficit” (Leonard, 2014, p. 19).

Apart from diagnostic difficulties caused by the heterogeneity of the disorder, identifying LI in bilingual children is made far more complex by the great variability in their (typical) language development, which is influenced by a multitude of child internal and external factors (Paradis, 2011; Hamann, 2012). The latter include age of onset (AoO) of systematic (sustained) exposure to the second language (L2), length of exposure (LoE), quantity and quality of linguistic input (poor or enriched), L1-L2 typological proximity, status of the home language (high prestige, minority, or heritage language), and socioeconomic status (SES). The interplay of these factors makes it notoriously difficult to establish what is typical for bilingual language development (Tuller et al., 2018). Depending on the timing of exposure, bilingual children could be classified as simultaneous (AoO < 3), early ($3 \leq \text{AoO} < 4$) or late ($\text{AoO} \geq 4$) sequential child bilinguals (also referred to as child L2, Meisel, 2009). Even in simultaneous bilingual language acquisition, bilingual children “have their input space divided” (Paradis and Genesee, 1996, p. 9) and are likely to receive less exposure to each language, on average, than monolingual age peers acquiring the respective languages. As a result, bilingual children often develop unbalanced command of their two languages, i.e., their linguistic abilities are unevenly distributed both within and across language domains at a given age (e.g., Döpke, 2000; Yip and Matthews, 2006; Kohnert, 2010). The language with the more advanced state of development within the process of language acquisition (Deuchar and Muntz, 2003; Genesee and Nicoladis, 2007; Gathercole, 2016) or the language to which the child receives more exposure on a regular basis (Pearson et al., 1997) is commonly described as the dominant (stronger) language as opposed to the weaker or non-dominant one (see also Meisel, 2007). In this sense, dominance is associated with language exposure/use (Grosjean, 2016) and/or with the degree of proficiency in either language (Petersen, 1988; Deuchar and Muntz, 2003; Genesee and Nicoladis, 2007). In the present study, we adopt Argyri and Sorace’s (2007, p. 83) definition of dominance as “the language in which the bilingual child obtains more input on a regular basis” (see also Grosjean, 2010). Language dominance can also shift over time due to changes in patterns of use and exposure resulting from “changes in family structure, child-care arrangements, schooling, or place of residence” (Paradis, 2010: p. 652). For example, in case of early sequential child bilinguals, who start acquiring the societal (second language L2) while their home language (first language, L1) is still at an early developmental stage, a change in the degree of dominance is frequently observed with schooling (cf. Flores, 2015; de Houwer and Bornstein, 2016). Diagnostic problems particularly occur when bilingual children are solely assessed using monolingual norm-referenced tests in the majority/societal language, which might still be their weaker, i.e., non-dominant language at the time

of assessment. In many cases, performance below monolingual average, especially on standardized measures for vocabulary and morphosyntax, is taken as evidence for LI leading to overdiagnosis with SLI (Bedore and Peña, 2008; Grimm and Schulz, 2014).

In addition to the aforementioned quantitative performance differences, a growing body of research has shown that the developmental trajectory of bilingual child language acquisition may show (persistent) delays (Tuller et al., 2015; Paradis et al., 2016) or temporary overlap with that of monolingual children with SLI (MoSLI), particularly in the area of morphosyntax (see Paradis, 2010 for an overview). The overlap in linguistic error patterns of bilingual typically developing children (BiTD) and error patterns serving as diagnostic markers for SLI in a particular language, e.g., extended use of infinitives in English (Rice and Wexler, 1996), object clitic omission in French (Paradis et al., 2003; Paradis, 2010; Hamann, 2012) and problems with SVA combined with the use of infinitives and verb placement errors in German (Clahsen, 1991; Hamann et al., 1998; Rothweiler et al., 2012) complicates the diagnosis of SLI in bilingual children. The delayed or deviant linguistic development of a bilingual child may be erroneously ascribed to bilingualism (underdiagnosis), while a child L2 learner may be overdiagnosed with SLI if such deficits are viewed as a token for SLI (Genesee et al., 2004; Grimm and Schulz, 2014; Armon-Lotem and de Jong, 2015), which could have costly consequences for the child and the society (Zurer-Pearson, 2010).

To avoid cases of misdiagnosis, it has been recommended to evaluate a bilingual child at least in her dominant language (Fredman, 2006) and ideally in both of her languages (American Speech-Language-Hearing Association [ASHA], 2004; Royal College of Speech and Language Therapists Specific Interest Group in Bilingualism [RCSLT], 2007; International Association of Logopedics and Phoniatrics [IALP], 2011), as genuine LI affects both. However, L1-assessment is often not feasible due to the lack of standardized language tests for (bilingual) children in their L1. Even if available, results may be unreliable due to incomplete L1-acquisition and/or L1-attribution, which are often reported for heritage language speakers (Montrul, 2008; Benmamoun et al., 2013). Not to mention that evaluation in two languages is time-consuming and that some of the immigrant L1 varieties undergo language change as a result of contact with the majority/societal language (L2), e.g., Immigrant Turkish in Germany (see Schroeder and Dollnick, 2013; Chilla and Şan, 2017). Hamann and Abed Ibrahim (2017) showed that even when dominance-adjusted bilingual cut-off criteria (Thordardottir, 2015) were applied to the standardized L1 tests, more than a quarter of the L1-dominant children in their sample were classified as SLI by the L1-tests. The fact that the latter children performed within aged-expectations on the L2-tests albeit being dominant in their heritage language questions the applicability of L1 tests in heritage contexts (even with norm adjustments) and suggests that direct assessment measures in the L2 are more reliable for identifying LI in bilingual populations, especially in case of heritage language speakers. This in turn makes it crucial to develop reliable tools that

could disentangle effects of bilingualism and LI in bilingual contexts.

The LITMUS Tools for Bilingual Language Assessment

In an attempt to cope with the diagnostic challenges in bilingual populations, a battery of tools was designed during COST Action IS0804 “Language Impairment in a Multilingual Society: Linguistic Patterns and the Road to Assessment” according to a set of linguistic principles that allow cross-linguistic comparability. These tools aim at minimizing the effect of factors related to bilingualism, so that SLI can be reliably identified in bilingual children with different language combinations. The latter tools are known as the LITMUS tools (Language Impairment Testing in Multilingual Settings, see Armon-Lotem et al., 2015), among which are sentence repetition (SRTs) and non-word repetition tasks (NWRs) and the Questionnaire for Parents of Bilingual Children (PaBiQ; Tuller, 2015). The latter was developed for gathering background information on factors related to bilingualism as well as information about risk factors for SLI. Such information is invaluable for the interpretation of performance results on linguistic tasks. In the current study, we concentrate on sentence repetition and non-word repetition (NWR) since they have been shown to reliably identify SLI in monolinguals (Conti-Ramsden et al., 2001) and to be less reliant on prior language experience than other language measures in bilinguals, e.g., receptive vocabulary (Chiat et al., 2013; Thordardottir and Brandeker, 2013). Depending on their construction, SRTs and NWRs can be designed to not only assess (phonological) working memory (Archibald and Gathercole, 2006), but also the command of syntactic and phonological representations/derivations (see Polišenská et al., 2015 for sentence-repetition; Gallon et al., 2007 for non-word-repetition). Such linguistic representations/derivations, especially their complexity, have been shown to crucially influence performance in these tasks (e.g., Ferré et al., 2012; Friedmann et al., 2015) so that it has been argued that they are not mere measures of working memory (Vinther, 2002; Polišenská et al., 2015). Because of this versatility, they are ideal for targeting language-specific (LS) as well as cross-linguistically challenging syntactic/phonological structures while minimizing avoidance strategies (see Hamann et al., 2017 for SRT).

Sentence repetition taps morphosyntactic abilities as recalling a sentence involves processing of the incoming input string, analysis and reconstruction thereof, especially when the sentences are long enough to prevent mere phonological reiteration (Baddeley, 2000; Marinis and Armon-Lotem, 2015). Furthermore, compared to other types of tasks, it is less constrained by pragmatic and discourse factors (Polišenská et al., 2015; Hamann et al., 2017), and is thus often used in clinical assessment as a measure of sentence-level abilities. The German LITMUS-SRT (Hamann et al., 2013) under investigation here was constructed according to the LITMUS principles (Marinis and Armon-Lotem, 2015) and builds on the notion of linguistic computational complexity. Within the generative framework, computational complexity can be determined by the number

and nature (e.g., merge vs. movement, distance of dependencies, and depth of embedding) of syntactic operations necessary for deriving a syntactic structure (Gibson, 1998; Jakubowicz, 2005; Hamann et al., 2007; Jakubowicz and Tuller, 2008; Friedmann et al., 2009). Children with atypical language acquisition are proposed to have a greater deficit on constructions with a higher degree of computational complexity, as the latter are more taxing to working memory capacities (Chomsky, 2005; Hamann et al., 2007; Jakubowicz and Tuller, 2008). A particular difficulty for children with SLI has been reported for structures involving movement along with intervening elements between the source of the moved constituent and its landing site, e.g., object *Which*-questions and object relative clauses with a lexical subject (Rizzi, 2004; Friedmann et al., 2015). Unlike the problems encountered by children with SLI, bilingual children with typical language development (BiTD) might struggle with vocabulary and uninterpretable features, i.e., grammatical features lacking semantic content like number agreement on the verb (Tsimpli and Dimitrakopoulou, 2007), or might even avoid complexity (Tuller et al., 2015). They are; however, assumed to have an intact language faculty and WM. Thus, having been acquired in the L1, syntactic operations such as recursion, embedding and movement do not have to be acquired again and should not be problematic for them given sufficient exposure to the L2 (Roeper, 2011). Accordingly, the German LITMUS-SRT incorporates a set of syntactically complex, i.e., computationally more demanding structures identified as difficult for children with SLI cross-linguistically in addition to a set of structures reported to be challenging for German MoSLI children such as topicalization and the sentence bracket, which represent crucial milestones in the acquisition of German word-order properties. The complex structures involve computational operations like syntactic movement (measured, for example by number of overt movement operations), in particular *Wh*-movement, i.e., fronting of interrogative or relative pronouns (Hamann et al., 1998; van der Lely, 1998; Marinis and van der Lely, 2007; Jakubowicz, 2011), and/or clausal embedding, e.g., relative clauses (Friedmann and Novogrodsky, 2011; Hamann and Tuller, 2014; Scheidenes and Tuller, 2018).

It has been recently shown that SRTs eliciting structures involving the latter operations can be reliably used to tease apart typically developing bilingual children from monolingual and bilingual children with SLI, not only in bilingual but also in bilingual settings (e.g., Armon-Lotem and Meir, 2016; Meir et al., 2016, 2017 for LITMUS-SRT in Russian and Hebrew; de Almeida et al., 2017; Fleckstein et al., 2018 for French; Lein et al., 2016; Abed Ibrahim and Hamann, 2017; Hamann et al., 2017; Hamann and Abed Ibrahim, 2017 for German; Theodorou et al. (2017) for Cypriot-Greek; see also Marinis et al., 2017 for an overview). In particular, Armon-Lotem and Meir (2016) showed that although the highest level of diagnostic accuracy can be achieved using a combination of SRTs in the child’s L1/Russian and L2/Hebrew (applying bilingual cut-offs), good diagnostic accuracy can still be achieved if SRT is only administered in the societal language (L2-Hebrew). In the same vein, Abed Ibrahim et al. (2018) and Chilla et al. (in press) looked into the potential influence of L1-L2 typological differences on the performance of bilingual children

with Arabic, Portuguese, and Turkish as L1 on German LITMUS-SRT. L1-influence surfaced neither in the overall performance nor in the performance on the individual structures included in the task or in the expected L1-driven error patterns confirming the applicability of the task to bilingual children with diverse L1-backgrounds. It should be; however, noted that most of the studies on LITMUS-SRT report lower-cut-off scores separating TD from SLI in the bilingual groups, and that the task can only be used to assess bilinguals who had at least 12 months of exposure to the L2 (see Tuller et al., 2018).

Non-word repetition belongs to the core assessment measures used for diagnosing LI and has been identified as a reliable clinical marker of SLI in monolingual children (Conti-Ramsden et al., 2001; Gathercole, 2006). An advantage of NWR over other language measures is that it is less affected by prior knowledge of vocabulary and morphosyntax (Thordardottir and Brandeker, 2013; Chiat, 2015) and counts as a relatively culturally fair measure, which could be used for the assessment of children with diverse linguistic and socio-economic backgrounds (Engel et al., 2008; Chiat and Polišenská, 2016). As such, NWR tasks offer promising tools for the identification of SLI especially in bilingual children with limited exposure to the L2.

Measured by increasing numbers of syllables, NWR has traditionally been used to assess phonological working memory (Archibald and Gathercole, 2007; Coady and Evans, 2008). However, the ability to repeat non-words does not only rely on phonological working memory but also requires phonological skills like speech perception, phonological encoding, storage and retrieval of phonological representations, phonological assembly and articulation, which also relate to the capacity of learning new words (Gathercole, 2006). Each of these skills can be deficient in language-impaired children (Coady and Evans, 2008; Marshall, 2014). Recent studies have shown that children with SLI are not only sensitive to the amount of phonological material, i.e., number of syllables in the non-words, but also to phonological complexity such as the presence of consonant clusters, which comprise a particular source of difficulty for children with (phonological) SLI in many languages (Barlow, 2001; Gallon et al., 2007; Marshall and van der Lely, 2009; Ferré et al., 2012; Tamburelli and Jones, 2013; Leonhard, 2014).

Designing an NWRT that identifies LI in bilingual children without disadvantaging those with less experience with the L2 is not straightforward. Despite being less reliant on LS knowledge, there is substantial evidence that performance on NWR (both within and across languages) is affected by the characteristics of the non-words such as word-likeness, length, complexity, prosodic structure, phonotactic probability, and neighborhood density. For instance, children are found to perform significantly better on non-words that are more wordlike, carry LS stress patterns, contain LS-morphemes or have higher phonotactic probability (Jones et al., 2010; Messer et al., 2010; Leclercq et al., 2013; for an overview see Chiat, 2015). These findings imply that “experience and knowledge of lexical phonology contribute to NWR” (Chiat and Polišenská, 2016), which, depending on the nature of the non-words, is generally shown to relate to vocabulary size in monolingual (Gathercole, 2006) and bilingual children (e.g., Engel de Abreu et al., 2013).

Departing from that, different LITMUS-NWRTs manipulating factors shown to influence performance on NWRTs such as length, prosody and/or syllable complexity were constructed within the COST IS0804 framework for NWR (see Chiat, 2015 for details).

Similar to the LITMUS *Crosslinguistic* (Quasi-Universal) NWR test (CL-NWRT, Chiat, 2015), the German LITMUS-NWRT (Grimm et al., 2014) was constructed parallel to the French LITMUS-NWRT (dos Santos and Ferré, 2018) within the COST Action IS0804 framework for NWR tests. Unlike the CL-NWRT, e.g., the Dutch Quasi-Universal NWRT (Boerma et al., 2015; Boerma and Blom, 2017), which primarily tests phonological short-term memory and comprises phonologically simple non-words compatible with the phonological properties of any language, the German LITMUS-NWRT was devised to tap more directly into phonological abilities by focusing on phonological complexity. The latter was found to be a promising marker for assessing phonological impairment (Marshall et al., 2002; Ferré et al., 2012; for German, see Ott et al., 2006). LITMUS-NWRTs of this type systematically vary segmental (articulatory difficulty), syllabic (presence or absence of clusters) and sequential complexity (types of consonant and syllable sequences) combining them into non-words of increasing phonological complexity. At the same time, LS phonological properties are controlled as far as possible to avoid penalizing bilingual children. In order to limit effects of lexical knowledge, the non-words were constructed to be maximally distinct from real words in the target language (German) and were created using elementary blocks (segments and syllables) that are cross-linguistically well-attested (Maddieson et al., 2011). In line with the COST Action IS0804 framework (Chiat, 2015), the latter blocks were combined and manipulated in two sets, a set of phonologically complex items with phonological properties common in most of the world’s languages (the quasi language-independent part, LI_part), and an additional set of items containing the same building blocks of the LI_part in addition to the extrametrical /s/ as a complexity variable specific to German and some other languages (the language dependent³ part, LD_part). The maximum non-word length is limited to three syllables in both parts in order to minimize working memory load, which could undermine the effect of phonological complexity. Various studies reported negative effects of language specific properties of the NWRTs on performance of bilingual children resulting in insufficient diagnostic accuracy, e.g., Kohnert et al. (2006), Windsor et al. (2010), Boerma et al. (2015), and Armon-Lotem and Meir (2016). However, since the construction of the LD_Part in the German LITMUS-NWRT varies considerably from other LS NWRTs (see section “The German LITMUS Non-word Repetition Task”), bilingual children are not expected to be disadvantaged by the LD_part of this particular task. Although they might encounter more difficulties with the LD items, both monolingual and bilingual children with SLI are

³Here, language-dependency is viewed as “an abstract phonological property rather than a lexical or sub-lexical property” (Grimm and Hübner, in press).

anticipated to disproportionately struggle with the structurally more complex LD items since both SLI groups are assumed to have similar underlying deficits (Paradis et al., 2011a,b). Indeed, studies by Ferré et al. (2015), dos Santos and Ferré (2018), Grimm and Hübner (in press), as well as Abed Ibrahim and Hamann (2017) have pointed to the fact that the structurally more complex LD_part of the NWRT did not disadvantage the BiTD children, who performed on par with their monolingual peers. On the contrary, compared to the LI part, the gap between SLI and TD was larger for the LD_part leading to better diagnostic accuracy in both monolingual and bilingual populations. These results corroborate that phonological complexity is vulnerable to phonological deficits not only in monolingual but also in bilingual children.

Several recent studies (e.g., Armon-Lotem and Meir, 2016; Meir et al., 2016; Meir and Armon-Lotem, 2017; Boerma and Blom, 2017; Tuller et al., 2018; Chilla et al., in press) investigated the diagnostic potential and impact of different variables related to bilingualism on the performance in LITMUS-SRTs and NWRTs. Here, we report on three studies of direct relevance to the present research that were conducted within the joint German-French project (BiLaD) using similar methodology with bilingual groups (Arabic/Portuguese/Turkish as L1) in Germany and France, who vary in their sociolinguistic settings. De Almeida et al. (2017) investigated the diagnostic accuracy of French LITMUS-SRT and NWRT and examined whether factors of L2 language use and exposure had an influence on the bilingual children's performance. Although both tasks significantly discriminated between SLI and TD in both monolingual and bilingual children, reduced specificity of SRT was observed for children not dominant in French. Significant correlations were found between SRT-performance and language use and dominance in the BiTD but not in the BiSLI group suggesting that dominance might be responsible for the variation observed in the BiTD group. To avoid cases of overdiagnosis and enhance diagnostic accuracy, the authors recommend combining SRT with NWRT, which did not correlate with any of the L2-exposure variables.

Tuller et al. (2018) report on direct comparisons of German and French LITMUS-NWRTs and SRTs. Their results showed good to excellent diagnostic accuracy in monolinguals, whereas the diagnostic accuracy for bilinguals was fair to good, i.e., the tasks generally distinguished bilingual children likely to be language-impaired from those likely to be typically developing. The authors further explored whether performance on the two tasks was mainly ascribed to developmental risk factors for SLI or to factors related to bilingualism. Results show that a sizable proportion of the variance in the performance of the bilinguals (BiSLI and BiTD collapsed together) in the German and French LITMUS-SRTs and NWRTs was explained by risk factors of SLI as measured by the index of Positive_Early_Development (see section "The LITMUS-Questionnaire for Parents of Bilingual Children" for details). Exposure and use variables such as current L2-richness accounted for additional 4% of the variance in the French-SRT and

11% of the variance in the German SRT. For the German NWRT, early L2-exposure weighed negatively to account for a further 7% of the variance. Since current L2-richness and early exposure to L2 both contribute to establishing language dominance based on the PaBiQ (see section "The LITMUS-Questionnaire for Parents of Bilingual Children"), this raises the question of whether language dominance has a negative impact on the diagnostic accuracy of the LITMUS-tools, especially on the LITMUS-SRT.

This question was further pursued in Hamann and Abed Ibrahim (2017), who used k-means cluster analysis to group bilingual children based on their performance scores on German LITMUS-SRT and NWRT as language impaired or not without access to their clinical group membership based on standardized assessment. In order to measure diagnostic accuracy, the children's k-means cluster membership based on SRT and NWRT scores was compared to the likelihood of a child to have SLI or TD based on standardized assessment in each of the child's languages (see section "Participants" for details). Whereas the sensitivity rates for both SRT (scored by identical repetition, SRT_Id) and NWRT were excellent, the specificity rates were only suggestive, as several bilinguals were assigned to the clinical cluster based on their global NWRT and SRT_Id scores. In line with previous studies on German LITMUS-SRT, this study showed that using the rating measure "target structure" (SRT_Tar), which focuses on the mastery of the constructions targeted by the task, resulted in better specificity and better overall diagnostic accuracy than SRT_Id in the bilingual groups. The individual scores of the children likely to be BiTD were plotted against language dominance for each of the tasks. While NWRT appeared to be rather unaffected by language dominance; 25% of the L1-dominant children performed below cut-off even on SRT_Tar. Finally, the study showed that a combination of SRT and NWRT helps to avoid cases of over-identification.

Given that assessment of bilingual children is usually exclusively carried out in the societal language, the finding that dominance appears to influence the SRT performance of BiTD children, especially those dominant in their L1, raises concerns whether this task is suited for the identification of SLI in L1-dominant children when administered in their weaker language German. However, the three studies above have their limitations: in all of them, diagnostic accuracy of the tools was measured against established clinical status based on standardized evaluation in the L1 and L2, which does not take into account cases of selective impairment or problems with L1 standardized tests in heritage contexts. This, in turn, might be responsible for the reduced accuracy rates (see de Almeida et al., 2017 and Hamann and Abed Ibrahim, 2017 for a discussion). Hamann and Abed Ibrahim (2017) showed that using an alternative procedure that takes into account selective impairments and problems with L1-assessment in minority contexts minimized the slight overlap between BiTD and BiSLI and enhanced diagnostic accuracy. A further limitation is that in both of de Almeida et al. (2017) and Hamann and Abed Ibrahim (2017), dominance was not factored in as a variable into a regression analysis model and might have been confounded by other variables. Hence, the assumed influence of

dominance remains a conjuncture that needs to be statistically validated.

The Present Study

In line with much recent research and building upon our own research, this study investigates the identification of LI in bilingual populations using sentence and nonword repetition tasks. Since both LITMUS-SRT and NWRT were designed to minimize bias against bilingual populations while being indicative of the presence or absence of LI, the following research questions emerge in the light of previous findings:

- i. Upon sufficient exposure to the L2, how robust are German LITMUS-SRT and NWRT against language dominance? Are they only sensitive to risk factors for SLI or could background variables related to bilingualism, in particular the degree of language dominance (estimated by relative amount of use of and exposure to L1/L2), compromise their diagnostic accuracy?
- ii. Since a combination of tools evaluating different aspects of language ability such as morphosyntax and phonology is recommended to acknowledge the heterogeneity within the SLI population and avoid cases of over- and underdiagnosis, does a combination of LITMUS-SRT (especially when scored by correct target structure) and NWRT yield higher accuracy rates than those estimated for each of the tasks in isolation?
- iii. Does a combination of SRT_Tar, which evaluates the mastery of complex constructions and the phonologically more complex LD part of the NWRT provide better diagnostic accuracy for identifying SLI in (monolingual) and bilingual children than other combinations of measures?

