Check for updates

Methods for Identifying Specific Language Impairment in Bilingual Populations in Germany

Cornelia Hamann* and Lina Abed Ibrahim

Department of English, University of Oldenburg, Oldenburg, Germany

This study investigates the performance of 22 monolingual and 54 bilingual children with and without specific language impairment (SLI), in a non-word repetition task (NWRT) and a sentence repetition task (SRT). Both tasks were constructed according to the principles for LITMUS tools (Language Impairment Testing in Multilingual Settings) developed within COST Action IS0804 and incorporated phonological or syntactic structures that are linguistically complex and have been shown to be difficult for children with SLI across languages. For phonology these are in particular (non)words containing consonant clusters. In morphosyntax, complexity has been attributed to factors such as embedding and/or syntactic movement. Tasks focusing on such structures are expected to identify SLI in bilinguals across language combinations. This is notoriously difficult because structures that are problematic for typically developing bilinguals (BiTDs) and monolingual children with SLI (MoSLI) often overlap. We show that the NWRT and the SRT are reliable tools for identification of SLI in bilingual contexts. However, interpretation of the performance of bilingual children depends on background information as provided by parental questionnaires. To evaluate the accuracy of our tasks, we recruited children in ordinary kindergartens or schools and in speech language therapy centers and verified their status with a battery of standardized language tests, assessing bilingual children in both their languages. We consider a bilingual child language impaired if she shows impairments in two language domains in both her languages. For assessment, we used tests normed for monolinguals (with one exception) and adjusted the norms for bilingualism and for language dominance. This procedure established the following groups: 10 typical monolinguals (MoTD), 12 MoSLI, 46 BiTD, and 8 bilingual children with SLI (BiSLI). Our results show that both tasks target relevant structures: monolingual children are classified with 100% accuracy. Crucially, both our tasks distinguish BiTDs from MoSLIs and BiTDs from BiSLIs. The NWRT shows high accuracy and only minimal influence of language dominance. The SRT can be scored as "identical repetition" or as "target structure," the latter aiming for scoring the mastery of a syntactic structure, ignoring lexical and specific case or gender errors. Focusing on the latter measure, we examine individual cases of BiTDs with unexpected, low scores. We identify first-language dominance as a factor influencing performance but crucially find that testing in the home language in a heritage context might lead to unreliable classifications and that our procedure for determining the clinical group of bilinguals missed cases of selective impairments such as syntactic SLI.

Keywords: bilingualism, specific language impairment, sentence repetition, non-word repetition, linguistic complexity

INTRODUCTION

Bilingual Language Development and Language Impairment

Recent linguistic research on (specific) language impairment (SLI) has focused on bilingual populations because more and more children grow up bilingually and the challenges of identifying what is typical in bilingual language development and what should be considered an impairment are notorious, see Armon-Lotem et al. (2015) and Marinis et al. (2017) for recent overviews. One such challenge is the finding that SLI may have different manifestations in different languages so that clinical markers widely differ. Extended use of infinitives has been described as a marker of SLI for English (Rice and Wexler, 1996), omission of object clitics for French (Jakubowicz et al., 1998; Paradis et al., 2003) and problems with subject–verb agreement (SVA) together with the use of infinitives and errors in verb placement for German (Clahsen, 1991; Hamann et al., 1998), to mention only some results from well-studied languages. The bigger challenge is, however, that there is an overlap in the linguistic structures that are difficult to master for bilingual children with those structures that are considered clinical markers for SLI in a particular target language; Håkansson and Nettelbladt (1996) were the first to point this out for Swedish, Paradis (2010), Hamann (2012), and Grimm and Schulz (2014) give more recent overviews of similarities and differences. This overlap in error patterns leads to over- and underdiagnosis, see Genesee et al. (2004).

Underdiagnosis occurs if difficulties are ignored based on the argument that delays or deficits in one or both languages often occur in bilingual development, as is the case for bilingual lexical development (Cobo-Lewis et al., 2002; Goldberg et al., 2008; Thordardottir, 2011), the bilingual acquisition of case in German (Schönenberger et al., 2012), or of grammatical gender in Dutch (Cornips and Hulk, 2008). See also Paradis et al. (2016) for a description of long lasting delays in bilingual language development. If, however, such difficulties are taken as evidence for language impairment, overdiagnosis is particularly likely when monolingual norms are applied in tests of the majority language, which might be the weaker language for a child at the time of assessment. Since SLI should be manifest in both languages of a bilingual child, the overlap problem can arguably be avoided if a child's language abilities are assessed in both her languages, the majority language (second language, L2) and the home language (first language, L1). The home language, L1, when spoken most of the time to and by the child in various communicative situations and with various speakers, will be the dominant language before the child is systematically exposed to the L2 in kindergarten or school. This situation often holds for simultaneous, but also for early sequential bilingual children. Even though it has been recommended (Fredman, 2006) that a child be tested in both or, at least, in the dominant language, testing a child in her L1 is often not practicable: there might be no normed tests available for the L1 or the speech language therapist (SLT) cannot administer or evaluate the test in this particular language. In the case of simultaneous bilingual children, it also has to be taken into account that the home language is often a heritage language, i.e., the parents are second or third generation immigrants and speakers of the

language. Heritage situations add further complications: L1 tests, if available, might not be appropriate because the immigrant language might have changed due to contact phenomena as in the case of Immigrant Turkish in Germany (Schroeder and Dollnick, 2013), or, independent of the L1, early acquisition of an L2 might lead to attrition phenomena (Köpke et al., 2004; Montrul, 2008).