To address these questions, we will use an unsupervised machine learning algorithm, the Partitioning Around Medoids (PAM, Kaufman and Rousseeuw, 2009) for deriving a clinical category (clustering) for the children as \pm language-impaired based on their performance scores on SRT and NWRT (in isolation and combined) while withholding information about their clinical status based on standardized assessment in L1 and L2. Subsequently, we will calculate diagnostic accuracy of the tasks (separately and combined) by verifying the goodness of the fit against the clinical groups we can establish for bilinguals by their scores in norm-referenced L1 and L2 tests (see section “Participants”), and use regression analysis to investigate which background variables (age, AoO, LoE, degree of language dominance, SES, and risk factors for SLI) best explained clinical-group-membership based on the children’s NWRT and SRT performance scores. Our premise is that if the PAM-cluster membership can be predicted by the presence of risk factors for SLI but not by any of the other background variables known to influence performance of bilingual children on language tests, particularly the degree of language dominance, then clustering of cases cuts across the SLI/TD dimension confirming that the LITMUS-SRT and NWRT are sensitive to LI and are not biased against bilingual children regardless of their language dominance.

MATERIALS AND METHODS

Establishing Language Dominance in Child Bilinguals

A number of methods have been put forward for measuring and operationalizing language dominance in bilingual children. These measures fall into two categories: performance-based measures and experiential-based measures (Unsworth, 2016; Unsworth et al., 2018). Estimates of language dominance obtained by performance-based measures are based on quantitative differences in proficiency measurements between the two languages of a bilingual. These measures are usually extracted from (a) spontaneous speech data, such as mean length of utterance (MLU), upper bound (UB, length of the longest utterance in a speech sample), multi-morphemic utterances (MMU), lexical diversity measures (number of different word types, verbs, and nouns) and directionality of code-mixing (see Cantone et al., 2008; Kupisch, 2008; Bedore et al., 2012 for an overview), and (b) proficiency measures based on standardized tests for vocabulary and grammar. Experiential measures, on the other hand, rely on biographical information and estimates of language use and exposure to predict dominance in bilingual children. The rationale behind the latter approach is that the (relative) proficiency of bilingual children in each of their languages is “in some sense a function of the amount of language to which they are exposed in these two languages” (Unsworth, 2016, p. 156). Accordingly, experiential variables like the relative amount of language use and exposure can be used as a predictor for the degree of bilingual language dominance.

Bedore et al. (2012), Unsworth (2016) as well as Unsworth et al. (2018) found that relative amount of exposure and use reliably predicted dominance group membership as determined by proficiency measures, confirming that relative amount of use and exposure can be used as a proxy for language dominance in bilingual children. For the purposes of the present study and building upon the findings of Bedore et al. (2012), Unsworth (2016), and Unsworth et al. (2018), we use experiential-based measures to establish language dominance for our participants and calculate this based on the information obtained by the PaBiQ as outlined in “The LITMUS-Questionnaire for Parents of Bilingual Children”.

The LITMUS-Questionnaire for Parents of Bilingual Children

Bilingual children vary considerably in properties of their language exposure and use, which in turn influence the rate and outcome of their language development (e.g., Gathercole and Thomas, 2009; Chondrogianni and Marinis, 2011; Paradis, 2011; Hoff et al., 2012). Thus, having a clear idea about the relative amount of exposure and use for each of the bilingual child’s languages should help professionals to interpret language performance in L1 and L2 adequately and determine whether a child’s (poor) language performance is linked to possible risk factors for LI or to factors related to bilingualism such as the

timing, quality and quantity of exposure to the L1/L2, and degree of language dominance.

In order to gather relevant background information, the Questionnaire for Parents of Bilingual Children (PaBiQ; Tuller, 2015) developed during COST Action IS0804 on the basis of the Alberta Language and Development Questionnaire (ALDeQ, Paradis et al., 2010) and the Alberta Language Environment Questionnaire (ALEQ, Paradis, 2011) was used to interview the parents/legal guardians of the participating children. The parents of participants in the study were interviewed orally in their language of preference by trained native bilingual interviewers familiar with the respective culture.

The PaBiQ incorporates questions about developmental risk factors for SLI, which are synthesized into a global No Risk Index, for which a maximum of 23 points can be attained. This index is arrived at by collapsing the scores of the Positive Early Development index, which is associated with the timing of early language developmental milestones, and the Family History index, which is associated with the presence of oral/written language disabilities in the family. The Positive Early Development index (/14 pts) is calculated by adding up the sub-scores for age of first word (≤ 15 mo = 6 pts; 16–24 mo = 4 pts; > 25 mo = 0 pts), age of first multiword utterances (≤ 24 mo = 6 pts; 25–30 mo = 4 pts; > 31 mo = 0 pts) and early parental concerns (yes = 0 points; no = 2 points). The familiar risk for SLI (/9 pts) is indexed by the existence of first-degree relatives (mother, father, siblings) with reading/writing problems, difficulties understanding others when they speak or difficulties expressing themselves orally. Children with a negative family history of language problems are awarded a maximum of 9 points (3×3 : 1 point per family member per type of language difficulty). Boerma and Blom (2017) investigated the influence of LI and bilingualism on the latter two indices and looked into their diagnostic accuracy. In line with Paradis et al. (2010), they reported strong negative effects of LI on Early Language Development and showed that it was a strong predictor of LI in both monolingual and bilingual children confirming previous findings that a late onset of first words and sentences in at least one language is a risk factor for SLI (cf. de Houwer, 2009; Reilly et al., 2010). With regard to the Family History index, Boerma and Blom (2017) observed a negative effect for LI in the monolingual group but not in the bilingual one and concluded that, due to cultural factors, “Family History as reported by parents may [...] be less reliable as an index of LI in bilingual children than in monolingual children” (p. 73). The Positive Early Development Index also yielded promising diagnostic results in the study by Tuller et al. (2018), who found it be the leading factor explaining performance differences between BiSLI and BiTD in both of the German and French LITMUS non-word and sentence repetition tasks.

The PaBiQ further allows the calculation of a Language Dominance Index (LDI) as a differential between the L1 Exposure Index (relative amount of exposure to the L1) and the L2 Exposure Index (relative amount of exposure to the L2, i.e., German). For each of the child’s languages a total of 50

exposure/use⁴ points could be attained using the German PaBiQ⁵. The Exposure Index is calculated for each of the child’s languages separately based on AoO, LoE⁶, frequency of early language use and exposure⁷, i.e., before the age of four, language richness before the age of four as measured by diversity of language exchange contexts, current language exposure/use within the family, current language use/exposure during different activities within an average week and in exchanges with playmates and family friends. The latter composite score also counts as an estimate of current language richness. An Exposure Index (/50 points) for L1 and L2 emerges by adding up the aforementioned sub-scores. A visual representation of the relative contribution of each of the sub-scores toward establishing the Exposure Index is given in **Figure 1**. As can be seen in **Figure 1**, current language use/exposure contributes the lion share (60%) to the calculation of the Exposure Index and consequently the LDI. This converges with the findings of Bedore et al. (2012) in their large-scale study, in which estimates of current language use (a composite score based on children’s amount of exposure and language output) accounted for 60% of the variance in language dominance patterns of bilingual children.

The language dominance index is then obtained by subtracting the L1 Exposure Index from that of the L2 yielding an estimate of the child’s degree of L2-dominance on a scale from -50 (extremely dominant in the L1) to $+50$ (extremely dominant in the L2). De Almeida et al. (2017, p. 5) compared multiple LDI cut-offs around LDI = 0 (optimal balanced bilingual) against impressions of bilingual investigators of the individual children after interacting with them and their families in both of their languages, and defined cut-off points for language dominance in attempt to explore the use of this variable. An LDI between -5 and $+5$ was set as a cut-off separating dominant from balanced bilinguals. Children with LDIs ranging from -5 to $+5$ are classified as “balanced,” children whose LDI is below -5 are considered to be dominant in the home language, while children with an LDI above $+5$ are classified as dominant in the societal language German.

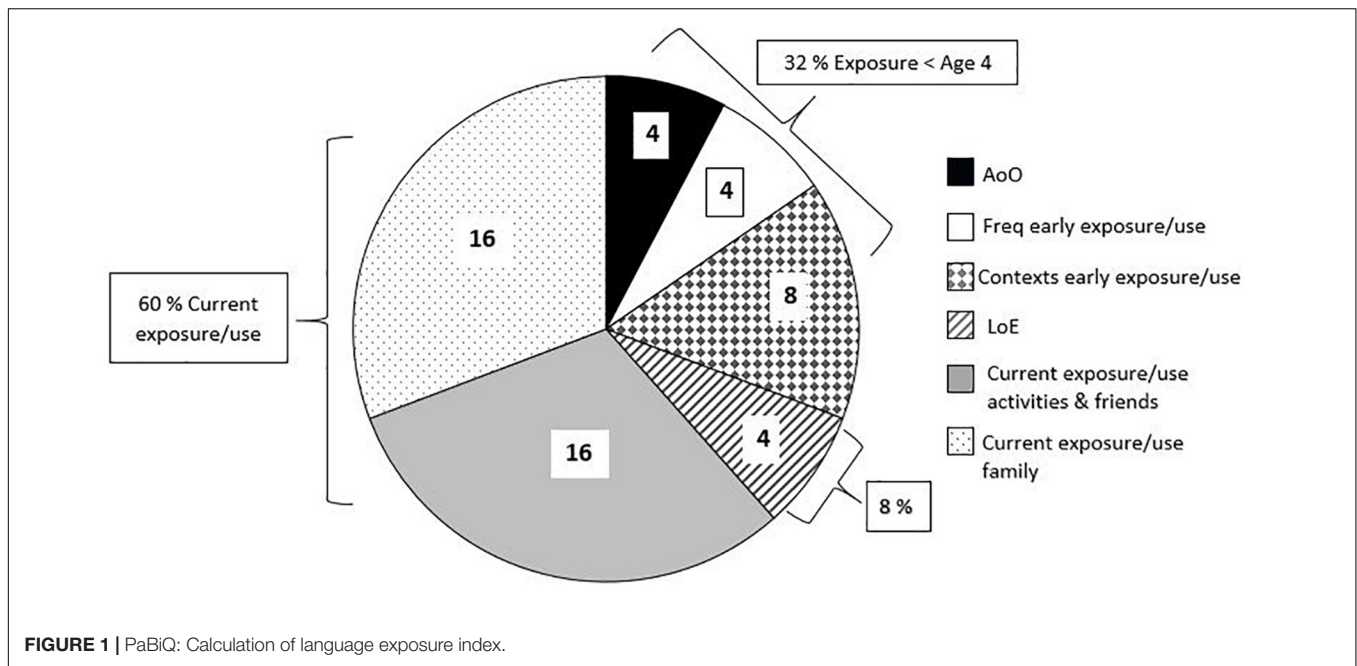
The questionnaire further allows determining the family’s socio-economic status (SES) based on the mother’s and the father’s educational levels. For the purposes of the current paper, maternal rather than paternal educational level (as measured by years of education of the mother) is used as a metric for SES, since the former is reported to be a strong predictor of language development, especially for expressive vocabulary levels, in both monolinguals (Hoff, 2003, 2006) as well as child bilinguals (Paradis, 2009; Calvo and Bialystok, 2014; Paradis and Jia, 2016; Meir and Armon-Lotem, 2017). SES-related language

⁴Input and output are collapsed together (interaction from interlocutor to child and from child to interlocutor).

⁵Five additional points were allotted to number of years in elementary school as part of the exposure indices in France, but not in Germany, where children join elementary school between the age of 6 and 8.

⁶Total length of exposure (LoE) is calculated by subtracting age of onset of systematic sustained exposure to the respective language from the chronological age.

⁷Contrary to Unsworth (2016) and Unsworth et al. (2018), PaBiQ’s language use estimates were not only limited to the “inside home context” but also cover the “outside home context.”



deficits⁸ are reported to have a negative effect on performance in tasks with rich linguistic load, e.g., SRTs and NWRTs with word-like items (Roy et al., 2014; Chiat and Polišenská, 2016).

The German LITMUS Sentence Repetition and Non-word Repetition Tasks

The German LITMUS Sentence Repetition Task

The German LITMUS-SRT (Hamann et al., 2013) used in this study was constructed in close parallel to the French LITMUS-SRT (de Almeida et al., 2017; Fleckstein et al., 2018). It consists of 45 sentences divided in three levels of syntactic complexity (five conditions per level controlled for syllable number, three test items per condition). The degree of an item’s structural complexity relies on the presence of syntactic operations such as Wh-movement, clausal embedding, intervention⁹ – where the latter may add difficulty to the presence of two propositions. Accordingly, level 1 consists of simple declaratives (7–9 syllables) and focuses on Subject-Verb-Agreement (SVA), tense and the sentence bracket[see (1)]. Level 2 (9–13 syllables) includes two types of object questions: bare Wh-questions with the non-D-linked *wh*-operator (*Wen* “*who-masc.-acc.*”), and *Which* NP-questions with the discourse-linked *wh*-operator (*Welchen* “*which-masc.-acc.*”) followed by an intervening lexical noun phrase [see (2a) & (2b)]. Bare Wh-questions are considered to be structurally less complex since

they do not involve intervention. Level 2 further contains non-finite and finite [see (3)] complement clauses. The latter are contrasted with coordinate structures, which serve as control items (two propositions but no embedding). Level 3 (11–12 syllables) comprises the most complex constructions and tests long passives, topicalizations [see (4)] as manifestations of the V2-property¹⁰ of German, subject relative clauses as well as object relative clauses with [see (5)] and without intervening lexical determiner phrases.

Note that German has morphological case marking on accusative masculine singular pronouns, such as the interrogative and relative pronouns in examples 2a, 2b, and 5. **Table 1** gives an overview of test conditions. For more details on German

¹⁰German is a verb-second (V2) language, which requires the finite verb (marked for tense and agreement) to be the second constituent of main (root) clauses. Like other Germanic languages (except for English), it allows different types of constituents (e.g., complements, adverbials) to occupy sentence initial position in main clauses (e.g., topicalization), where only one constituent can linearly precede the finite verb (cf. Chomsky, 1986; Eisenberg, 1999).

TABLE 1 | German LITMUS-SRT: Overview of test conditions.

Level 1	Level 2	Level 3
SVO-Present	Object who-question (bareWH)	Passive
SVO-Simple past	Object which-question (Wh-NP)	Topicalization
Sentence bracket (Aux)	Coordination (Coord)	Subject relatives (SR)
Sentence bracket (Particle)	Non-finite complement Cl	Object relatives without intervener (OR -intv.)
“Werden” control	Finite complement Cl (CompFin)	Object relatives with intervener (OR+intv.)

⁸A number of studies such as Balladares et al. (2016) showed that the influence of SES on repetition tasks diminishes after controlling for vocabulary sizes indicating that the effect of SES is primarily ascribed to smaller vocabulary sizes in children with lower SES.

⁹Intervention can be defined syntactically as in Rizzi (2004, 2013) modeled on work by Gibson (1998) or Gordon et al. (2001).

LITMUS-SRT, we refer to Hamann et al. (2017) and Hamann and Abed Ibrahim (2017).

(1) Sentence bracket:

Der Prinz **hat** die Prinzessin **umarmt**
 The/nom. prince has the/acc. princess hugged
 “The prince hugged the princess”

(2a) Bare WH

Wen beißt der große Löwe immer?
 Who/acc. bites the/nom. big lion always?
 “Who(m) does the big lion always bite?”

(2b) Which-NP

Welchen Bauern ärgert der Affe?
 Which/acc. peasant annoys the/nom. monkey?
 “Which peasant does the monkey annoy?”

(3) Finite complement clause:

Der Wikinger glaubt, **dass** die hexe ihn **mag**.
 The/nom. viking believes, that the/nom. witch him likes
 “The viking believes that the witch likes him”

(4) Topicalization

Den Arzt fotografiert der Bauer gerne
 The/acc. doctor photographs the/nom. peasant gladly
 “The doctor, the peasant photographs gladly”

(5) Object relative with intervention:

Ich sehe den Vogel, **den der Pinguin** weckt.
 I see the/acc. bird who/acc. the/nom. penguin wakes up
 “I see the bird who(m) the penguin wakes up”

The test stimuli are pre-recorded, pseudo-randomized and integrated into a child friendly PowerPoint Presentation. The administration of the task takes about 10 minutes. The task is scored both by identical repetition of test items (SRT_Id), i.e., whole item accuracy, where only phonological errors are disregarded, and by correct target structure (SRT_Tar), which measures whether a particular structure has been mastered or not (see Marinis and Armon-Lotem, 2015 for scoring measures). Although scoring by SRT_Id is faster and easier, L2-errors not affecting the realization of the targeted structure such as lexical substitutions, omissions and systematic recurrent case¹¹ as well as gender errors could surface using this scoring method and penalize bilingual children. Comparison of these scoring methods has indeed shown that SRT_Tar leads to higher diagnostic accuracy of the test for German (see Hamann and Abed Ibrahim, 2017 for particulars).

The German LITMUS Non-word Repetition Task

The German LITMUS-NWRT (Grimm et al., 2014) employed in this study is composed of two parts: a structurally less complex (quasi-) language independent part (NWRT_LI) and

a language dependent part (NWRT_LD) incorporating more complex structural aspects. In both parts the item length ranges from one to three syllables with constant word-initial stress. The 30 items of the LI part were constructed using phonemes and phonotactic constraints attested in the vast majority of the world's languages (Maddieson et al., 2011), i.e., phonemes that are “compatible with cross-linguistically diverse constraints on lexical phonology” (Chiat, 2015, p. 138). Unlike the non-words of the Quasi-Universal-NWRT discussed in Chiat and Polišenská (2016), the non-words of the German LITMUS-NWRT are shorter and are not only composed of simple CV sequences, but also include syllables with initial consonant clusters “#CCV” or closed syllables of the type “CVC#,” which are typologically well-attested albeit their relative complexity (Maddieson, 2006). Throughout the task, phonological complexity is systematically varied at the segmental (consonantal), syllabic (presence of branching onsets or coda) or sequential (position of cluster within the non-word) levels (see dos Santos and Ferré, 2018; Grimm and Hübner, in press for details). The LD part contains 36 items adhering to the same construction principles of the LI part in addition to the extrametrical /s, ʃ/ in word initial and final positions as a complexity feature specific to German (and some other languages, e.g., English and Russian). Such sC sequences violate the Sonority Sequencing Principle and are considered phonologically more complex than other types of onset clusters. Constructed as such, the LD_part is considered to be structurally more complex compared to the LI_part, yet less dependent on LS knowledge than the more traditional Language-Specific NWRTs, e.g., Rispens and Baker (2012), which draw on the full phoneme inventory (consonants and vowels) and include many more properties specific to the target language (Chiat, 2015; Chiat and Polišenská, 2016).

Although structures with higher phonological complexity are generally more error-prone in TD children, they are “disproportionately difficult” for children with SLI (Chiat, 2015, p. 137), who struggle with phonological complexity (Archibald and Gathercole, 2006; Jones et al., 2010; dos Santos and Ferré, 2018). Thus, a greater performance gap between TD and SLI is expected for both monolingual and bilingual children on NWRT_LD, which contains trilateral sCC onset clusters, where /s/ and /ʃ/ represent an appendix to the prosodic word. The latter has been shown to be deficient in phonologically impaired monolingual German children (Ott et al., 2006). An overview of segments and syllable types is given in **Table 2**.

Task administration takes about 5 min and the non-words are presented to the child in a pseudo-randomized order via an animated PowerPoint Presentation. At the beginning of the task, children are provided with noise-canceling headphones and are told that an alien from another planet would appear on the screen and try to teach them his language (format adapted from Engel de Abreu et al., 2013). The test is scored by whole item accuracy (percentage of items correct), since this scoring method is better suited for clinical purposes and has been shown to be informative (Roy and Chiat, 2004; Boerma et al., 2015). A response is rated as correct if all consonants and vowels in addition to their sequencing correspond to the target form. Phoneme omissions, substitutions or additions are regarded as incorrect. Systematic

¹¹Case errors are not disregarded if they are crucial for the realization of the targeted structure, e. g., object relatives and topicalized sentences.

TABLE 2 | Overview of segments and syllable types in German LITMUS-NWRT.

	Vowels	Consonants	Syllable types	Examples
Language-Independent part (LI)	/a, i, u/	/p, k, f, l/	CV CCV CVC#	kapi plaklu pukif
30 items 23 test items 7 controls (e.g., faku, paf)				
Language-Dependent part (LD)	/a, i, u/	/p, k, f, l/	same syllable types plus	
36 items		plus	#sCV	sfikupla
32 test items		/s/	#sCCV	sklipafu
4 controls (e.g., kif, sapi)		/ʃ/	Cs# internal /s/	kapifaps fikuspa

phoneme replacements reflecting articulatory difficulties, e.g., /t/ for /k/ (/kafip/→/tafip/) are not counted as errors. Since the task mainly targets bilingual children, L2-errors such as voicing of consonants (/pilu/→/bilu/) or vowel alternations (/faku/→/fako/) are disregarded. Furthermore, substitution of extrametrical /ʃ/ through [s] or an interdental pronunciation of extrametrical /s/ are not counted as errors since this does not result in a phonemic contrast in extrametrical positions in German (Grimm and Hübner, in press).

Participants

The present study was conducted in line with the compliance form, transaction number 20120416505890730506, of the German Science Foundation and the recommendation of the “Kommission für Forschungsfolgenabschätzung und Ethik” (commission for the evaluation of research consequences and ethics) of the Carl-von-Ossietzky University of Oldenburg (rf. Drs. 21/16/2013). Parents or legal guardians of all participating minors provided written informed consent for both data collection and analysis. The research protocol was approved by the “Kommission für Forschungsfolgenabschätzung und Ethik” of the Carl-von-Ossietzky University of Oldenburg.

Except for 3 children, the current study used the same participant sample as Hamann and Abed Ibrahim (2017), including 77 children, 21 German monolinguals and 56 L2-German bilinguals with Arabic, European Portuguese or Turkish as L1. The latter L1s were chosen because a sizable proportion of immigrants residing in Germany are of Arab, Portuguese and Turkish origin. Furthermore, the typological differences between them and the children’s L2 (German) enable cross-group comparisons, e.g., Abed Ibrahim et al. (2018) and Chilla et al. (in press). The age range of the participants was 5;6–9;0 years covering the last year of kindergarten and the crucial first 2–3 years of primary school. As inclusion criteria for bilingual children, children had to have a minimum L2 exposure of 18 months and be at least functionally bilingual. Thus, children who failed to complete even receptive subtests in the L1 were excluded from the study. 49/56 children were simultaneous bilinguals, while 7 were sequential bilinguals, whose systematic

exposure to L2 mainly started upon kindergarten entry at approximately age three. Almost all of the bilingual participants had a LoE to German of more than 24 months at the time of testing with a mean LoE of 5;1 years (*SD* = 1;10). Children likely to have SLI, i.e., with a clinical diagnosis of SLI, were recruited from specialized speech-language pathology centers and kindergartens with special inclusion programs from different parts of Germany. Given the high rates of over- and under-referral of bilingual children to speech language therapy (Grimm and Schulz, 2014), an extensive procedure based on standardized evaluation in each of the child’s languages was applied in order to verify the clinical status of all recruited bilingual children as ± language-impaired. The verification of clinical status was done in accordance with the recommendations of the COST Action IS0804 assessment committee as outlined in Thordardottir (2015, p. 343) and began with a control for non-verbal intelligence using the German version of Raven’s Colored Progressive Matrices (CPM; Bulheller and Häcker, 2002). Only Children who had a non-verbal IQ score ≥ 80 were included in the study. In addition to standardized assessment, narrative samples were collected from each child in both of her languages using the picture materials provided by the LITMUS-Multilingual Assessment Instrument for Narratives (MAIN, Gagarina et al., 2015). The collection of the narrative samples was done in accordance with the MAIN protocol (story telling). However, for the purposes of the current study, the latter samples were not analyzed in terms of narrative macro- and microstructure, but were rather used as spontaneous speech samples. Especially in borderline cases, the latter samples were consulted in order to gain an impression¹² about the child’s expressive language abilities in both of her languages and look for clinical markers for SLI, e.g., SVA errors, the use of infinitives and verb placement errors in German (Clahsen, 1991; Rice et al., 1997; Hamann et al., 1998; Lindener, 2002).

As to assessment using formal tests, in our previous work, e.g., Hamann and Abed Ibrahim (2017), Tuller et al. (2018), and Chilla et al. (in press), we adapted the criteria outlined in Leonard (2014) to bilinguals using Thordardottir’s (2015) recommendations and assigned a child to the BiSLI group if she scored below dominance-adjusted¹³ norms in two language domains (on norm-referenced tests) in both of her L1 and L2. Five language areas relevant in this context were evaluated in each of the child’s languages (except for Turkish): phonology, morphosyntax comprehension and production as well as receptive and expressive vocabulary (see also Tomblin et al., 1996). Since expressive vocabulary is a notorious locus of difficulty for bilingual children, we counted lexicon as a single domain and considered the child unimpaired in this domain

¹²The narrative samples of the children were evaluated by linguistically trained native speakers (L1 and L2) according to certain markers: e.g., subject-verb-agreement and sentential complexity, i.e., presence of embeddings.

¹³Following the recommendation of Thordardottir (2012, 2015), the monolingual –1.25 SD cut-off criterion used by Tomblin et al. (1996) was adapted according to the dominance status of the language being assessed. Accordingly, we used a criterion of –1,5 SD if the child was evaluated in her dominant language, and a cut-off of –2,25 SD if the child’s weaker language was being assessed. In case of balanced bilinguals, the cut-off criterion was set at –1,75 SD for both languages.

TABLE 3 | Overview of norm-referenced tests employed for standardized language assessment in Arabic, German, European Portuguese, and Turkish.

Language	Test	Language skill evaluated					Method of scoring	Age range
		Phonology	Reception vocabulary	Expression vocabulary	Morphosyntax comprehension	Morphosyntax production		
Arabic	ELO-L ^a	Word repetition	Picture selection	Picture naming	Picture-sentence matching	Sentence completion	Individual subtest scores	3;0–7;11
German	WWT 6–10 ^b	–	Picture selection	Picture naming	–	–	Individual subtest scores	5;6–10;11
	LiSe-DaZ ^c	–	–	–	Picture-sentence matching, TVJT	Story, sentence completion, lead-in questions	Individual subtest scores	Monolinguals: 3;0–6;11 Bilingual: 3;0–7;11
	PLAKSS-II ^d	Picture naming	–	–	–	–	Individual subtest scores	2;6–7;11
European Port.	PALPA-P ^e	Non-word repetition	Picture selection	Picture naming	Picture selection	Sentence repetition	Individual subtest scores	5;0–9;11 (with missing norms for some age ranges for all tasks)
	GOL-E ^f	–	Word definition	Antonyms naming	–	Complex S from two simple S's	Individual subtest scores and global score	5;07–10;00
Turkish	TEDIL ^g	–	Picture selection	Picture naming	Picture Selection	Sentence completion/ construction	2 composite scores, 1 production and 1 comprehension	2;0–7;11

^aZebib et al. (2017); ^bGlück (2011); ^cSchulz and Tracy (2011); ^dFox-Boyer (2014); ^eCastro et al. (2007); ^fSua-Kay and Santos (2014); and ^gTopbaş and Güven (2013); TVJT, truth value judgment task.

if only receptive vocabulary was above the respective cut-off. For the assessment of L1 and L2, we chose norm-referenced L1 and L2 tests frequently used by speech language pathologists and cover the age range¹⁴ under investigation (see **Table 3** for a detailed overview of standardized assessment tools). For German, we selected the LiSe-DaZ (Schulz and Tracy, 2011), which provides bilingual and monolingual norms, for assessing morphosyntax. The short form of the WWT (Glück, 2011) was used to assess receptive and expressive vocabulary, and the screening version of the PLAKSS-II (Fox-Boyer, 2014) was used to evaluate phonology. We tried to assess the same language domains in Arabic, Portuguese and Turkish. For Arabic, this was possible using the comprehensive test battery ELO-L (Zebib et al., 2017), which offers norms for Lebanese Arabic and was adapted to a number of other varieties of Arabic¹⁵ by the test authors in collaboration with linguistically trained native speakers of the respective varieties (Algerian, Iraqi, Libyan, Moroccan, Palestinian, Syrian, and Tunisian). We used the PALPA-P test battery (Castro et al., 2007) for Portuguese. One major limitation

of the PALPA-P is that it lacks norms for some of the age ranges we are investigating for the lexical domain. As a result, we chose to assess receptive and expressive vocabulary using subtests of the GOL-E (Sua-Kay and Santos, 2014), which covers our entire age range, and used subtests of the PALPA-P to assess phonology and morphosyntax. For Turkish, we chose the TEDIL (Topbaş and Güven, 2013), which measures morphosyntactic comprehension and production as well as lexical semantics. The test; however, does not include a subtest for phonology and does not offer norms for the individual subdomains. Instead, a composite score exists for each of comprehension and production collapsing morphosyntax and lexical semantics together. As the Turkish test merely offers a single production and a single comprehension score, encompassing two domains each, a child was assigned to the BiSLI group if she scored below cut-off in either production or comprehension. For a detailed description of standardized assessment L1-L2-tests and a complete overview of recruitment and classification procedure of bilingual children into TD vs. SLI, we refer to Hamann and Abed Ibrahim (2017).

Following the argumentation in Hamann and Abed Ibrahim (2017, p. 16) about problems encountered with standardized L1 tests in heritage contexts, and since our previous classification procedure did not isolate subgroups of SLI and might have missed cases of selective impairment such as grammatical/syntactic, phonological or lexical SLI (cf. Friedmann and Novogrodsky, 2008), we adopted Hamann and Abed Ibrahim's (2017) modified

¹⁴In case of children older than the norming sample, we consulted with the test authors concerning the possibility of norm-extension, e.g., LiSe-DaZ and ELO-L.

¹⁵Norms are only available for the Lebanese version. Due to linguistic proximity between Lebanese, Syrian, Jordanian and Palestinian Arabic, the Lebanese norms can be applied to the latter varieties with the caveat that the socio-cultural context may differ. In case of Maghreb (Moroccan, Tunisian, and Algerian) dialects, norms should be viewed with caution, especially in borderline cases.

TABLE 4 | Participants including monolingual controls and bilinguals after verification of clinical status (Mean, SD and range¹⁶).

Simult./total	MoTD (n = 10)	MoSLI (n = 11)	BiTD (n = 44)				BiSLI (n = 12)
			BiTD-A (n = 10)	BiTD-P (n = 18)	BiTD-T (n = 16)	Total (n = 44) 38/44	
Age at testing (in months)	75.90 (8.99) 66–92	79.00 (9.89) 68–98	88.60 (13.5) 70–108	83.44 (14.72) 66–108	86.68 (13.01) 70–104	85.6 (13.60) 66–108	80.66 (15.05) 64–108
Age of onset (in months)	0	0	39 (13.61) 24–75	17.61 (24.39) 0–90	25.31 (15.52) 0–48	25.27 (20.65) 0–90	16.00 (17.03) 0–36
Length of exposure (Gr.) (in months)	75.90 (8.99) 66–92	79.00 (9.89) 68–98	53.2 (19.68) 32–97	67.97 (24.59) 18–101	61.37 (19.01) 34–96	61.73 (21.87) 18–101	62.00 (22.77) 30–88
CPM (PR)	81.20 (13.98) 56–100	53.72 (24.50) 25–99	44 (29.13) 9–93	73.77 (17.69) 42–94	71.18 (22.23) 38–100	66.06 (25.01) 9–100	55.08 (26.69) 27–98
LDI (/50)	N/A	N/A	–7 (9.34) –25–11	1.44 (13.76) –25–23	–5.75 (12.57) –27–18	–3.09 (12.77) –27–23	0.58 (12.37) –21–24
L1-dominant (no./total)	N/A	N/A	(5/10)	(8/18)	(8/16)	(21/44)	(3/12)
Balanced (no./total)	N/A	N/A	(1/10)	(3/18)	(4/16)	(8/44)	(5/12)
L2-dominant (no./total)	N/A	N/A	(4/10)	(7/18)	(4/16)	(15/44)	(4/12)
Yrs. educ. mother	N/A	N/A	14.2 (1.68) 12–18	13 (3.91) 4–18	13.68 (5.23) 5–20	13.52 (4.05) 4–20	11.83 (2.75) 8–16

¹⁶When applicable.

“criteria for the identification of the bilingual clinical group” in this paper. Accordingly, we assigned a child to the BiSLI group if she had a selective impairment in the L2, i.e., if she performed below the dominance-adjusted cut-off in either morphosyntax or receptive vocabulary or phonology (not necessarily two domains in combination), and scored below norms in two domains in her L1 (one domain for Turkish) or showed poor performance of spontaneous production in both of her L1 and L2. **Table 4** gives a participant overview based on clinical status as verified by the modified procedure described above and also includes the two monolingual control groups MoSLI and monolingual typically developing children (MoTD). By applying the modified classification criteria, the clinical status of 4 children who were initially classified as BiTD in Hamann and Abed Ibrahim (2017) changed to BiSLI¹⁶. In **Table 4**, the BiTD children are divided into subgroups based on their L1: Arabic = BiTD-A, Portuguese = BiTD-P and Turkish = BiTD-T. The BiSLI group is composed of 12 children (4 with L1 Arabic, 3 with L1 Portuguese and 5 with L1 Turkish). Due to the relatively small sample size, the BiSLI children are grouped together regardless of their home language. The bilingual children are further classified according to language dominance as measured by the PaBiQ (see section

“The LITMUS-Questionnaire for Parents of Bilingual Children”). As can be seen in **Table 4**, almost half of the children in the BiTD group (21/44) are dominant in their L1, whereas the majority of the BiSLI children (9/12) are either balanced or L2-dominant¹⁷.

The four groups (MoTD, MoSLI, BiTD, and BiSLI) were comparable in terms of non-language variables such as chronological age and non-verbal intelligence. Concerning age, the overall effect of Group was not significant, as revealed by Kruskal–Wallis test [$\chi^2(3, N = 77) = 5.505, p = 0.138, \eta^2 = 0.034$]. This also holds when the BiTD group is split into three subgroups by L1 [$\chi^2(5, N = 77) = 6.758, p = 0.239, \eta^2 = 0.051$]. In terms on non-verbal intelligence, the overall effect of Group was significant [$\chi^2(3, N = 77) = 8.448, p = 0.038, \eta^2 = 0.075$]. However, subsequent pairwise comparisons using Mann–Whitney U tests controlling for false positives, that is Type I error, revealed only one marginally significant comparison, namely MoSLI vs. MoTD ($U = 19.00, p = 0.06, r = 0.553$, Bonferroni-corrected). Yet, all of the children belonging to the MoSLI group have normal non-verbal intelligence. We further checked whether the bilingual groups were comparable concerning SES, AoO, LoE, and degree of L2-dominance (LDI). No significant differences emerged between BiTD and BiSLI concerning SES [$\chi^2(1, N = 56) = 2.228, p = 0.135, \eta^2 = 0.041$], AoO [$\chi^2(1, N = 56) = 3.261, p = 0.071,$

¹⁶Two of them were cases of selective impairment in the L2, i.e., they performed below cut-off in the L1, and only showed deficits in morphosyntax in the L2 (grammatical SLI) plus SLI markers in the speech samples in both languages. The other two were L2-dominant and performed below-cut-off in all of the domains in the L2, but slightly above cut-off in the L1. Since SLI markers were present in their speech samples, we classified them as BiSLI

¹⁷This reflects the advice frequently given to parents of bilingual children with atypical language development that they should restrict parent-child interactions to the societal language to avoid aggravating the existing language difficulties, which in turn means less exposure to the L1.

$\eta^2 = 0.059$], LoE [$\chi^2(1, N = 56) = 0.615, p = 0.433, \eta^2 = 0.011$], and LDI [$\chi^2(1, N = 56) = 1.912, p = 0.167, \eta^2 = 0.035$]. This also holds when the BiTDs are split by L1 SES [$\chi^2(3, N = 56) = 3.216, p = 0.360, \eta^2 = 0.06$], LoE [$\chi^2(3, N = 56) = 3.640, p = 0.303, \eta^2 = 0.07$] and LDI [$\chi^2(3, N = 56) = 4.457, p = 0.216, \eta^2 = 0.08$]. With respect to AoO, the overall effect of Group was significant when BiTDs were divided by L1 into three subgroups L1 [$\chi^2(3, N = 56) = 11.833, p = 0.008, \eta^2 = 0.17$]. Mann-Whitney U tests applying Bonferroni-adjustment of p -values revealed significant differences in AoO between BiTD-A and BiTD-P ($U = 33.00, p < 0.05, r = 0.531$) as well as between BiTD-A and BiSLI ($U = 17.00, p < 0.05, r = 0.617$). Nevertheless, the overall effect of Group was not significant when the BiTD groups were collapsed together [$\chi^2(1, N = 56) = 3.261, p = 0.071, \eta^2 = 0.059$].

Data Analysis

The children's responses on the SRT and NWRT were recorded using special dictaphones. Data transcription, verification and coding for errors were done offline by two independent linguistically trained raters (percentage of agreement was at least 90%). For each repetition measure, the percentage of correct responses was used as basis for data analysis. Null reactions were counted as errors, unless they were due to technical problems or errors by the investigators (missing data, less than 1% of the overall data).

IBM SPSS 24 (2016) and R-Studio (2012) were used to conduct statistical analyses. Non-parametric tests were used for group comparisons due to unequal sample sizes and the violation of the normality assumption, checked by the Shapiro-Wilk test. Since we wanted to investigate whether the LITMUS repetition tools are suitable for assessment of bilingual children in their weaker language, we first checked for group differences between L1-dominant BiTDs and their monolingual, balanced and L2-dominant TD peers, and whether performance of L1-dominant BiTDs overlapped with that of MoSLIs and BiSLIs. Here, we split the BiTDs into three subgroups based on LDI as established in the section "The LITMUS-Questionnaire for Parents of Bilingual Children".¹⁸ and ran Kruskal-Wallis tests and Mann-Whitney U tests with Bonferroni-adjustment. Recall that BiSLIs were collapsed into a single group due to the small sample size. Since performance of BiTDs on SRT appeared to be influenced by dominance, we ran partial correlation analysis controlling for age on their SRT_Id and SRT_Tar. In addition to language dominance, we also checked for correlations with AoO, LoE and SES, since they are factors known to influence performance on linguistic tasks. Next, linear regression models for predicting performance of the BiTDs on SRT_Id and SRT_Tar were built using the variables that yielded significant correlations.

Secondly, we applied cluster analysis to the data in order to automatically group the children into \pm language-impaired based on their performance scores on the SRT (SRT_Id, SRT_Tar) and NWRT (NWRT_global, NWRT_LI, NWRT_LD), separately and then in combination. A clustering algorithm classifies a dataset into several meaningful homogenous sub-categories - so-called clusters (i.e., TD vs. SLI in this study) - based on the values of their

attributes (i.e., linguistic variables in the present study) such that the similarity¹⁹ among objects within a category is larger than that between categories. We opted for unsupervised learning (cluster analysis) for verifying diagnostic accuracy and establishing cut-off points separating TD from SLI on the tasks, since it does not use predefined clinical status during the statistical analysis, and is thus unbiased by any given classification of participants.

Because children were measured based on performance scores on LITMUS-SRT and NWRT designed to identify SLI without penalizing bilinguals, our premise was that SLI-cases would be similar to each other, and hence group together, while TD-cases would form their own cluster regardless of bilingualism. Different from Hamann and Abed Ibrahim (2017), we chose the PAM (Partitioning Around Medoids) non-hierarchical k-medoid clustering method (Kaufman and Rousseeuw, 1987, 2009) over k-means, because it is a suitable method for small datasets with up to approximately 60 objects, and because it can handle noisy data and outliers (Kaufman and Rousseeuw, 1987, 2009; Kashef and Kamel, 2008; Patel and Singh, 2013; Soni and Patel, 2017). Variables were scaled for normalization purposes in the course of the PAM-analysis. We used the function *pam* of the *cluster* R package (Maechler et al., 2017).

We used *Hopkins statistic* (H) based on the *factoextra* R package (Kassambara and Mundt, 2017) as a measure of cluster tendency to assess clusterability (Hopkins and Skellam, 1954). If the H-value is close to zero, and far below 0.5, then the dataset is clusterable (Kassambara and Mundt, 2017; Krishna et al., 2018). Because H is run on the created random dataset every time, we get fluctuations in the H-values if we run the statistics multiple times. Banerjee and Davé (2004) demonstrate that random data sets, clustered data sets and regularly spaced data sets show H-values of around 0.5, 0.7–0.99 and 0.01–0.3, respectively.

Because the k-medoid algorithm requires that the number of clusters should be pre-defined, we first ran the *Gap Statistic* (Tibshirani et al., 2001) to determine the optimal number of clusters. The *Gap Statistic* compares the change in within-cluster dispersion for each clustering solution (at each number of clusters) to that expected at random distribution. We used the functions *fviz_nbclust* of the *Factoextra* R package (Kassambara and Mundt, 2017) and *NbClust* of the *NbClust* R package (Charrad et al., 2014) to determine the optimal number of clusters.

The k-medoid algorithm selects one of the members of the cluster as the most representative object, named *cluster medoid*, so that each cluster has only one medoid. By choosing an actual case (i.e., an SLI or a TD child) as the cluster medoid, the k-medoid method is less sensitive to outliers, as mentioned before. The optimal cluster is achieved by minimizing the sum of squared Euclidean distances to the medoid in each cluster, also called the error sum of squares (Kaufman and Rousseeuw, 1987). First, in the so-called "Build-step," the k-medoid algorithm selects k medoids randomly, with k being the optimal number of clusters. Next, a matrix of dissimilarity is calculated from the raw data and the algorithm assigns every object to either of the

¹⁸Note that language dominance was used as a categorical variable in this step.

¹⁹The notion of similarity in the clustering approach is operationalized as Euclidean distance.

k clusters based on their distance to the nearest medoid (Patel and Singh, 2013). The sum of absolute error in the clustering procedure is equal to the sum of the distances between data points and their medoids. In the so-called “Swap-step,” each non-medoid object is tested as a potential medoid within each cluster by checking if the sum of within-cluster distances gets smaller if that object is used as the new medoid. If this is the case, then that configuration is used. The algorithm checks at each iteration step, if the solution is better than the previous one. If the medoids do not change, the algorithm terminates (see Patel and Singh, 2013 for details).

Because the medoid of each cluster can be seen as a prototype of that cluster, identifying the medoid can serve as a cue to interpret the cluster. For example, if the medoid of a cluster was originally diagnosed as an SLI-case, then that cluster represents most probably the SLI-cases. We expected the SLI-cluster to contain the majority of the children classified originally as SLI based on standardized assessment, while the majority of TD-cases would reside in the other larger cluster. Our further premise was that the cluster with the lower scores on the linguistic variables would represent the cluster with LI, since language-impaired cases score lower on the linguistic variables.

After clustering the sample, we determined the estimated cut-offs on the linguistic variables (i.e., SRT and NWRT) between the SLI- and TD-clusters based on the clustering result. A cut-off is a value of a variable which can be seen as the best threshold score to separate the cases belonging to the two categories using that variable. If the two categories can be best separated along multiple variables simultaneously, e.g., SRT and NWRT combined, then cases can be predicted (as TD vs. SLI in our study) based on multiple cut-offs on these variables. To this end, we employed conditional inference tree models (Tagliamonte and Baayen, 2012). Conditional inference trees (ctrees) are non-parametric regression models visualized as decision trees. They are suitable for our dataset because of the presence of high-order interactions among the variables and the overall small sample-size compared to the number of predictors (Levshina, 2015). Besides determining the cut-off for the linguistic variables, ctrees can also give information about the hierarchical structure of the relevant predictors of cluster membership, i.e., about variable importance. For instance, if clustering is based on several linguistic variables such as SRT_Id, SRT_Tar and NWRT_global, decision trees can show which one contributed the most toward predicting cluster membership as TD or SLI. The higher the variable in the hierarchy, the more important it is, with the highest-level variable being the most important. If there are multiple variables in the ctree, then a multi-hierarchy predicts the outcome (i.e., cluster membership as TD or SLI). Ctrees were implemented with the *party* R package using the *ctree* function (see Hothorn et al., 2006 for details).

In order to address research questions (ii) and (iii), we calculated diagnostic accuracy²⁰ for the SRT and NWRT measures separately and combined. Sensitivity and specificity levels were

estimated by comparing cluster membership of each of the children as TD vs. SLI as assigned by PAM on the basis of LITMUS-SRT and NWRT results to their clinical status (as established by the standardized assessment procedure described in “Participants”). Sensitivity is determined by the proportion of children with LI identified as such by LITMUS SRT and NWRT or subtests thereof (i.e., assigned to the clinical cluster in our case), while specificity is computed based on the proportion of children with typical language development identified as such by our tests, i.e., assigned to the non-clinical cluster (Oetting et al., 2008; Dollaghan and Horner, 2011). In addition, likelihood ratios²¹ (LRs) were calculated based on the obtained sensitivity and specificity levels. An advantage of LRs is that they are less likely to be affected by variations in the properties of the test sample (Dollaghan and Horner, 2011). LR+, positive likelihood ratio [sensitivity/(1-specificity)], indicates how likely it is that a score below a cut-off criterion to be present in language-impaired children, whilst an LR-, negative likelihood ratio ((1-sensitivity)/specificity), is indicative of the likelihood of a score above a cut-off criterion to belong to a child without LI.

To answer research question (i), we investigated which of the background information variables provided by the PaBiQ as cogent confounders predicted cluster membership following each clustering procedure based on SRT and NWRT measures or combinations thereof. The hypothesis to be tested was that cluster membership as TD or SLI based on performance scores in LITMUS-NWRT and/or SRT can only be explained by variables concerning risk factors to SLI and not by background information variables related to bilingualism, particularly the degree of language dominance. If this hypothesis is confirmed, then the clustering of the cases cuts across the SLI/TD dimension rather than any of the background information variables unrelated to risk factors for SLI validating that the diagnostic accuracy of the tasks is not compromised by language dominance. To that end, we ran Firth’s Bias-Reduced Binary Logistic Regression (Firth, 1993), which uses penalized ML²². Cluster membership (TD or SLI) served as the dependent measure. Models with Firth’s correction were built using the *Brglm2* R package (Kosmidis, 2018). We included only a maximum of four background information variables as fixed factors in the model to avoid over-parametrization given the overall small sample size. Because regression analysis provides a way of adjusting for potentially confounding covariates included in the model, we entered the covariates into the model at once. To examine

²¹LR+ values ≥ 10 are highly indicative of the presence of language impairment, LR- values ≤ 0.10 highly indicate the absence of an impairment, LR+ values ≥ 3.0 and LR- values ≤ 0.3 are considered to be clinically suggestive, while LR+ values < 3.0 and LR- values > 0.3 are viewed as clinically uninformative (cf. Dollaghan and Horner, 2011).

²²We used Firth’s Bias-Reduced Logistic Regression to step around the following statistical concerns. Given our small sample, Long (1997) advises against maximum likelihood estimation in logistic regression with less than 100 cases. A second confound is that in small samples maximum likelihood estimates in binary logistic regression models are not powerful because there can be a complete- or quasi-complete separation along one covariate (Rainey, 2016). That is, for a given combination of covariates the outcome can be predicted perfectly. To avoid separation, Peduzzi et al. (1996) suggest that the number of positive outcome events (i.e., the smaller number of binary outcomes) divided by the number of independent variables should be more than 10.

²⁰Following Plante and Vance (1994), good diagnostic accuracy is given when specificity and/or sensitivity rates are $\geq 90\%$. Rates between 80 and 90% are viewed as fair.

whether the diagnostic accuracy of the tasks is not compromised by language dominance and is only sensitive to risk factors for SLI, we built several regression models using Firth’s correction with PAM cluster membership as TD or SLI as the dependent variable. In each model, we entered LDI and the index of Positive_Early_Development (risk factors for SLI) in addition to two further background information variables reported to explain performance on LITMUS-SRT and NWRT (see Tuller et al., 2018) as covariates. The latter variables included AoO, LoE, SES. We also included chronological age as a covariate since working memory and cognitive capacities are rapidly growing in children and since language abilities of children tend to improve over time.

Background information variables were first scaled by the mean of their original variable to remove potential *non-essential* multi-collinearity between them (Dalal and Zickar, 2011) and to adjust the interpretation of the coefficients. Multi-collinearity among covariates was checked using the *Variance Inflation Factor (VIF)* after scaling, with a VIF value above 10 indicating serious multi-collinearity (Kutner et al., 2004). Correlations between the background variables are given in the **Appendix**.

RESULTS

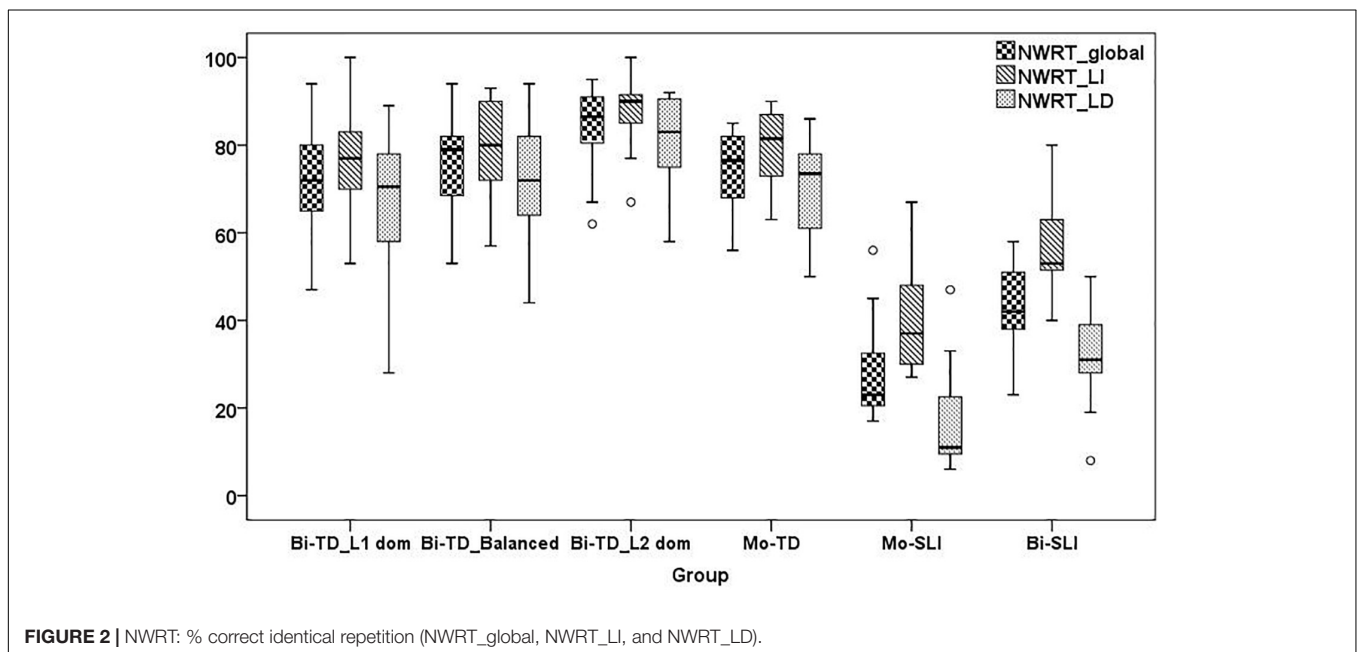
Overall Results on the German LITMUS NWRT and SRT

Kruskal-Wallis tests comparing performance scores of L1-dominant BiTDs to the other groups (MoTD, balanced-BiTD, L2-dominant-BiTD, MoSLI, and BiSLI) on NWRT_global, NWRT_LI, NWRT_LD, SRT_Id and SRT_Tar yielded significant results for all measures as shown in **Table 5**.

Subsequently, pairwise comparisons were carried out using Mann-Whitney U tests with Bonferroni-adjustment. Typically

developing children performed significantly better than their language-impaired counterparts on all measures. All measures distinguish between MoTDs and MoSLIs as well as between BiTDs and BiSLIs regardless of language dominance: Moreover, all of the BiTD groups significantly outperformed MoSLIs. The comparisons yielded no significant differences between MoSLIs and BiSLIs on any of the aforementioned measures. Comparing MoTDs to the BiTDs split by dominance revealed no significant differences between MoTDs and balanced as well as L2-dominant BiTDs on either measure. Nevertheless, significant differences with large effect sizes were found between MoTDs and L1-dominant BiTDs as well as between L1-dominant and L2-dominant BiTDs for both SRT_Id and SRT_Tar but not for any of the NWRT measures (see **Table 5**). It should, however, be stressed that despite the observed significant differences in SRT_Id and SRT_Tar, L1-dominant BiTDs performed significantly better than MoSLIs and BiSLIs on both SRT measures. **Figures 2 and 3** depict the overall performance of the groups in NWRT and SRT, respectively. An overview of significant pairwise comparisons is provided in **Table 5**.

In the next step, we collapsed all of the BiTDs into one group and ran partial correlation analysis controlling for age on their performance in SRT_Id and SRT_Tar and variables shown to influence performance on LITMUS-SRT including language dominance (see Tuller et al., 2018). Moderate positive correlations were found between LDI and performance on SRT_Id ($r = 0.542, p < 0.001$) and SRT_Tar ($r = 0.586, p < 0.001$), as well as SES and SRT_Id ($r = 0.478, p = 0.001$) and SRT_Tar ($r = 0.431, p = 0.004$). The analysis revealed a weak positive correlation between SRT_Id and LoE ($r = 0.364, p < 0.05$) and a weak negative correlation between SRT_Id and AoO ($r = -0.348, p < 0.05$), whereas the latter two correlations were not significant in case of SRT_Tar. Two multiple linear regression models were built for predicting performance of the BiTDs on SRT_Id and



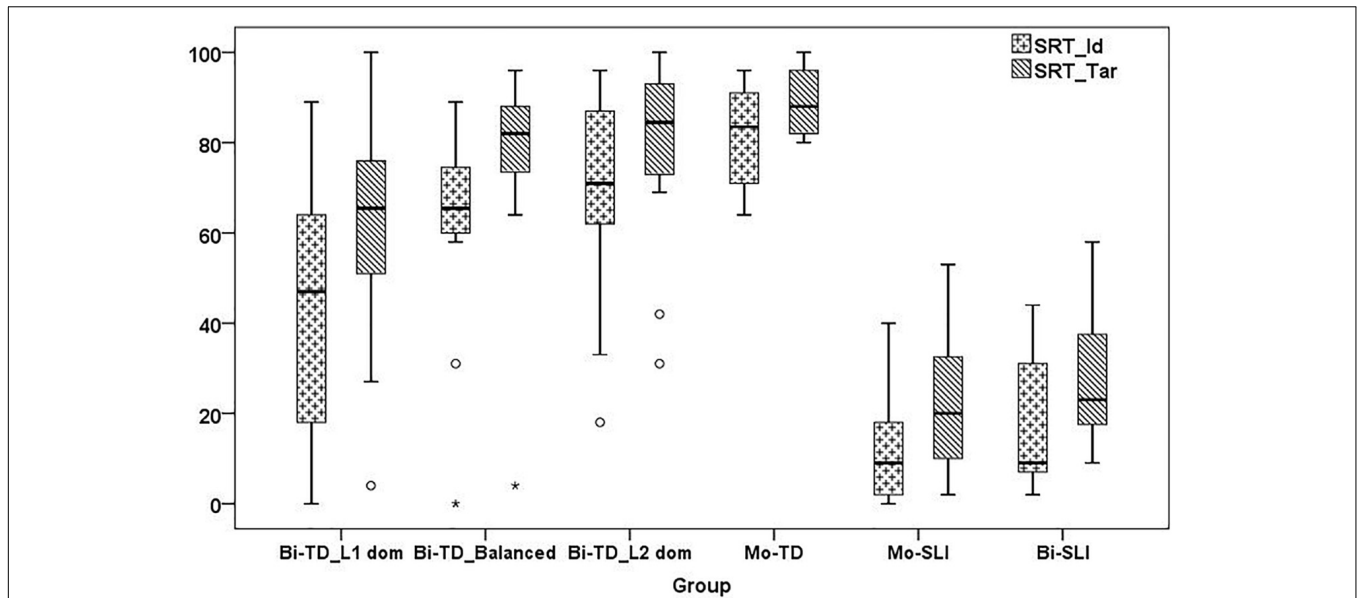


FIGURE 3 | SRT: % identical repetition (SRT_Id) and correct production of target structure (SRT_Tar).

TABLE 5 | The effect of (clinical) group membership on LITMUS-NWRT and SRT (Kruskal-Wallis tests) and pairwise comparisons (Mann-Whitney U tests).

	NWRT_global	NWRT_LI	NWRT_LD	SRT_Id	SRT_Tar
	$\chi^2(5, N = 77) = 39.600, p < 0.001$	$\chi^2(5, N = 77) = 35.044, p < 0.001$	$\chi^2(5, N = 77) = 38.714, p < 0.001$	$\chi^2(5, N = 77) = 43.365, p < 0.001$	$\chi^2(5, N = 77) = 43.086, p < 0.001$
MoTD vs. MoSLI	$U = 0.500, p < 0.001, r = 0.838$	$U = 1.500, p < 0.001, r = 0.824$	$U = 0.000, p < 0.001, r = 0.847$	$U = 0.000, p < 0.001, r = 0.847$	$U = 0.500, p < 0.001, r = 0.846$
BiTD_L1 dom. vs. MoSLI	$U = 2.000, p < 0.001, r = 0.804$	$U = 6.500, p < 0.001, r = 0.775$	$U = 3.000, p < 0.001, r = 0.797$	$U = 33.500, p < 0.001, r = 0.592$	$U = 24.500, p < 0.001, r = 0.652$
BiTD_Balanced vs. MoSLI	$U = 7.000, p < 0.001, r = 0.775$	$U = 8.5000, p < 0.001, r = 0.763$	$U = 10.000, p < 0.001, r = 0.736$	$U = 11.500, p < 0.001, r = 0.717$	$U = 10.000, p < 0.001, r = 0.735$
BiTD_L2 dom. vs. MoSLI	$U = 3.000, p < 0.001, r = 0.845$	$U = 4.500, p < 0.001, r = 0.830$	$U = 3.500, p < 0.001, r = 0.841$	$U = 5.000, p < 0.001, r = 0.824$	$U = 5.000, p < 0.001, r = 0.822$
BiTD_L1_dom. vs. BiSLI	$U = 6.000, p < 0.001, r = 0.765$	$U = 25.000, p < 0.001, r = 0.643$	$U = 8.000, p < 0.001, r = 0.752$	$U = 44.000, p < 0.05, r = 0.512$	$U = 29.500, p < 0.001, r = 0.608$
BiTD_balanced. vs. BiSLI	$U = 2.000, p < 0.001, r = 0.802$	$U = 10.500, p < 0.001, r = 0.688$	$U = 2.000, p < 0.001, r = 0.804$	$U = 3.000, p < 0.001, r = 0.798$	$U = 0.000, p < 0.001, r = 0.829$
BiTD_L2_dom. vs. BiSLI	$U = 9.500, p < 0.05, r = 0.648$	$U = 12.000, p < 0.05, r = 0.616$	$U = 7.000, p < 0.001, r = 0.686$	$U = 4.000, p < 0.001, r = 0.726$	$U = 4.500, p < 0.001, r = 0.718$
BiTD_L1_dom. vs. MoTD	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	$U = 16.500, p < 0.001, r = 0.672$	$U = 21.000, p < 0.001, r = 0.638$
BiTD_L1_dom. vs. BiTD_L2_dom.	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	$U = 44.500, p < 0.05, r = 0.559$	$U = 53.000, p < 0.05, r = 0.501$

SRT_Tar. The following variables were entered into the model as independent variables: AoO, LoE, LDI, and SES. The results show that performance on SRT_Id in the BiTD group is predicted by LDI ($\beta = 3.724, T = 2.922, p = 0.001$), followed by LoE ($\beta = 3.846, T = 2.287, p = 0.01$), and SES ($\beta = 3.424, T = 2.829, p = 0.001$). However, for SRT_Tar only LDI and SES had significant effects in the full model: LDI ($\beta = 4.480, T = 3.360, p = 0.001$), SES ($\beta = 2.914, T = 2.301, p = 0.01$). The independent variables did not show multi-collinearity in the models ($VIF < 3$ for all independent variables).

Comparison of global performance of L1-dominant BiTDs to their monolingual, balanced, and L2-dominant peers as well as results of regression analyses show that language dominance was the first predictor to explain performance of the BiTDs on both SRT_Id and SRT_Tar, and point to the possibility that language dominance could compromise the diagnostic accuracy of the SRT if administered to bilinguals in their non-dominant language, here German. In order to examine this, we ran the k-medoid PAM-clustering to group the children into SLI vs. TD based on their performance scores on SRT and NWRT, determined the cut-off points between the clusters for each of the repetition measures and calculated the diagnostic accuracy for different combinations of sub-measures of the two. Next, regression analyses using Firth's correction were carried out to examine whether language dominance contributed to results of PAM-clustering, i.e., assigning the children to the clinical vs. non-clinical cluster based on performance scores on the LITMUS repetition tasks. We examined this for SRT and NWRT separately as well as combined. LDI and Positive_Early_Development

were entered as predictors for PAM cluster membership into all regression models in addition to two further background variables (age, AoO, LoE, SES).

Before applying the PAM clustering to our bilinguals, we first tested it on our monolingual data set. The following variables were entered in the cluster analysis simultaneously: NWRT_global, NWRT_LI, NWRT_LD, SRT_Id and SRT_Tar. The Hopkins statistic yielded a value of around 0.23 indicating clusterable non-random data, and the Gap Statistic revealed that the optimal cluster solution is 2. The clustering procedure resulted in a clear separation into two homogenous groups with two cluster medoids. The cut-off points (see section "Data Analysis") separating the monolingual clinical cluster from the non-clinical one in our data sample were as follows: SRT_Id: 40%, SRT_Tar: 53.3% and NWRT_global: 45.45%, NWRT_LI: 60%, NWRT_LD: 47.22%. Figure 4 gives a visual representation of the k-medoid PAM-cluster analysis on monolingual data using the two-cluster solution. Cases belonging to the cluster on the right are identified as TD cases, while those in the cluster on the left as SLI cases. To facilitate computing sensitivity and specificity of the task, case numbers were combined with the clinical status as assigned by our classification procedure based on standardized assessment. As can be seen in Figure 4, all of the monolingual children assigned to the MoSLI group based on standardized test procedures belong to the clinical cluster, yielding a sensitivity of 100%, whereas all of the monolingual subjects classified as MoTD based on standardized test procedures belonged to the non-clinical cluster, which yields a specificity of 100%. We also ran the clustering procedure

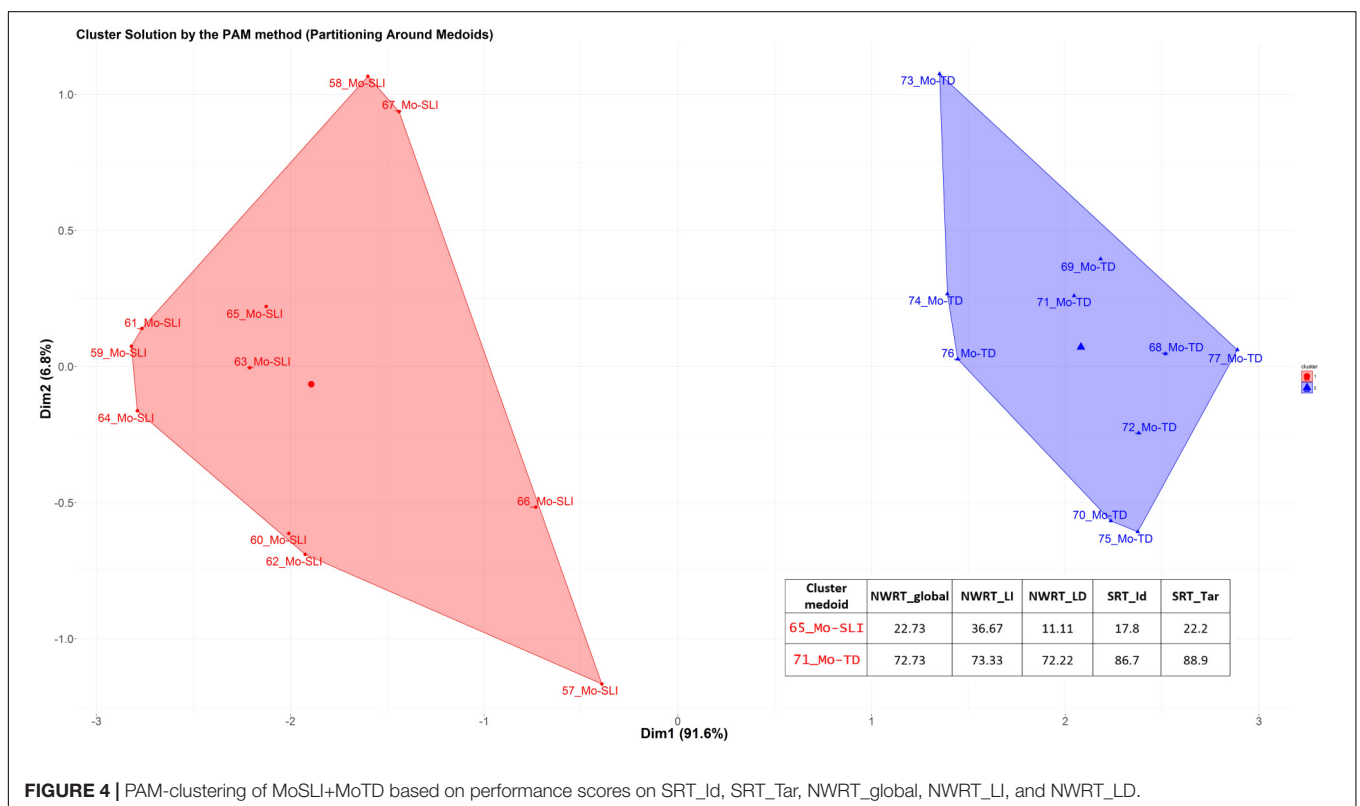


FIGURE 4 | PAM-clustering of MoSLI+MoTD based on performance scores on SRT_Id, SRT_Tar, NWRT_global, NWRT_LI, and NWRT_LD.

on SRT and NWRT separately, i.e., (SRT_Id+SRT_Tar) and (NWRT_global+NWRT_LI+NWRT_LD), respectively and obtained similar results. In a next step, we used regression analysis entering age²³ as a single variable to check whether chronological age could explain cluster membership. Results of the latter analysis indicate that there is no association between age and the cluster variable (Firth: $\beta = -0.03167$, $Z = -0.646$, $p = 0.519$).

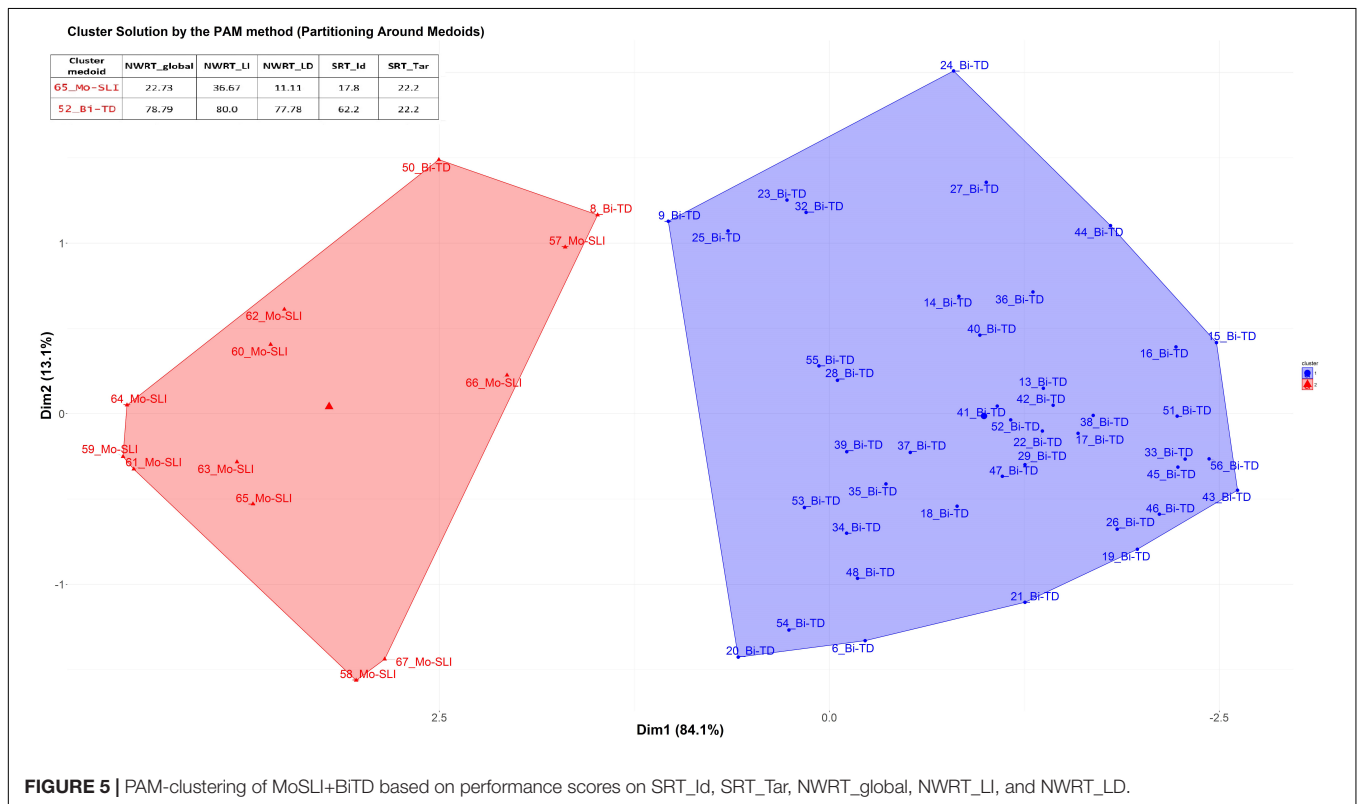
In order to check for overlap between BiTD and MoSLI, we applied the PAM-analysis to the MoSLI and BiTD children collapsed together using performance scores on both SRT and NWRT. The data yielded an H-value of around 0.18, which indicates clusterable non-random data with 2 as the optimal number of clusters. Before entering all five variables into the clustering procedure, we first carried out the clustering procedure based on performance on SRT_Id+SRT_Tar. Ctree models showed that SRT_Tar but not SRT_Id predicted cluster membership with a threshold of 53.3% separating the two clusters. All MoSLI children scored below cut-off and were thus assigned to the clinical cluster by the PAM algorithm, i.e., sensitivity = 100% with an LR+ = 6.29, whereas 37/44 BiTD children performed above threshold (specificity = 84.1%, LR- = 0.00). Age as a single variable in the regression model did not prove to be a predictor for cluster membership (Firth: $\beta = -0.02849$, $Z = -1.238$, $p = 0.216$). In the next step, we ran the PAM-analysis on NWRT_global, NWRT_LI and NWRT_LD. The clustering resulted in two clusters separated

by a cut-off of 33.33% on NWRT_LD, which also is the primary predictor of clustering membership. 10/11 MoSLI children were assigned to the clinical cluster, yielding a sensitivity rate of 91% and 43/44 BiTD children performed above cut-off and were assigned to the non-clinical cluster giving a specificity of 98% with an LR+ = 39.56 and LR- = 0.092. Again, age was not a significant predictor for the cluster variable (Firth: $\beta = -0.03618$, $Z = -1.293$, $p = 0.196$).

Finally, both LITMUS-tasks were included in the PAM-analysis using the measures NWRT_global, NWRT_LI, NWRT_LD, SRT_Id and SRT_Tar. After entering all SRT and NWRT measures at once into the clustering procedure, 5 of the 7 BiTDs, who were assigned to the clinical cluster based on scores on SRT alone, changed membership from the clinical to the non-clinical cluster. A combination of both SRT and NWRT measures yielded 100% sensitivity (all MoSLIs belong to the clinical cluster) and 95% specificity (42/44 BiTDs belong to the non-clinical cluster) with an LR+ of 20 and an LR- of 0.00. An illustration of the result of the PAM cluster analysis on the MoSLI and BiTD data is given in Figure 5. Age at testing as a single variable did not play a significant role in predicting PAM cluster membership (Firth: $\beta = -0.05145$, $Z = -1.824$, $p = 0.0681$).

A visual representation, a ctree, of the hierarchical structure of the most relevant linguistic variables for predicting PAM cluster membership illustrated in Figure 5 is provided in Figure 6. Within a ctree, only those variables serving as relevant to explaining the clustering results appear in the

²³SES information is only available for the bilingual participants.

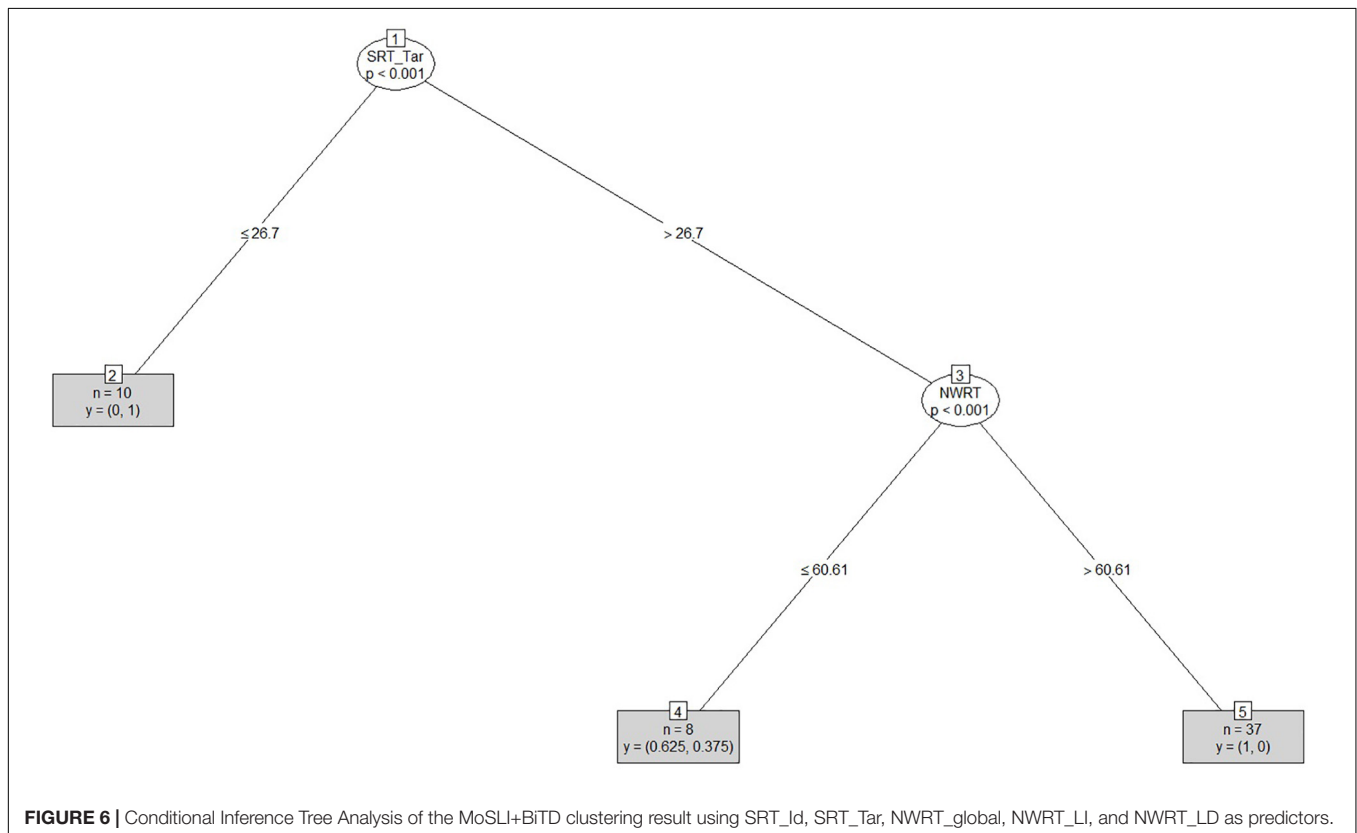


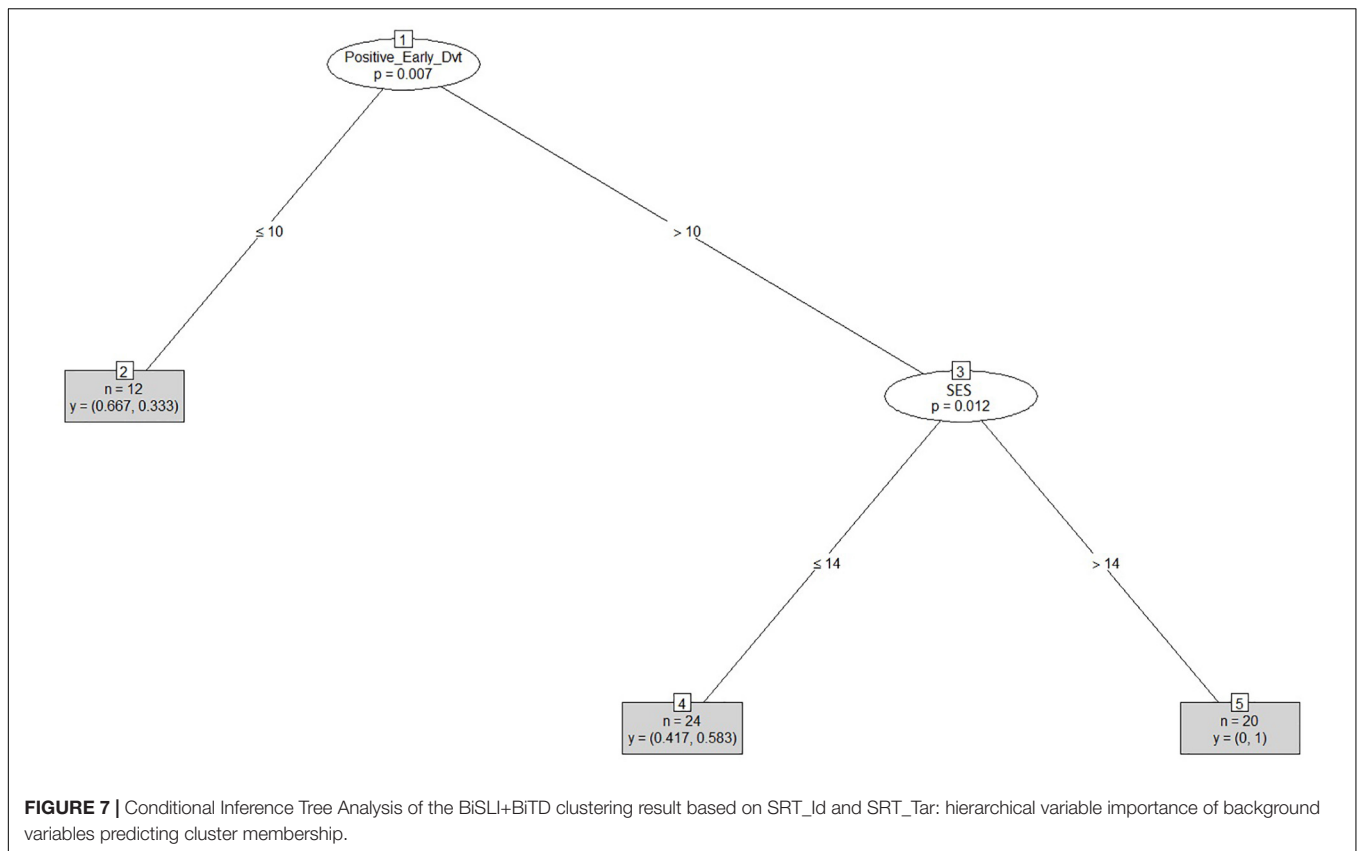
graph, where each relevant variable is represented by an oval circle and classification rules are represented by thresholds. Classification of cases starts at the top node (root). The second most important variable is one level below the top node. Classification then proceeds by moving down the branch until we arrive at a terminal node representing classification accuracy according to PAM clustering²⁴, where classification accuracy is represented in squares (y). The two numbers next to “y” show the proportion of cases successfully classified and misclassified as SLI. The number of cases on that route is represented by “n.” Each classification route can be expressed in the form of if-then conditions with cut-offs. As can be seen in **Figure 6**, when all five measures are included in the clustering procedure, both SRT_Tar and NWRT_global are identified as significant contributors toward predicting PAM cluster membership. Classification of cases start at the top node occupied by SRT_Tar followed by the second most important variable “NWRT_global,” which is one level below the top node. Based on the hierarchical variable structure depicted in the ctree below, it becomes visible that 10 children whose scores on SRT_Tar were $\leq 26\%$ were assigned to the clinical cluster. For subjects performing above 26% correct on SRT_Tar, performance on NWRT_global was taken into

account giving rise to two roots: (a) if subject performs $>26\%$ on SRT_Tar and $>60.61\%$ on NWRT_global then assign to non-clinical cluster (TD), (b) if subject performs $>26\%$ on SRT_Tar but $\leq 60.61\%$ on NWRT_global then assign to clinical cluster (SLI).

Turning now to results of bilingual children, we performed the PAM-analysis on all BiSLI and BiTD groups collapsed together based on the performance scores in the SRT and NWRT. The *Hopkins statistic* indicated regularly spaced data that are neither clustered nor random (H-value of around 0.2) and the Gap statistic suggested the two-cluster solution. Results of the PAM clustering based on performance of BiTDs and BiSLIs on SRT_Id and SRT_Tar were similar to those we obtained for BiTDs and MoSLIs (see **Figure 5**). 11/12 BiSLIs were assigned to the clinical cluster yielding a sensitivity of 91.7%, whereas 37/44 BiTDs were assigned to the non-clinical cluster giving a specificity of 84.1%, $LR+ = 5.76$, $LR- = 0.10$. The thresholds separating the bilingual clinical cluster from the non-clinical one were 33.3% for SRT_Id and 53.3% for SRT_Tar, whereby SRT_Tar was the main predictor for the clustering result (with the same cut-off of 53.3%). Regression analysis as well as ctree analysis showed that Positive_Early_Development (Firth: $\beta = 1.0636$, $Z = 2.614$, $p = 0.001$) followed by SES ($\beta = 0.7843$, $Z = 2.033$, $p = 0.01$) were significant predictors for cluster membership. Variables related to bilingualism, i.e., AoO, LoE and LDI, did not explain cluster membership. An illustration of hierarchical structure of variable importance with classification thresholds is depicted in **Figure 7**.

²⁴The classification accuracy given in squares within the ctree refers only to that of PAM clustering and does not represent diagnostic accuracy of cluster membership as measured by comparing cases identified as SLI or TD by the PAM to the clinical status assigned based on our classification procedure outlined in section “Participants.”





We ran the same clustering procedure on the bilinguals' performance in NWRT using the variables NWRT_global, NWRT_LI and NWRT_LD. All BiSLI children were assigned to the clinical cluster yielding a 100% sensitivity; however, 9 BiTD children were assigned to the clinical cluster, i.e., only 35/44 BiTDs were assigned to the non-clinical cluster (specificity = 80%), LR+ = 5, LR- = 0.00. Ctree analysis showed that NWRT_global was the main variable predicting cluster membership with a cut-off 66.7%. Next, we ran Firth's biased regression analysis on the clustering results for NWRT entering age, Positive_Early_Development, SES and LDI as fixed factors. Results showed that neither language dominance nor SES explained cluster membership based on NWRT_global. As expected, Positive_Early_Development was the main variable explaining the clustering result (Firth: $\beta = 0.38996$, $Z = 2.626$, $p = 0.001$). The other significant predictor for NWRT_global was chronological age (Firth: $\beta = 0.05931$, $Z = 2.150$, $p = 0.01$).

Since NWRT_global is a composite score obtained by adding up scores of both of the language independent (NWRT_LI) and language dependent parts (NWRT_LD), we wanted to verify whether both of them were affected by the age factor. To achieve this, we ran the PAM-analysis on each of them separately. The results show that if clustering is solely based on performance on NWRT_LI upon a threshold of 73.3% (as established by ctree analysis), 10 BiTD children would be over-identified as having SLI, yielding a specificity of only 77%,

LR+ = 4.385, LR- = 0.00. Both Positive_Early_Development (Firth: $\beta = 0.38996$, $Z = 2.626$, $p = 0.01$) and age (Firth: $\beta = 0.05591$, $Z = 2.266$, $p = 0.01$) were significant predictors for the clustering results (variables entered in the regression model: Positive_Early_Development, SES, LDI and age). Ctree analysis showed that the age threshold separating the two clusters based on NWRT_LI was 87 months (7;3 years). On the other hand, if the bilingual children in our data set are clustered based on performance in NWRT_LD alone, the diagnostic accuracy drastically improves: upon a 50% cut-off score, only 2/44 BiTD children are assigned to the clinical cluster, while all BiSLI children are classified as SLI, which yields 95% specificity and 100% sensitivity (LR+ = 20, LR- = 0.00). Positive_Early_Development was singled out as a predictor explaining cluster membership based on NWRT_LD (Firth: $\beta = 0.30611$, $Z = 2.946$, $p = 0.001$), i.e., the variables age, LDI and SES did not explain cluster membership. In the following step, we included all SRT and NWRT measures (SRT_Id, SRT_Tar, NWRT_global, NWRT_LI and NWRT_LD) in the clustering procedure. As can be seen in **Figure 8**, combing SRT with NWRT enhances the diagnostic accuracy: all of the BiSLI children (12/12) were assigned to the clinical cluster (100% sensitivity), while 39/45 BiTD children were assigned to the non-clinical cluster (87% specificity, LR+ = 7.692, LR- = 0.00).

Figure 9 shows that both NWRT_global and SRT_Tar were significant contributors toward predicting PAM

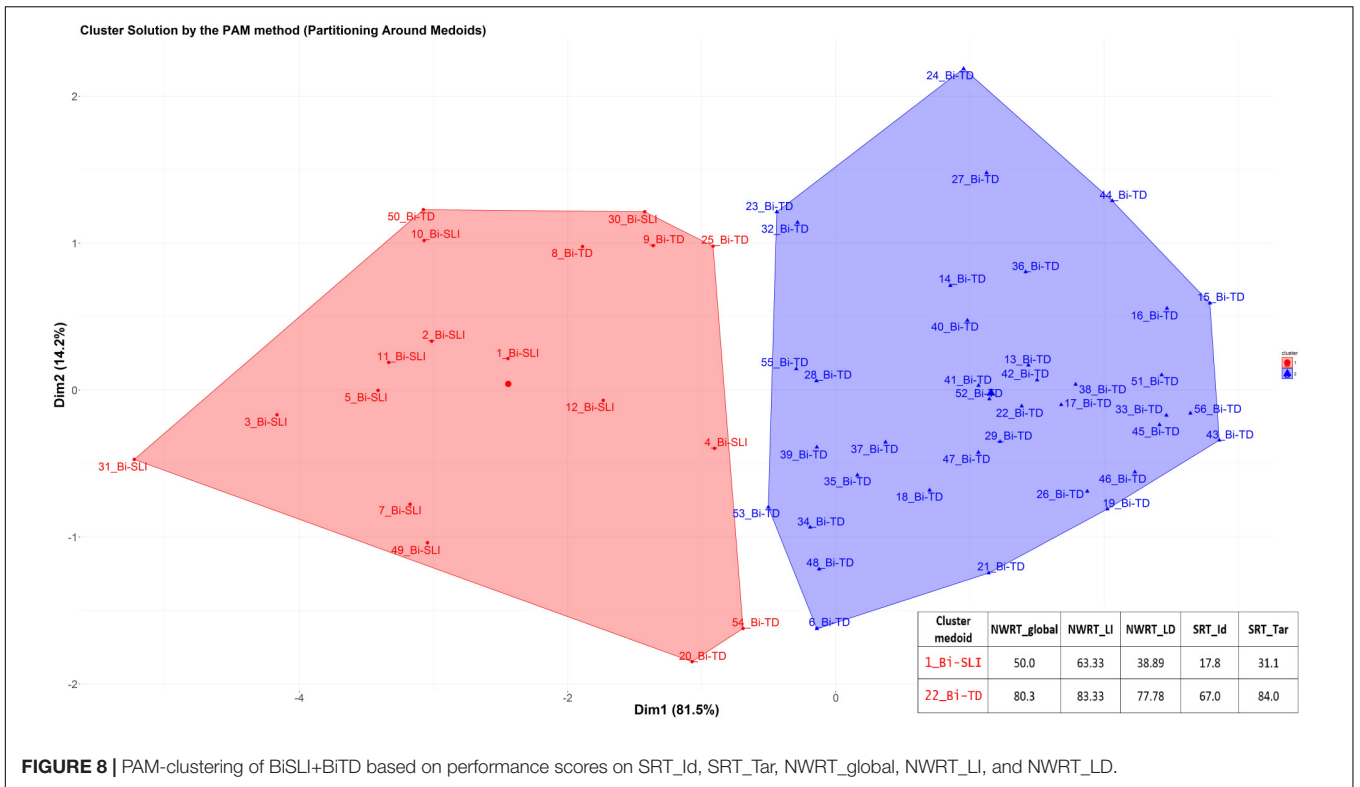


FIGURE 8 | PAM-clustering of BiSLI+BiTD based on performance scores on SRT_Id, SRT_Tar, NWRT_global, NWRT_LI, and NWRT_LD.

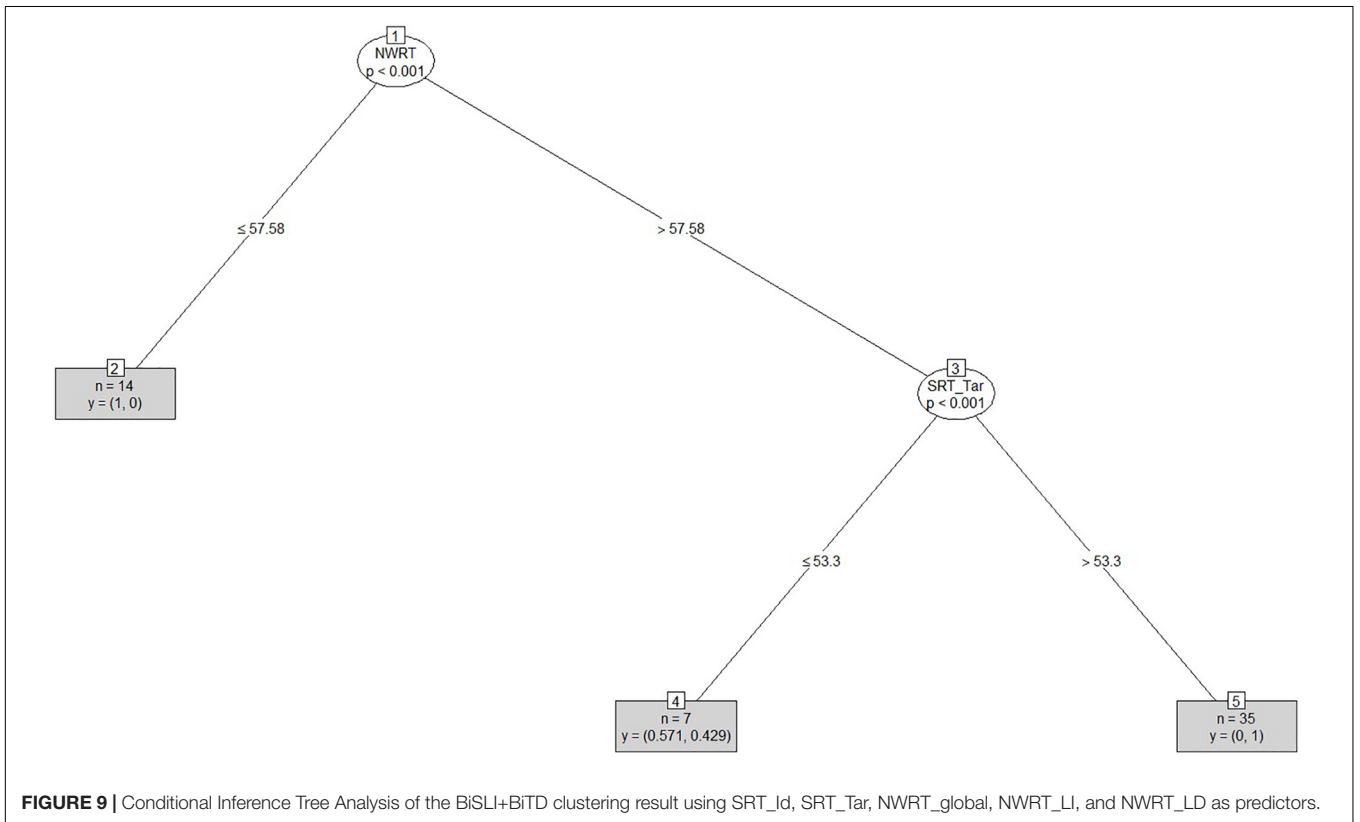


FIGURE 9 | Conditional Inference Tree Analysis of the BiSLI+BiTD clustering result using SRT_Id, SRT_Tar, NWRT_global, NWRT_LI, and NWRT_LD as predictors.

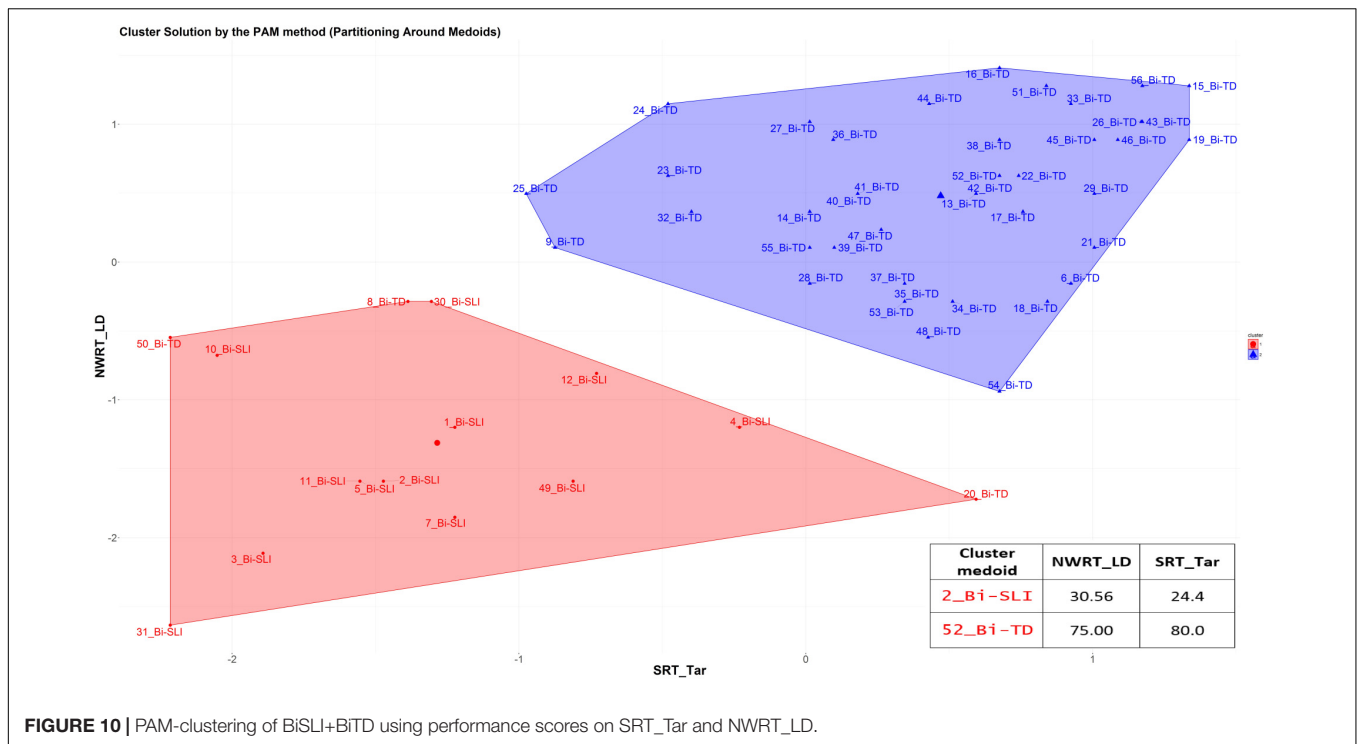


FIGURE 10 | PAM-clustering of BiSLI+BiTD using performance scores on SRT_Tar and NWRT_LD.

cluster membership when all 5 variables are included in the clustering procedure. Classification of cases started at top node “NWRT_global” followed by the second relevant variable “SRT_Tar,” which is one level below the top node. According to the hierarchical variable structure illustrated in **Figure 9**, 14 children whose scores on NWRT_global were $\leq 57.58\%$ were classified as SLI. In case of children with scores above 57.58% on NWRT_global, performance on SRT_Tar was taken into consideration leading to two roots: (a) if subject performs $> 57.58\%$ on NWRT_global and $> 53.3\%$ SRT_Tar, then assign subject to non-clinical cluster (TD), (b) if subject performs $> 57.58\%$ on NWRT_global but $\leq 53.3\%$ on SRT_Tar then assign to clinical cluster (SLI). Regression analysis using the previous four variables revealed that only Positive_Early_Development was a significant predictor for the clustering outcome (Firth: $\beta = 0.39394$, $Z = 2.907$, $p = 0.001$).

To address research question (iii), we ran PAM clustering on bilingual data using scores of NWRT_LD and SRT_Tar in order to examine whether a combination thereof yielded the best diagnostic accuracy rates. Indeed, only 2/44 BiTD children were over-identified as having SLI (95% specificity) and all of the 12 BiSLI children were assigned to the clinical cluster (100% sensitivity) with an LR+ of 20 and an LR- of 0). The cut-off scores were 52.78% for NWRT_LD and 53.3% for SRT_Tar, with NWRT LD being the primary predictor for clustering results followed by SRT_Tar. Only Positive_Early_Development was a significant predictor of cluster membership (Firth: $\beta = 0.39394$, $Z = 2.907$, $p = 0.001$). The clustering results are depicted in **Figure 10**.

DISCUSSION AND CONCLUSION

The purpose of this study was to evaluate the robustness of two LITMUS tools, German LITMUS-SRT and NWRT, against the influence of language dominance on their diagnostic accuracy for SLI in bilingual children. Since both tasks were designed to minimize bias against bilingual populations while being indicative of the presence of LI, we wanted to specifically verify whether the tasks were only sensitive to risk factors for SLI or whether background variables related to bilingualism, particularly, the degree of language dominance (as measured by relative amount of use and exposure to L1/L2) could influence the performance of BiTDs to an extent that would compromise their diagnostic accuracy. The second aim of the study was to investigate whether combining LITMUS-SRT (especially when scored by correct target structure) with NWRT yielded better diagnostic accuracy than single measures and helped avoid cases of misdiagnosis. Following our own research (e.g., Abed Ibrahim and Hamann, 2017; Hamann and Abed Ibrahim, 2017; Grimm and Hübner, in press), we particularly wanted to check whether a combination of German SRT_Tar and the language dependent part of the NWRT yielded higher diagnostic accuracy for identifying SLI in bilingual children than other combinations of measures. The former was found to be a fairer method than identical repetition for scoring SRT as it compensates for typical L2-errors such as lexical substitutions, while the latter was shown to maximize the performance gap between SLI and TD not only in monolinguals but also in bilinguals given its higher level of structural complexity.

In order to examine this, we first compared global performance of L1-dominant BiTDs to that of MoTDs, balanced

and L2-dominant-BiTDs as well as to MoSLIs and BiSLIs. Results showed that although all three BiTD groups (regardless of their dominance) significantly outperformed MoSLIs and BiSLIs on all SRT and NWRT measures, L1-dominant-BiTDs were significantly outperformed by MoTDs and L2-dominant-BiTDs on both SRT_Id and SRT_Tar with large effect sizes (see **Figures 2, 3**). This echoes the findings of Meir (2018), who reported similar results for performance of Russian-Hebrew bilinguals on LITMUS-SRTs in their weaker heritage or societal language. Our results further showed that the performance gap between monolingual and bilingual SLI and TD groups was larger for NWRT_LD as opposed to the structurally less complex language independent part of the NWRT and the composite score of the two parts “NWRT_global” (see **Figure 2**). This is in line with previous work showing that the complexity factors involved in the NWRT_LD part (i.e., presence of trilateral onset clusters and /sC/ clusters violating the Sonority Sequencing Principle) is particularly challenging for language impaired children regardless of lingual status (Ferré et al., 2015; dos Santos and Ferré, 2018; Grimm and Hübner, in press).

Since language dominance was used as a categorical variable to classify BiTDs in our between-group comparisons, we had to entertain the possibility that the assumed dominance effect for L1-dominant children might have been caused by confounding variables such as age of onset of exposure to L2 (AoO), length of exposure to L2 (LoE) and SES. As for SRT_Id and SRT_Tar, we found moderate correlations between performance and language dominance as well as SES, in addition to weak correlations for SRT_Id with LoE and AoO. Regression analysis showed that language dominance was the key predictor explaining variance in the performance of the BiTDs on SRT_Id and SRT_Tar followed by SES. That AoO and LoE did not predict performance of BiTDs on the SRT was an expected outcome since the vast majority of the participants in our bilingual sample were either simultaneous or early successive and were exposed to German for at least 24 months at the time of assessment (see also Armon-Lotem, 2011 for similar results on L2-Hebrew-SRT).

The finding that language dominance influenced the performance of BiTDs on both measures of the LITMUS-SRT questioned its applicability for the identification of SLI in L1-dominant children when administered in their weaker language (German). To answer this, we used a prominent unsupervised machine learning technique, the Partitioning Around Medoids (PAM) for establishing an automatic classification of the monolingual and bilingual children in our data set as TD vs. SLI directly from their performance scores on SRT and NWRT without using information about their clinical status. Subsequently, we compared the participants’ clinical group membership revealed by PAM-clustering to their clinical status based on standardized assessment in L1/L2, and calculated sensitivity and specificity (diagnostic accuracy) levels of the tasks in isolation and combined. We also explored which combinations of the measures obtained from SRT_Id, SRT_Tar, NWRT_global, NWRT_LI, and NWRT_LD yielded the highest diagnostic accuracy. Finally, we conducted regression analysis to investigate whether background variables other than risk factors for SLI, in particular language dominance (LDI),

explained PAM-cluster membership as TD or SLI based on performance scores on SRT and/or NWRT. Since the index of Positive_Early_Development was shown to be a strong predictor for SLI in bilinguals (Boerma and Blom, 2017; Tuller et al., 2018), our premise was that if PAM-cluster membership can only be predicted by this index and not by language dominance or other background variables known to influence performance on repetition tasks (age, AoO, LoE, SES), then clustering of cases cuts across the SLI/TD dimension confirming that the LITMUS-SRT and NWRT are only sensitive to the presence of SLI and are not biased against bilingual children, who are non-dominant in the societal language.

In Hamann and Abed Ibrahim (2017), unsupervised (clustering) machine learning algorithms were only applied to the bilingual data, while Receiver Operating Characteristic curve (ROC) analysis was used to calculate sensitivity and specificity levels for the monolingual data. Given that ROC analysis uses “clinical status” (as assigned by standardized test procedures) as a dependent variable for predicting the sensitivity and specificity of a test, we wanted to verify this finding for the monolinguals using a method independent of “clinical status.” PAM-clustering solely based on scores in SRT_Id, SRT_Tar, NWRT_global, NWRT_LI, and NWRT_LD yielded even higher diagnostic accuracy than that in Hamann and Abed Ibrahim (2017). The fact that all of the subjects identified as MoSLI by standardized assessment belonged to the lower performing cluster, while all of the MoTDs belonged to the higher performing cluster (100% sensitivity and 100% specificity) provides additional evidence that these linguistically motivated tasks are very sensitive to the presence of LI in monolinguals and tap the core morphosyntactic and phonological deficits in SLI. The source of the improved diagnostic accuracy as compared to results based on ROC-analysis in Hamann and Abed Ibrahim (2017) is most likely the simultaneous inclusion of both tasks into the clustering procedure and the lower cut-off points obtained by applying ctrees to the PAM clustering. This is reminiscent of Armon-Lotem and Meir’s (2016) study, which reported an increase in diagnostic accuracy when LITMUS-SRT is supplemented by NWRT for Hebrew and Russian monolinguals. A further important result was that chronological age could not predict cluster membership for the age range in our monolingual data set.

After establishing that both LITMUS-SRT and NWRT were sensitive to SLI in monolinguals, we proceeded to address the frequently reported overlap between MoSLI and BiTD children (e.g., Håkansson and Nettelbladt, 1996; Armon-Lotem, 2010; Paradis, 2010; Hamann, 2012). PAM-clustering conducted on SRT scores entering both measures SRT_Id and SRT_Tar yielded good overall diagnostic accuracy (100% sensitivity and 84.1%²⁵ specificity) with SRT_Tar being the leading variable for predicting cluster membership since it led to a better separation between

²⁵It is worthwhile mentioning that three of the 7 BiTD children assigned to the clinical cluster started acquiring L2 German after the age of five and spent their first years in an exclusive L1-environment. The latter children demonstrated high rates of determiner errors, especially case and gender errors which could affect the realization of target structure of a sizable proportion of the test items of the LITMUS-SRT (see also Abed Ibrahim et al., 2018).

the BiTD and MoSLI clusters. Several studies found this scoring method better suited for assessing morphosyntactic abilities in bilingual children, since it only focuses on the mastery of syntactic structure and does not penalize bilingual children for frequent L2-errors such as lexical substitutions (Armon-Lotem and Meir, 2016; Hamann and Abed Ibrahim, 2017; Hamann et al., 2017; Abed Ibrahim et al., 2018; Meir, 2018).

Next, we checked whether the overlap problem between MoSLI and BiTD could be overcome by using SRT in combination with NWRT. Indeed, including NWRT scores into the clustering procedure resulted in much better diagnostic accuracy with almost no overlap between MoSLI and BiTD (100% sensitivity and 95% specificity). As also reported in Armon-Lotem and Meir (2016), de Almeida et al. (2017), Boerma and Blom (2017), and Hamann and Abed Ibrahim (2017), the latter finding corroborates that a combination of LITMUS instruments assessing different areas of language ability helps to avoid cases of misdiagnosis. Among the five measures SRT_Id, SRT_Tar, NWRT_global, NWRT_LI and NWRT_LD, both SRT_Tar and NWRT_global were main predictors for clustering results with SRT_Tar being the more important contributor (see **Figure 6**). We further demonstrated that chronological age did not predict cluster membership here either.

As to the diagnosis of bilinguals, PAM clustering based on scores in SRT_Id and SRT_Tar resulted in good overall accuracy rates (91.7% sensitivity and 84.1% specificity). Interestingly, the same 7 BiTDs previously assigned to the clinical cluster upon comparison with MoSLIs were classified as SLI by the PAM as well showing that changing the reference group had no influence on the individual classification of the BiTDs. Again, SRT_Tar, which compensates for L2-errors, was the primary contributor toward the clustering results with a cut-off 53.3%, which is very close to the threshold obtained by *k*-means clustering in Hamann and Abed Ibrahim (2017). Of the five background variables considered for regression analysis, just two variables unrelated to bilingualism emerged as significant predictors for clustering membership: Positive_Early_Development followed by SES. The influence of language dominance, which was a significant predictor explaining the variance in the performance of the BiTDs in SRT_Id and SRT_Tar, was outweighed by the presence of risk factors for SLI and was rendered insignificant once the BiSLIs became part of the equation. This is consistent with the findings of Tuller et al. (2018), who found for the German children that Positive_Early_Development was the leading predictor for performance in SRT (followed by SES) over variables related to bilingualism.

The clustering solution based on NWRT_global, NWRT_LI and NWRT_LD scores yielded only fair diagnostic accuracy rates due to reduced specificity (specificity = 80%). NWRT_global emerged as the main predictor for clustering results. Regression analysis revealed that not only Positive_Early_Development (most important predictor) but also chronological age were significant predictors for clustering results based on performance scores in NWRT_global. Given that NWRT_global is a composite score computed by adding up performance scores in NWRT_LI, and NWRT_LD, and since Grimm and Hübner (in press) reported an overlap between MoSLI and BiTD on NWRT_LI

and better discriminatory power for NWRT_LD in children aged 8;0 to 10;0 years, we ran cluster analyses on both subparts of the NWRT separately to check for age effects. The analysis revealed that in addition to Positive_Early_Development, cluster-membership based on NWRT_LI was predicted by chronological age with a threshold of 7;3 years, whereas cluster-membership based on NWRT_LD was not predicted by age and was only sensitive to risk factors for SLI. On the other hand, neither bilingualism related factors nor SES predicted cluster membership derived by performance scores on NWRT_global or subtests thereof. The latter result echoes what has been found for this type of NWRT in de Almeida et al. (2017) as well as in Tuller et al. (2018).

We have also shown that including all five SRT and NWRT measures in the clustering procedure enhances diagnostic accuracy for SLI in bilingual children, where NWRT_global and SRT_Tar were the main contributors explaining the results of the cluster solution. Interestingly, once SRT is combined with NWRT, only Positive_Early_Development emerges as a significant predictor for clustering results and SES does not play a role anymore, which is in line with the findings of Chiat and Poliženská (2016).

Given that clustering by scores on NWRT_LI appeared to be influenced by age, while NWRT_LD was only sensitive to risk factors of SLI (Positive_Early_Development) and since the SRT_Tar was the chief contributor toward clustering results when both SRT_Id and SRT_Tar were included in any clustering procedure on bilingual performance, we expected a combination of SRT_Tar and NWRT_LD to yield better diagnostic accuracy rates than other combinations of measures. Indeed, clustering based on performance scores on SRT_Tar and NWRT_LD yielded the highest diagnostic accuracy, where only Positive_Early_Development predicted clustering results. The crucial contribution of the structurally more complex NWRT_LD toward diagnostic accuracy is consistent with the robust effects of phonological complexity found in the respective studies (e.g., Gallon et al., 2007; Ferré et al., 2012), with clinical implications that phonological complexity can be used as a reliable indicator for SLI in both monolingual and bilingual children (see Grimm and Hübner, in press). Our results concerning the NWRT_LD part might seem at odds with results of other studies showing better diagnostic accuracy for *Crosslinguistic*-NWRTs over *Language-Specific*-NWRTs in bilingual populations, e.g., Boerma et al. (2015), Armon-Lotem and Meir (2016), and Boerma and Blom (2017). This can clearly be ascribed to differences in the construction of the tasks, which, as described in the section “The German LITMUS Nonword Repetition Task”, tap different aspects vulnerable in SLI (i.e., phonological working memory vs. phonological complexity), and differ considerably from each other, especially in their language dependent parts. Another possible reason for the poor diagnostic accuracy reported for the *Language-Specific*-NWRTs in the latter three studies might be relatively young age of their participants (5;0–6;0) compared to the age range in our sample (5;6–9;0), which covers the last year of preschool and the first 2–3 primary school years. A study by Rispens and Baker (2012) demonstrated that

both lexical knowledge and discrimination ability significantly influenced performance on NWRT in 5-year-old MoTDs, while this kind of relation could not be attested for 8-year olds.

In line with our previous research, the results presented here and the fact that they emerge from unsupervised PAM-clustering clearly indicate that the German LITMUS- SRT and NWRT are promising tools for the identification of LI in bilingual populations with diverse dominance profiles. We replicated the finding that SRT_Tar is better suited than SRT_Id for the assessment of language abilities of bilingual children with German as L2 on a slightly larger group of children with a statistical method better suited for our data set. Even though dominance influences the performance of BiTDs, especially in the SRT, we demonstrated that the diagnostic accuracy of these tools is not compromised by language dominance: while risk factors for SLI were significant predictors for clinical status in all models, language dominance did not contribute at all to explaining results of any of the clustering procedures. Moreover, our results confirmed that using a combination of tasks, each emphasizing a different aspect of language ability, enhances diagnostic accuracy and helps avoid cases of misdiagnosis. As a last promising result, we showed that using SRT_Tar in conjunction with NWRT_LD renders the best diagnostic accuracy so far obtained in studies on similarly constructed tasks, where the combination of measures is only sensitive to risk factors for SLI, but not to language dominance nor to SES, which is not achieved by many tasks. We therefore feel confident in pursuing these investigations in order to be able to provide useful and easy to administer L2-tools for clinical use in bilingual contexts. Finally, it should be noted that vast majority of the bilingual children in our sample were either simultaneous or early successive bilinguals, who had at least 2 years of exposure to the L2. Thus,

future research should focus on testing the applicability of this particular combination of tasks to bilinguals with less exposure to the L2.

AUTHOR CONTRIBUTIONS

LAI developed the theoretical framework, gathered and evaluated the data. IF conducted the statistical analysis. All authors wrote the manuscript.

FUNDING

The BiLaD project (Bilingual Language Development: Typically Developing Children and Children with Specific Language Impairment) was financed by a joint grant (German DFG: HA 2335/6-1, CH 1112/2-1, and RO 923/3-1) and French ANR grant (ANR-12-FRAL-0014-01) to Laurice Tuller and her team.

ACKNOWLEDGMENTS

We deeply thank Cornelia Hamann for her continued support and constructive comments. We are grateful to Solveig Chilla, Monika Rothweiler, Hilal Şan, Tatjana Lein, and our French BiLaD partners for their support with data collection and analysis. Our special thanks go to the participating children and their parents as well as to the collaborating kindergartens and speech-language therapy centers. We also thank the three reviewers for their constructive comments and suggestions on previous drafts of the manuscript.

REFERENCES

- Abed Ibrahim, L., and Hamann, C. (2017). "Bilingual Arabic-German & Turkish-German children with and without specific language impairment: comparing performance in sentence and Nonword repetition tasks," in *Proceedings of BUCLD 41*, eds M. LaMendola and J. Scott (Somerville, MA: Cascadilla Press), 1–17.
- Abed Ibrahim, L., Hamann, C., and Öwerdieck, D. (2018). "Identifying specific language impairment (SLI) across different bilingual populations: German sentence repetition task (SRT)," in *Proceedings of the 42nd Annual Boston University Conference on Language Development*, eds A. B. Bertolini and M. J. Kaplan (Somerville, MA: Cascadilla Press), 1–14.
- American Speech-Language-Hearing Association [ASHA] (2004). *Knowledge and Skills Needed by Speech-Language Pathologists and Audiologists to Provide Culturally and Linguistically Appropriate Services [Knowledge and Skills]*. Available at: www.asha.org/policy
- Archibald, L. M., and Gathercole, S. E. (2006). Nonword repetition: a comparison of tests. *J. Speech Lang. Hear. Res.* 49, 970–983. doi: 10.1044/1092-4388(2006/070)
- Archibald, L. M., and Gathercole, S. E. (2007). The complexities of complex memory span: storage and processing deficits in specific language impairment. *J. Mem. Lang.* 57, 177–194. doi: 10.1016/j.jml.2006.11.004
- Argyri, E., and Sorace, A. (2007). Crosslinguistic influence and language dominance in older bilingual children. *Biling. Lang. Cogn.* 10, 79–99. doi: 10.1017/S1366728906002835
- Armon-Lotem, S. (2010). Instructive bilingualism: can bilingual children with specific language impairment rely on one language in learning a second one? *Appl. Psycholinguist.* 31, 253–260. doi: 10.1017/S0142716409990385
- Armon-Lotem, S. (2011). Design and data of English-Hebrew and Russian-Hebrew sentence repetition tasks. *Paper Presented at COST IS0804 WG and MC Meeting*, Malta.
- Armon-Lotem, S., and de Jong, J. (2015). "Introduction," in *Disentangling Bilingualism from Language Impairment*, eds S. Armon-Lotem, J. de Jong, and N. Meir (Bristol: Multilingual Matters). doi: 10.21832/9781783093137
- Armon-Lotem, S., de Jong, J., and Meir, N. (eds). (2015). *Assessing Multilingual Children: Disentangling Bilingualism From Language Impairment*. Bristol: Multilingual Matters. doi: 10.21832/9781783093137
- Armon-Lotem, S., and Meir, N. (2016). Diagnostic accuracy of repetition tasks for the identification of specific language impairment (SLI) in bilingual children: evidence from Russian and Hebrew. *Int. J. Lang. Commun. Disord.* 51, 715–731. doi: 10.1111/1460-6984.12242
- Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends Cogn. Sci.* 4, 417–423. doi: 10.1016/S1364-6613(00)01538-2
- Balladares, J., Marshall, C., and Griffiths, Y. (2016). Socio-economic status affects sentence repetition, but not non-word repetition, in Chilean preschoolers. *First Lang.* 36, 338–351. doi: 10.1177/0142723715626067
- Banerjee, A., and Davé, R. N. (2004). "Validating clusters using the Hopkins statistic," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, Piscataway, NJ, 149–153. doi: 10.1109/FUZZY.2004.1375706
- Barlow, J. A. (2001). The structure of /s/ sequences: evidence from a disordered system. *J Child Lang.* 28, 291–324. doi: 10.1017/S0305000901004652
- Bedore, L. M., and Peña, E. D. (2008). Assessment of bilingual children for identification of language impairment: current findings and implications for practice. *Int. J. Biling. Educ. Biling.* 11, 1–29. doi: 10.2167/beb392.0

- Bedore, L. M., Peña, E. D., Summers, C. L., Boerger, K. M., Resendiz, M. D., Green, K., et al. (2012). The measure matters: language dominance profiles across measures in Spanish-English bilingual children. *Biling. Lang. Cogn.* 15, 616–629. doi: 10.1017/S1366728912000090
- Benmamoun, E., Montrul, S., and Polinsky, M. (2013). Heritage languages and their speakers: opportunities and challenges for linguistics. *Theor. Linguist.* 39, 129–181. doi: 10.1515/tl-2013-0009
- Bishop, D. V., Adams, C. V., and Norbury, C. F. (2006). Distinct genetic influences on grammar and phonological short-term memory deficits: evidence from 6-year-old twins. *Genes Brain Behav.* 5, 158–169. doi: 10.1111/j.1601-183X.2005.00148.x
- Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., and the Catalise-2 consortium. (2017). Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: terminology. *J. Child Psychol. Psychiatry* 58, 1068–1080. doi: 10.1111/jcpp.12721
- Boerma, T., and Blom, E. (2017). Assessment of bilingual children: what if testing both languages is not possible? *J. Commun. Disord.* 66, 65–76. doi: 10.1016/j.jcomdis.2017.04.001
- Boerma, T., Chiat, S., Leleman, P., Timmermeister, M., Wijnen, F., and Blom, E. (2015). A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning dutch as a second language. *J. Speech Lang. Hear. Res.* 58, 1747–1760. doi: 10.1044/2015_JSLHR-L-15-0058
- Bortolini, U., and Leonard, L. (2000). Phonology and children with specific language impairment: the status of structural constraints in two languages. *J. Commun. Disord.* 33, 131–150. doi: 10.1016/S0021-9924(99)00028-3
- Bulheller, S., and Häcker, H. (2002). *Coloured Progressive Matrices*. Frankfurt: Pearson Assessment.
- Calvo, A., and Bialystok, E. (2014). Independent effects of bilingualism and socioeconomic status on language ability and executive functioning. *Cognition* 130, 278–288. doi: 10.1016/j.cognition.2013.11.015
- Cantone, K. F., Müller, N., Schmitz, K., and Kupisch, T. (2008). Rethinking language dominance in bilingual children. *Linguistische Berichte* 215, 307–343.
- Castro, S. L., Caló, S., Gomes, I., Kay, J., Lesser, R., and Coltheart, M. (2007). *PALPA-P Provas de Avaliao da Linguagem e da Afasia em Portugues*. Lisbon: CEGOC.
- Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* 61, 1–36. doi: 10.18637/jss.v061.i06
- Chiat, S. (2015). “Non-word repetition,” in *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*, eds S. Armon-Lotem, J. de Jong, and N. Meir (Bristol: Multilingual Matters), 123–148.
- Chiat, S., Armon-Lotem, S., Marinis, T., Polišenská, K., Roy, P., and Seeff-Gabriel, B. (2013). “The potential of sentence imitation tasks for assessment of language abilities in sequential bilingual children,” in *Issues in the Assessment of Bilinguals*, ed. V. Mueller-Gathercole (Bristol: Multilingual Matters), 56–89.
- Chiat, S., and Polišenská, K. (2016). A framework for crosslinguistic nonword repetition tests: effects of bilingualism and socioeconomic status on children’s performance. *J. Speed. Lang. Hear. Res.* 59, 1179–1189. doi: 10.1044/2016_JSLHR-L-15-0293
- Chilla, S., Hamann, C., Prévost, P., Abed Ibrahim, L., Ferré, S., dos Santos, C., et al. (in press). “The influence of different first languages on L2 LITMUS-NWR and L2 LITMUS-SRT in French and German: a crosslinguistic approach”, in *LITMUS in Action: Comparative Studies across Europe, TILAR*, eds K. Grohmann and S. Armon-Lotem (Amsterdam: John Benjamins Publishing Company).
- Chilla, S., and Şan, H. (2017). “Möglichkeiten und Grenzen der Diagnostik erstsprachlicher Fähigkeiten: Türkisch-deutsche und türkisch-französische Kinder im Vergleich,” in *Sprachen 2016. Russisch und Türkisch im Fokus*, eds C. Yildiz et al. (Berlin: Peter Lang), 175–205.
- Chomsky, N. (1986). *Knowledge of Language: its Nature, Origin and Use*. New York, NY: Praeger.
- Chomsky, N. (2005). Three factors in language design. *Linguist. Inq.* 36, 1–22. doi: 10.1162/0024389052993655
- Chondrogianni, V., and Marinis, T. (2011). Differential effects of internal and external factors on the development of vocabulary, tense morphology and morpho-syntax in successive bilingual children. *Linguist. Approaches Biling.* 1, 318–345. doi: 10.1075/lab.1.3.05cho
- Clahsen, H. (1991). *Child Language and Developmental Dysphasia: Linguistic Studies of the Acquisition of German*. Amsterdam: John Benjamins. doi: 10.1075/sspcl.2
- Coady, J., and Evans, J. (2008). Uses and interpretations of nonword repetition tasks in children with and without specific language impairments. *Int. J. Lang. Commun. Disord.* 43, 1–40. doi: 10.1080/13682820601116485
- Conti-Ramsden, G., Botting, N., and Faragher, B. (2001). Psycholinguistic markers for specific language impairment. *J. Child Psychol. Psychiatry* 42, 741–748. doi: 10.1111/1469-7610.00770
- Crutchley, A., Botting, N., and Conti-Ramsden, G. (1997). Bilingualism and specific language impairment in children attending language units. *Int. J. Lang. Commun. Disord.* 32, 267–276. doi: 10.3109/13682829709020409
- Dalal, D., and Zickar, M. (2011). Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organ. Res. Methods* 15, 339–362. doi: 10.1177/1094428111430540
- de Almeida, L., Ferré, S., Morin, E., Prévost, P., dos Santos, C., Tuller, L., et al. (2017). Identification of bilingual children with specific language impairment in France. *Linguist. Approaches Biling.* 7, 331–358. doi: 10.1075/lab.15019.alm
- de Houwer, A. (2009). *Bilingual First Language Acquisition*. Bristol: Multilingual Matters.
- de Houwer, A., and Bornstein, M. H. (2016). “Balance patterns in early bilingual acquisition: a longitudinal study of word comprehension and production,” in *Language Dominance in Bilinguals: Issues of Measurement and Operationalization*, eds C. Silva-Corvalán and J. Treffers-Daller (Cambridge: Cambridge University Press), 134–155. doi: 10.1017/CBO9781107375345.007
- Deuchar, M., and Muntz, R. (2003). “Factors accounting for code-mixing in an early developing bilingual,” in *(In)vulnerable Domains in Multilingualism*, ed. N. Müller (Amsterdam: John Benjamins), 161–190.
- Dollaghan, C., and Horner, E. (2011). Bilingual language assessment: a meta-analysis of diagnostic accuracy. *J. Speech Lang. Hear. Res.* 54, 1077–1088. doi: 10.1044/1092-4388(2010/10-0093)
- Döpke, S. (2000). *One Parent-One Language: An Interactional Approach*. Amsterdam: Benjamins.
- dos Santos, C., and Ferré, S. (2018). A non-word repetition task to assess bilingual children’s phonology. *Lang. Acquis.* 25:1. doi: 10.1080/10489223.2016.1243692
- Eisenberg, P. (1999). *Der Satz. Grundriß der deutschen Grammatik*. Stuttgart: J. B. Metzler. doi: 10.1007/978-3-476-03765-7
- Engel, P. M. J., Santos, F. H., and Gathercole, S. E. (2008). Are working memory measures free of socioeconomic influence? *J. Speech Lang. Hear. Res.* 51, 1580–1587. doi: 10.1044/1092-4388(2008/07-0210)
- Engel de Abreu, P. M. J., Baldassi, M., Puglisi, M. L., and Befi-Lopes, D. M. (2013). Cross-linguistic and cross-cultural effects on verbal working memory and vocabulary: testing language minority children with an immigrant background. *J. Speech Lang. Hear. Res.* 56, 630–642. doi: 10.1044/1092-4388(2012/12-0079)
- Ferré, S., dos Santos, C., and de Almeida, L. (2015). “Potential Phonological Markers for SLI in Bilingual Children,” in *Proceedings of the 39th Annual Boston University Conference on Language Development* (Somerville, MA: Cascadia Press), 152–164.
- Ferré, S., Tuller, L., Sizaret, E., and Barthez, M.-A. (2012). “Acquiring and avoiding phonological complexity in SLI vs. typical development of French: the case of consonant clusters,” in *Consonant Clusters and Structural Complexity*, eds P. Hoole, L. Bombine, M. Pouplier, C. Mooshammer, and B. Kühnert (Berlin: de Gruyter), 285–308. doi: 10.1515/9781614510772.285
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80, 27–38. doi: 10.1093/biomet/80.1.27
- Fleckstein, A., Prévost, P., Tuller, L., Sizaret, E., and Zebib, R. (2018). How to identify SLI in bilingual children: a study on sentence repetition in French. *Lang. Acquis.* 25:1. doi: 10.1080/10489223.2016.1192635
- Flores, C. (2015). Losing a language in childhood: a longitudinal case study on language attrition. *J. Child Lang.* 42, 562–590. doi: 10.1017/S0305000914000233
- Fox-Boyer, A. V. (2014). *Psycholinguistischer Analyse kindlicher Aussprachestörungen-II: PLAKSS II*. Frankfurt: Pearson.

- Fredman, M. (2006). Recommendations for working with bilingual children – Prepared by the multilingual affairs committee of IALP. *Folia Phoniatri. Logop.* 58, 458–464.
- Friedmann, N., Belletti, A., and Rizzi, L. (2009). Relativized relatives: types of intervention in the acquisition of A-bar dependencies. *Lingua* 119, 67–88. doi: 10.1016/j.lingua.2008.09.002
- Friedmann, N., and Novogrodsky, R. (2008). “Subtypes of SLI: SySLI, PhoSLI, LeSLI, and PraSLI”, in *Language Acquisition and Development*, eds A. Garvaro and M. J. Freitas (Cambridge: Cambridge University), 205–217.
- Friedmann, N., and Novogrodsky, R. (2011). Which questions are most difficult to understand? The comprehension of wh-questions in three subtypes of SLI. *Lingua* 121, 367–382. doi: 10.1016/j.lingua.2010.10.004
- Friedmann, N., Yachini, M., and Szterman, R. (2015). “Relatively easy relatives: children with syntactic SLI avoid intervention,” in *Structures, Strategies and Beyond. Studies in Honour of Adriana Belletti*, eds E. Di Domenico, C. Hamann, and S. Matteini (Amsterdam: John Benjamins), 303–320. doi: 10.1075/la.223.14fri
- Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balciuniene, I., et al. (2015). “Assessment of narrative abilities in bilingual children,” in *Assessing Multilingual Children: Disentangling Bilingualism From Language Impairment*, eds S. Armon-Lotem, J. de Jong, and N. Meir (Bristol: Multilingual Matters), 243–276. doi: 10.21832/9781783093137-011
- Gallon, N., Harris, J., and Van der Lely, H. (2007). Non-word repetition: an investigation of phonological complexity in children with grammatical SLI. *Clin. Linguist. Phon.* 21, 445–455. doi: 10.1080/02699200701299982
- Gathercole, S. (2006). Complexities and constraints in nonword repetition and word learning. *Appl. Psycholinguist.* 27, 599–613. doi: 10.1017/S014271640606053X
- Gathercole, S., and Baddeley, A. (1990). Phonological memory deficits in language disordered children: is there a causal connection? *J. Mem. Lang.* 29, 336–360. doi: 10.1016/0749-596X(90)90004-J
- Gathercole, V. C. M. (2016). “Factors moderating proficiency in bilingual speakers,” in *Bilingualism Across the Lifespan: Factors Moderating Language Proficiency*, eds E. Nicoladis and S. Montanari (Washington DC: De Gruyter Mouton and American Psychological Association), 123–140. doi: 10.1037/14939-008
- Gathercole, V. C. M., and Thomas, E. M. (2009). Bilingual first-language development: dominant language takeover, threatened minority language take-up. *Biling. Lang. Cogn.* 12, 213–237. doi: 10.1017/S1366728909004015
- Genesee, F., and Nicoladis, E. (2007). “Bilingual first language acquisition,” in *Blackwell Handbook of Language Development*, eds E. Hoff and M. Shtatz (Oxford: Blackwell Publishing Ltd.), 324–342. doi: 10.1002/9780470757833.ch16
- Genesee, F., Paradis, J., and Crago, M. (2004). *Dual Language Development and Disorders: a Handbook on Bilingualism and Second Language Learning*. Baltimore: Brookes Publishing.
- Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1
- Glück, C. (2011). *Wortschatz- und Wortfindungstest für 6- bis 10-Jährige: WWT 6–10*, 2nd Edn. Munich: Elsevier, Urban & Fischer.
- Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol.* 27, 1411–1423. doi: 10.1037/0278-7393.27.6.1411
- Graf Estes, K., Evans, J. L., and Else-Quest, N. (2007). Differences in non-word repetition performance for children with and without specific language impairment: a meta-analysis. *J. Speech Lang. Hear. Res.* 50, 177–195. doi: 10.1044/1092-4388(2007/015)
- Grimm, A., Ferré, S., dos Santos, C., and Chiat, S. (2014). “Can nonwords be language-independent? Cross-linguistic evidence from monolingual and bilingual acquisition of French, German, and Lebanese,” in *Symposium Language Impairment Testing in Multilingual Setting (LITMUS): Disentangling bilingualism and SLI*, (Amsterdam: IASCL).
- Grimm, A., and Schulz, P. (2014). Specific language impairment and early second language acquisition: the risk of over- and underdiagnosis. *Child Indic. Res.* 7, 821–841. doi: 10.1007/s12187-013-9230-6
- Grimm, A., and Hübner, J. (in press). “Nonword repetition by bilingual learners of German: the role of language-specific complexity,” in *Bilingualism and Specific Language Impairment, Bi-SLI 201*, eds C. dos Santos and L. de Almeida (Amsterdam: Benjamins).
- Grosjean, F. (2010). *Bilingual: Life and Reality*. Cambridge, MA: Harvard University Press. doi: 10.4159/9780674056459
- Grosjean, F. (2016). “The complementarity principle and its impact on processing, acquisition and dominance,” in *Language Dominance in Bilinguals*, eds C. Silva-Corvalán and J. Treffers-Daller (Cambridge: Cambridge University Press), 66–85.
- Håkansson, G., and Nettelbladt, U. (1996). “Similarities between SLI and L2 children. Evidence from the acquisition of Swedish word order,” in *Children’s Language*, Vol. 9, eds C. Johnson and J. Gilbert (Philadelphia, PA: Gardner Press), 135–151.
- Hamann, C. (2012). “Bilingual development and language assessment,” in *Proceedings of BUCLD 36*, eds A. K. Biller, E. Y. Chung, and A. E. Kimball (Somerville, MA: Cascadilla Press), 1–28.
- Hamann, C., and Abed Ibrahim, L. (2017). Methods for identifying specific language impairment in bilingual populations in Germany. *Front. Commun.* 2:19. doi: 10.3389/fcomm.2017.00016
- Hamann, C., Chilla, S., Gagarina, N., and Abed Ibrahim, L. (2017). “Syntactic complexity and bilingualism: how (a)typical bilinguals deal with complex structures,” in *Complexity in Acquisition*, ed. E. Di Domenico (Cambridge: Scholars Publishing), 142–178.
- Hamann, C., Chilla, S., Ruigendijk, E., and Abed Ibrahim, L. (2013). “A German sentence repetition task: testing bilingual Russian/German children,” in *Poster Presented at the COST Meeting in Krakow, Kraków*.
- Hamann, C., Penner, Z., and Lindner, K. (1998). German impaired grammar: the clause structure revisited. *Lang. Acquis.* 7, 193–246. doi: 10.1207/s15327817la0702-4_5
- Hamann, C., and Tuller, L. (2014). Genuine versus superficial relatives in French: the depth of embedding factor. *Revisita Grammatica Generativa* 36, 146–181.
- Hamann, C., Tuller, L., Monjauze, C., Delage, H., and Henry, C. (2007). “(Un)successful subordination in French-speaking children and adolescents with SLI,” in *Proceedings of BUCLD 31*, eds H. Cauntulton, S. Kalatilake, and I. Woo (Somerville, MA: Cascadilla Press), 286–297.
- Hoff, E. (2003). The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech. *Child Dev.* 74, 1368–1378. doi: 10.1111/1467-8624.00612
- Hoff, E. (2006). How social contexts support and shape language development. *Dev. Rev.* 26, 55–88. doi: 10.1016/j.dr.2005.11.002
- Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., and Parra, M. (2012). Dual language exposure and early bilingual development. *J. Child Lang.* 39, 1–27. doi: 10.1017/S0305000910000759
- Hopkins, B., and Skellam, J. G. (1954). A new method for determining the type of distribution of plant individuals. *Ann. Bot.* 18, 213–227. doi: 10.1093/oxfordjournals.aob.a083391
- Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Stat.* 15, 651–674. doi: 10.1198/106186006X133933
- IBM SPSS 24 (2016). *IBM SPSS Statistics for Windows, Version 24.0*. Armonk, NY: IBM Corp.
- International Association of Logopedics and Phoniatrics [IALP] (2011). *Recommendations for Working With Bilingual Children*. Available at: http://www.specchioriflesso.net/media/162083/linea_guida_bilingui_ialp_may_2011.pdf
- Jakubowicz, C. (2005). The language faculty: (Ab)normal development and interface constraints. *Paper Presented at Generative Approaches to Language Acquisition, GALA, 2005*, Siena.
- Jakubowicz, C. (2011). Measuring derivational complexity: new evidence from typically developing and SLI learners of L1 French. *Lingua* 121, 339–351. doi: 10.1016/j.lingua.2010.10.006
- Jakubowicz, C., Nash, L., Rigaut, C., and Gérard, C.-L. (1998). Determiners and clitic pronouns in French-speaking children with SLI. *Lang. Acquis.* 7, 113–160. doi: 10.1080/02699206.2018.1437222
- Jakubowicz, C., and Tuller, L. (2008). “Specific language impairment in French,” in *Studies in French applied linguistics*, ed. D. Ayoun (Amsterdam: John Benjamins), 97–134.

- Jones, G., Tamburelli, M., Watson, S. E., Gobet, F., and Pine, J. M. (2010). Lexicality and frequency in specific language impairment: Accuracy and error data from two nonword repetition tests. *J. Speech Lang. Hear. Res.* 53, 1642–1655. doi: 10.1044/1092-4388(2010/09-0222)
- Kashef, R., and Kamel, M. S. (2008). “Efficient bisecting k-medoids and its application in gene expression analysis,” in *ICIAR 2008 Proceedings of the 5th International Conference on Image Analysis and Recognition, LNCS 5112*, eds A. Campilho and M. Kamel (Heidelberg: Springer), 423–434. doi: 10.1007/978-3-540-69812-8_42
- Kassambara, A., and Mundt, F. (2017). *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses, R Package Version 1.0.5*. Available at: <https://cran.r-project.org/web/packages/factoextra/factoextra.pdf> [accessed June 1, 2018].
- Kaufman, L., and Rousseeuw, P. J. (1987). “Clustering by means of Medoids,” in *Statistical Data Analysis Based on the L1-Norm and Related Methods*, ed. Y. Dodge (Basel: Birkhäuser Basel), 405–416.
- Kaufman, L., and Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons, Inc.
- Kohmert, K. (2010). Bilingual children with primary language impairment: issues, evidence and implications for clinical actions. *J. Commun. Disord.* 43, 456–473. doi: 10.1016/j.jcomdis.2010.02.002
- Kohmert, K., Windsor, J., and Yim, D. (2006). Do language-based processing tasks separate children with language impairment from typical bilinguals? *Learn. Disabil. Res. Pract.* 21, 19–29. doi: 10.1016/j.jcomdis.2010.02.002
- Kosmidis, I. (2018). *brglm2: Bias Reduction in Generalized Linear Models. R Package Version 0.1.7*. Available at: <https://github.com/ikosmidis/brglm2>
- Krishna, S., Yesu Babu, A., and Kiran Kumar, R. (2018). “Determination of optimal clusters for a Non-hierarchical clustering paradigm K-Means algorithm,” in *Proceedings of International Conference on Computational Intelligence and Data Engineering*, eds N. Devarakonda, N. Chaki, and A. Cortesi (Singapore: Springer), 301–316. doi: 10.1007/978-981-10-6319-0_26
- Kupisch, T. (2008). “Dominance, mixing and cross-linguistic influence: on their relation in bilingual development,” in *First Language Acquisition of Morphology and Syntax: Perspectives Across Languages and Learners*, eds P. Guijarro-Fuentes, P. Larrañaga, and J. Clibbens (Amsterdam: John Benjamins), 209–234.
- Kutner, M. H., Nachtsheim, C. J., and Neter, J. (2004). *Applied Linear Regression Models*, 4th Edn. Chicago, IL: McGraw-Hill/ Irwin.
- Leclercq, A. L., Maillart, C., and Majerus, S. (2013). Nonword repetition problems in children with specific language impairment: a deficit in accessing long-term linguistic representations? *Top. Lang. Disord.* 33, 238–254. doi: 10.1097/TLD.0b013e31829dd8c9
- Lein, T., Hamann, C., Rothweiler, M., Abed Ibrahim, L., Chilla, S., and San, H. (2016). “SLI in bilinguals: testing complex syntax and semantics in German,” in *Proceedings of the 2015 GASLA*, eds D. Stringer, J. Garrett, B. Halloran, and S. Mossman (Somerville, MA: Cascadilla Proceedings), 124–135.
- Leonard, L. (1998). *Children With Specific Language Impairment*. Cambridge: MIT Press.
- Leonard, L. (2007). Processing limitations and the grammatical profile of children with specific language impairment. *Adv. Child Dev. Behav.* 35, 139–171. doi: 10.1016/B978-0-12-009735-7.50009-8
- Leonard, L. (2014). *Children With Specific Language Impairment*, 2nd Edn. Cambridge: MIT Press.
- Leonhard, L. (2014). Children with specific language impairment and their contribution to the study of language development. *J. Child Lang.* 41, 37–48. doi: 10.1017/S0305000914000130
- Levshina, N. (2015). *How to do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins Publishing Company. doi: 10.1075/z.195
- Lindener, K. (2002). Finiteness and children with specific language impairment: an exploratory study. *Linguistics* 40, 797–847. doi: 10.1515/ling.2002.031
- Long, S. L. (1997). *Regression Models for Categorical and Limited Dependent Variables*. London: SAGE Publications.
- Maddieson, I. (2006). Correlating phonological complexity: data and validation. *Linguist. Typol.* 10, 106–123.
- Maddieson, I., Flavier, S., Marsico, E., and Pellegrino, F. (2011). *LAPSYd: Lyon-Albuquerque Phonological Systems Databases, Version 1.0*. Available at: <http://www.lapsyd.ddl.ish-lyon.cnrs.fr/>
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2017). *Cluster: Cluster Analysis Basics and Extensions. R package version 2.0.1.2015*.
- Marinis, T. (2011). On the nature and cause of specific language impairment: a view from sentence processing and infant research. *Lingua* 121, 463–475. doi: 10.1016/j.lingua.2010.10.010
- Marinis, T., and Armon-Lotem, S. (2015). “Sentence repetition,” in *Assessing Multilingual Children: Disentangling Bilingualism From Language Impairment*, eds S. Armon-Lotem, J. de Jong, and N. Meir (Bristol: Multilingual Matters), 95–124.
- Marinis, T., Armon-Lotem, S., and Pontikas, G. (2017). “Language impairment in bilingual children – state of the art 2017,” in *Linguistic Approaches to Bilingualism*, eds J. Rothman and S. Unsworth (Amsterdam: John Benjamins), 265–276.
- Marinis, T., and van der Lely, H. (2007). On-line processing of wh-questions in children with G-SLI and typically developing children. *Int. J. Lang. Commun. Disord.* 42, 557–582. doi: 10.1080/13682820601058190
- Marshall, C., Ebbels, S., Harris, J., and van der Lely, H. K. (2002). Investigating the impact of prosodic complexity on the speech of children with specific language impairment. *UCL Work. Pap. Linguist.* 14, 43–66.
- Marshall, C., Harris, J., and van der Lely, H. (2003). “The nature of phonological representations in children with grammatical specific language impairment,” in *Proceedings of the University of Cambridge First Postgraduate Conference in Language Research*, (Cambridge: Institute of Language Research Cambridge), 511–517.
- Marshall, C. R. (2014). Word production errors in children with developmental language impairments. *Philos. Trans. R. Soc. B Biol. Sci.* 369:20120389. doi: 10.1098/rstb.2012.0389
- Marshall, C. R., and van der Lely, H. (2009). Effects of word position and stress on onset cluster production: evidence from typical development, SLI and dyslexia. *Language* 85, 39–57. doi: 10.1353/lan.0.0081
- Meir, N. (2018). Morpho-syntactic abilities of unbalanced bilingual children: a closer look at the weaker language. *Front. Psychol.* 9:1318. doi: 10.3389/fpsyg.2018.01318
- Meir, N., and Armon-Lotem, S. (2017). Independent and combined effects of socioeconomic Status (SES) and bilingualism on children’s vocabulary and verbal short-term memory. *Front. Psychol.* 8:1442. doi: 10.3389/fpsyg.2017.01442
- Meir, N., Walters, J., and Armon-Lotem, S. (2016). Disentangling SLI and bilingualism using sentence repetition tasks: the impact of L1 and L2 properties. *Int. J. Biling.* 20, 421–452. doi: 10.1177/1367006915609240
- Meir, N., Walters, J., and Armon-Lotem, S. (2017). Bi-directional cross-linguistic influence in bilingual Russian-Hebrew children. *Linguist. Approaches Biling.* 7, 514–553. doi: 10.1075/lab.15007.mei
- Meisel, J. (2007). The weaker language in early child bilingualism: acquiring a first language as a second language? *Appl. Psycholinguist.* 28, 495–514. doi: 10.1017/S0142716407070270
- Meisel, J. (2009). Second language acquisition in early childhood. *Zeitschrift Sprachwissenschaft* 28, 5–34. doi: 10.1515/ZFSW.2009.002
- Messer, M. H., Leseman, P. P. M., Boom, J., and Mayo, A. Y. (2010). Phonotactic probability effect in nonword recall and its relationship with vocabulary in monolingual and bilingual preschoolers. *J. Exp. Child Psychol.* 105, 306–323. doi: 10.1016/j.jecp.2009.12.006
- Montrul, S. (2008). *Incomplete Acquisition in Bilingualism. Re-examining the Age Factor*. Amsterdam: Benjamins. doi: 10.1075/sibil.39
- Novogrodsky, R., and Kreiser, V. (2015). What can errors tell us about specific language impairment deficits? Semantic and morphological cuing in a sentence completion task. *Clin. Linguist. Phon.* 29, 812–825. doi: 10.3109/02699206.2015.1051239
- Oetting, J. B., Cleveland, L., and Cope, R. (2008). Empirically derived combinations of tools and clinical cutoffs: an illustrative case with a sample of culturally/linguistically diverse children. *Lang. Speech Hear. Serv. Schl.* 39, 44–53. doi: 10.1044/0161-1461(2008/005)
- Orsolini, M., Sechi, E., Maronato, C., Bonvino, E., and Corcelli, A. (2001). Nature of phonological delay in children with specific language impairment. *Int. J. Lang. Commun. Disord.* 36, 63–90. doi: 10.1080/13682820150217572

- Ott, S., van de Vijver, R., and Höhle, B. (2006). The effect of phonotactic constraints in German-speaking children with delayed phonological acquisition: evidence from production of word-initial consonant clusters. *Adv. Speech Lang. Pathol.* 8, 323–334. doi: 10.1080/14417040600970622
- Paradis, J. (2009). *Oral Language Development in French and English and the Role of Home Input Factors*. Conseil scolaire centre-nord: Edmonton.
- Paradis, J. (2010). The interface between bilingual development and specific language impairment. Keynote article for special issue with peer commentaries. *Appl. Psycholinguist.* 31, 3–28.
- Paradis, J. (2011). Individual differences in child English second-language acquisition: comparing child-internal and child-external factors. *Linguist. Approaches Biling.* 1, 834–847. doi: 10.1044/jslhr.4304.834
- Paradis, J., Crago, M., and Genesee, F. (2003). “Object clitics as clinical markers of SLI in French: evidence from French-English bilingual children,” in *Proceedings of the 27th Annual Boston University Conference on Language Development*, eds B. Beachley, A. Brown, and F. Conlin (Somerville: Cascadilla Press), 638–649.
- Paradis, J., Emmers, K., and Sorenson Duncan, T. (2010). Assessment of English language learners: using parent report on first language development. *J. Commun. Disord.* 43, 474–497. doi: 10.1016/j.jcomdis.2010.01.002
- Paradis, J., and Genesee, F. (1996). Syntactic acquisition in bilingual children: autonomous or interdependent? *Stud. Second Lang. Acquis.* 18, 1–25. doi: 10.1017/S0272263100014662
- Paradis, J., Genesee, F., and Crago, M. (2011a). *Dual Language Development and Disorders: A Handbook on Bilingualism and Second Language Learning*, 2nd Edn. Baltimore, MD: Brookes.
- Paradis, J., Nicoladis, E., Crago, M., and Genesee, F. (2011b). Bilingual children’s acquisition of the past tense: a usage-based approach. *J. Child Lang.* 38, 544–578. doi: 10.1017/S0305000910000218
- Paradis, J., and Jia, R. (2016). Bilingual children’s long-term outcomes in English as a second language: language environment factors shape individual differences in catching up with monolinguals. *Dev. Sci.* 20:1. doi: 10.1111/desc.12433
- Paradis, J., Tulpar, Y., and Arppe, A. (2016). Chinese L1 children’s English L2 verb morphology over time: individual variation in long-term outcomes. *J. Child Lang.* 43, 553–580. doi: 10.1017/S0305000915000562
- Patel, A., and Singh, P. (2013). New approach for K-mean and K-medoids algorithm. *Int. J. Comput. Appl. Technol. Res.* 2, 1–5.
- Pearson, B. Z., Fernald, S. C., Lewedag, V., and Oller, D. K. (1997). The relation of input factors to lexical learning by bilingual infants (ages 10 to 30 months). *Appl. Psycholinguist.* 18, 41–58. doi: 10.1017/S0142716400009863
- Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* 49, 1373–1379. doi: 10.1016/S0895-4356(96)00236-3
- Petersen, J. (1988). Word-internal code-switching constraints in a bilingual child’s grammar. *Linguistics* 26, 479–493. doi: 10.1515/ling.1988.26.3.479
- Plante, E., and Vance, R. (1994). Diagnostic accuracy of two tests of preschool language. *Am. J. Speech Lang. Pathol.* 4, 70–76. doi: 10.1044/1058-0360.0402.70
- Polišenská, K., Chiat, S., and Roy, P. (2015). Sentence repetition: what does the task measure? *Int. J. Lang. Commun. Disord.* 50, 106–118. doi: 10.1111/1460-6984.12126
- Rainey, C. (2016). Dealing with separation in logistic regression models. *Polit. Anal.* 24, 339–355. doi: 10.1093/pan/mpw014
- Reilly, S., Wake, M., Ukoumunne, O., Bavin, E., Prior, M., Cini, E., et al. (2010). Predicting language outcomes at 4 years of age: findings from early language in victoria study. *Pediatrics* 126, 1530–1537. doi: 10.1542/peds.2010-0254
- Rice, M., Noll, K., and Grimm, H. (1997). An extended optional infinitive stage in German-speaking children with specific language impairment. *Lang. Acquis.* 61, 255–295. doi: 10.1207/s15327817la0604_1
- Rice, M., and Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *J. Speech Lang. Hear. Res.* 39, 1239–1257. doi: 10.1044/jshr.3906.1239
- Rispens, J., and Baker, A. (2012). Nonword repetition: the relative contributions of phonological short-term memory and phonological representations in children with language and reading impairment. *J. Speech Lang. Hear. Res.* 55, 683–694. doi: 10.1044/1092-4388(2011/10-0263)
- Rizzi, L. (2004). “Locality and left periphery,” in *Structures and Beyond. The Cartography of Syntactic Structures*, Vol. 3, ed. A. Belletti (Oxford: OUP), 223–251.
- Rizzi, L. (2013). Locality. *Lingua* 130, 169–186. doi: 10.1016/j.lingua.2012.12.002
- Roeper, T. (2004). Diagnosing language variations: underlying principles for syntactic assessment. *Semin. Speech Lang.* 25, 41–56. doi: 10.1055/s-2004-824825
- Roeper, T. (2011). The acquisition of recursion: how formalism articulates in the child’s path. *Biolinguistics* 5, 57–87.
- Rothweiler, M., Chilla, S., and Clahsen, H. (2012). Subject verb agreement in specific language impairment: a study of monolingual and bilingual German-speaking children. *Biling. Lang. Cogn.* 15, 39–57.
- Roy, P., and Chiat, S. (2004). A prosodically controlled word and nonword repetition task for 2- to 4-year-olds: evidence from typically developing children. *J. Speech Lang. Hear. Res.* 47, 223–234. doi: 10.1044/1092-4388(2004/019)
- Roy, P., Chiat, S., and Dodd, B. (2014). *Language and Socioeconomic Disadvantage: From Research to Practice*. London: University of London.
- Royal College of Speech and Language Therapists Specific Interest Group in Bilingualism [RCSLT] (2007). *Good Practice for Speech and Language Therapists Working With Clients From Linguistic Minority Communities*. London: RCSLT.
- R-Studio (2012). *RStudio: Integrated Development Environment for R (Version 0.99.892)*. Boston, MA: R-Studio.
- Scheidenes, M., and Tuller, L. (2018). Using clausal embedding to identify language impairment in sequential bilinguals. *Biling. Lang. Cogn.* 1–19. doi: 10.1017/S1366728918000949
- Schroeder, C., and Dollnick, M. (2013). “Mehrsprachige gymnasien mit türkischem hintergrund schreiben auf türkisch,” in *Mehrsprachig in Wissenschaft und Gesellschaft. Tagungsband*, eds C. Riemer, H. Brandl, E. Arslan, and E. Langelahn (Bielefeld: University of Bielefeld), 101–114.
- Schulz, P., and Roeper, T. (2011). Acquisition of exhaustivity in wh-questions: a semantic dimension of SLI? *Lingua* 121, 383–407. doi: 10.1016/j.lingua.2010.10.005
- Schulz, P., and Tracy, R. (2011). *Linguistische Sprachstandserhebung – Deutsch als Zweitsprache (LiSe-DaZ)*. Göttingen: Hofgrefe Verlag.
- Soni, K. G., and Patel, A. (2017). Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. *Int. J. Comput. Intell. Res.* 13, 899–906.
- Sua-Kay, E., and Santos, M. E. (2014). *Grelha de Avaliacao da Linguagem – Nível Escolar (GOL-E) 2nd edição revista*. Alcoitao: Escola Superior de Saúde do Alcoitao.
- Tagliamonte, S. A., and Baayen, R. H. (2012). Models, forests, and tress of York English: was/ were variation as a case study for statistical practice. *Lang. Var. Change* 24, 135–178. doi: 10.1017/S0954394512000129
- Tamburelli, M., and Jones, G. (2013). Investigating the relationship between nonword repetition and syllable structure in typical and atypical language development. *J. Speech Lang. Hear. Res.* 56, 708–720. doi: 10.1044/1092-4388(2012/11-0171)
- Theodorou, E., Kambanaros, M., and Grohmann, K. K. (2017). Sentence repetition as a tool for screening morphosyntactic abilities of bilingual children with SLI. *Front. Psychol.* 8:2104. doi: 10.3389/fpsyg.2017.02104
- Thordardottir, E. (2012). *Proposed Diagnostic Procedures and Criteria for Bi-SLI COST Action Studies*. Available at: <http://www.bi-sli.org>
- Thordardottir, E. (2015). “Proposed diagnostic procedures for use in bilingual and cross-linguistic contexts,” in *Assessing Multilingual Children: Disentangling Bilingualism From Language Impairment*, eds S. Armon-Lotem, J. de Jong, and N. Meir (Bristol: Multilingual Matters), 331–358. doi: 10.21832/9781783093137-014
- Thordardottir, E., and Brandeker, M. (2013). The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. *J. Commun. Disord.* 46, 1–16. doi: 10.1016/j.jcomdis.2012.08.002
- Tibshirani, R., Guenther, W., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 63, 411–423. doi: 10.1111/1467-9868.00293
- Tomblin, B. J., Records, N. L., and Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *J. Speech Lang. Hear. Res.* 39, 1284–1294. doi: 10.1044/jshr.3906.1284
- Topbaş, S., and Güven, S. (2013). *Turkce Erken Dil Gelisim Testi (TEDYL) (Test of Early Language Development: TELD-3, Turkish Version)*. Ankara: Detay Yayıncılık.

- Tsimpli, I. M., and Dimitrakopoulou, M. (2007). The interpretability hypothesis: evidence from wh- interrogatives in second language acquisition. *Second Lang. Res.* 23, 215–242. doi: 10.1177/0267658307076546
- Tuller, L. (2015). “Clinical use of parental questionnaires in multilingual contexts,” in *Assessing Multilingual Children: Disentangling Bilingualism From Language Impairment*, eds S. Armon-Lotem, J. de Jong, and N. Meir (Bristol: Multilingual Matters), 229–328.
- Tuller, L., Abboud, L., Ferré, S., Fleckstein, A., Prévost, P., dos Santos, C., et al. (2015). “Specific language impairment and bilingualism: assembling the pieces,” in *Language Acquisition and Development. Proceedings of GALA 2013*, eds C. Hamann and E. Ruigendijk (Newcastle: Cambridge Scholars Press), 533–567.
- Tuller, L., Hamann, C., Chilla, S., Ferré, S., Morin, E., Prévost, P., et al. (2018). Identifying language impairment in bilingual children in France and in Germany. *Int. J. Lang. Commun. Disord.* 53, 888–904. doi: 10.1111/1460-6984.12397
- Unsworth, S. (2016). “Amount of exposure as a proxy for dominance in bilingual language acquisition,” in *Language Dominance in Bilinguals*, eds C. Silva-Corvalán and J. Treffers-Daller (Cambridge: Cambridge University Press), 156–173.
- Unsworth, S., Chondrogianni, V., and Skzarabela, B. (2018). Experiential measures can be used as a proxy for language dominance in bilingual language acquisition research. *Front. Psychol.* 9:1809. doi: 10.3389/fpsyg.2018.01809
- van der Lely, H. (1998). SLI in children: movement, economy, and deficits in the computational-syntactic system. *Lang. Acquis.* 7, 161–192. doi: 10.1207/s15327817la0702-4_4
- van der Lely, H. (2005). Domain-specific cognitive systems: insight from grammatical specific language impairment. *Trends Cogn. Sci.* 9, 53–59. doi: 10.1016/j.tics.2004.12.002
- Vinther, T. (2002). Elicited imitation: a brief overview. *Int. J. Appl. Linguist.* 12, 54–73. doi: 10.1111/1473-4192.00024
- Windsor, J., Kohnert, K., Lobitz, K. F., and Pham, G. T. (2010). Cross-language nonword repetition by bilingual and monolingual children. *Am. J. Speech Lang. Pathol.* 19, 298–310. doi: 10.1044/1058-0360(2010/09-0064)
- Yip, V., and Matthews, S. (2006). Assessing language dominance in bilingual acquisition: a case for mean length utterance differentials. *Lang. Assess. Q.* 3, 97–116. doi: 10.1207/s15434311laq0302_2
- Zebib, R., Henri, G., Khomsi, A., Messara, C., and Hreich, E. (2017). *Batterie d’Evaluation du Langage Oral chez l’enfant libanais (ELO-L)*. Kerserwan: LTE.
- Zurer-Pearson, B. (2010). We can no longer afford a monolingual norm. *Appl. Psycholinguist.* 31, 339–343. doi: 10.1017/S014271640999052X

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer AG and handling Editor declared their shared affiliation.

Copyright © 2019 Abed Ibrahim and Fekete. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

TABLE A1 | Correlations between background variables for BiTD and BiSLI groups collapsed.

			Age	Positive_Earl_Dvt	AcO	SES	LDI	LoE
Spearman-Rhc	Age	Corr. coeff.	1.000	0.032	0.207	0.237	0.086	0.445**
		sig. (2-tailed)		0.816	0.126	0.078	0.514	0.001
		N	56	56	56	56	56	56
	Positive_Early_Dvt	Corr. coeff.	0.032	1.000	0.103	0.064	-0.057	-0.146
		sig. (2-tailed)	0.816		0.450	0.639	0.676	0.282
		N	56	56	56	56	56	56
	AoO	Corr. coeff.	0.207	0.103	1.000	-0.051	-0.351**	-0.670***
		sig. (2-tailed)	0.126	0.450		0.711	0.008	0.000
		N	56	56	56	56	56	56
	SES	Corr. coeff.	0.237	0.064	-0.051	1.000	0.269*	0.231
		sig. (2-tailed)	0.078	0.638	0.711		0.045	0.087
		N	56	56	56	56	56	56
	LDI	Corr. coeff.	0.089	-0.057	-0.351**	0.269*	1.000	0.428**
		sig. (2-tailed)	0.514	0.679	0.008	0.045		0.001
		N	56	56	56	56	56	56
	LoE	Corr. coeff.	0.44**	-0.146	-0.670***	0.231	0.428**	1.000
		sig. (2-tailed)	0.001	0.282	0.000	0.087	0.001	
		N	56	56	56	56	56	56

*The correlation is significant at the level of 0.05 (2-tailed).

**The correlation is significant at the level of 0.01 (2-tailed).

***The correlation is significant at the level of 0.001 (2-tailed).