The diversity of bilingual profiles and the subtypes of SLI discussed in the literature (Leonard, 1998, 2014) also contribute to the diagnostic difficulties. Bilingual development is crucially influenced by age of onset (AoO), which leads to the definition of simultaneous (AoO $\leq$ 3) bilingualism, early (3 < AoO < 4) and late (AoO $\geq$ 4) sequential child bilingualism (also called child L2), as well as to a clear distinction of child and adult L2 speakers, see Meisel (2009) for a discussion of early and late child L2.[1] Length of exposure (LoE), quantity and quality of input, and socioeconomic status (SES) also contribute crucially to bilingual language development so that background information about these factors is essential for the assessment of language samples and the interpretation of test results. Though SLI frequently concerns both phonological and morphosyntactic development, selective impairments have been identified, such as grammatical/syntactic (van der Lely, 1998) or semantic SLI (Schulz and Roeper, 2011), see also Friedmann and Novogrodsky (2011). The diversity of subtypes of SLI contributes to the problems in identifying language impairment in bilingual children.

The Language Impairment Testing in Multilingual Settings (LITMUS) Tools for Crosslinguistic Research

Given these difficulties, several approaches can be explored. First, existing assessment tools can be normed for bilingual populations. Second, existing tools normed for monolinguals can be applied adjusting the norms for bilingualism and according to the status of the language being tested as the dominant or weaker language, see the recommendations by Thordardottir (2015) described in Section "Participants and Procedure for Verification of Clinical Status." Third, new tools can be constructed according to linguistic principles that allow crosslinguistic application, such as the tools developed during and following COST Action IS0804. These are called LITMUS tools and are described in Armon-Lotem et al. (2015). Of specific interest here are the LITMUS principles outlined in Chiat (2015) for non-word repetition tasks (NWRTs) and by Marinis and Armon-Lotem (2015) for sentence repetition tasks (SRTs) and the Questionnaire for Parents of Bilingual Children (PaBiQ) described by Tuller (2015). These three tasks were central in a French–German joint project (BiLaD – bilingual language development)[2] investigating monolingual and bilingual children with and without language impairment and with Arabic,

---

[1]Note that authors often use their own definitions, e.g., Schulz and Tracy (2011) define children with an AoO < 24 months as simultaneous bilinguals.

[2]The project was funded by DFG (German Science Foundation) grants HA 2335/6-1, RO 923/3-1, CH 1112/2-1 to Cornelia Hamann, Monika Rothweiler, and Solveig Chilla as well as an ANR (French Science Agency) to Laurice Tuller as principal investigator.

Portuguese and Turkish as home languages,[3] of which we report the German data here.

We focus on non-word repetition and sentence repetition since such tasks have been shown to reliably identify SLI in monolinguals (Conti-Ramsden et al., 2001) and are often part of standard assessment tools. Such tests usually assess working memory (WM), see Archibald and Gathercole (2006) but can be constructed so that they measure the command of phonological or syntactic representations/derivations (see Gallon et al., 2007 for non-word repetition; Polišenská et al., 2014 for sentence repetition). In SRTs, this can be achieved by taxing memory with number of words and vocabulary so that a successful parse of the sentence is a necessary condition for successful repetition. In addition, structurally minimal pairs should be incorporated to identify the locus of difficulty: in the case of embedding, a finite complement clause can be contrasted with a coordination structure, which also contains two propositions but does not embed one into the other. The LITMUS tasks incorporate linguistically complex (syntactic or phonological) structures and operations known to be difficult for children with SLI crosslinguistically or in a particular language, such as SVA or topicalization in German. For syntax, especially structures involving syntactic movement, particularly Wh-movement, i.e., fronted interrogative or relative pronouns (see Hamann et al., 1998; van der Lely, 1998; Friedmann and Novogrodsky, 2011), as well as embedding (Hamann and Tuller, 2014) have been crosslinguistically identified by recent research as vulnerable in children with SLI. A particular difficulty has been identified for structures that involve movement and contain intervening elements between the source of the moved element and its landing site (Rizzi, 2004; Friedmann et al., 2015). The latter difficulty occurs in object Which-questions and in object relative clauses containing a lexical subject. In contrast to the difficulties encountered by children with SLI, a typically developing bilingual child might have problems with vocabulary or grammatical features that do not have semantic content (uninterpretable linguistic features, such as number agreement on the verb, Tsimpli and Dimitrakopoulou, 2007) and might even avoid complexity, but should in principle not be overtaxed by structures involving movement or embedding. Recent results indicate that SRTs incorporating structures involving these operations can be successfully applied in bilingual settings for identifying SLI, see Marinis and Armon-Lotem (2015), Tuller et al. (2015), and Fleckstein et al. (2016). As to non-word repetition and phonological complexity, recent studies show that syllables containing branching onsets or a coda are particularly difficult for children with SLI, but are mastered by typical bilinguals (Marshall and van der Lely, 2009; Ferré et al., 2012; dos Santos and Ferré, 2016; Grimm and Hübner, in press). NWRTs can be constructed to incorporate quasi-universal non-words or non-words conforming to phonotactic and/or morphophonological constraints of a specific language. Especially the quasi-universal type can be used successfully with bilingual children after only a short time

of exposure to the target language, independent of SES and L2 experience, see Chiat and Polišenská (2016). Thordardottir and Brandeker (2013) compared performance of bilingual children in an NWRT and an SRT to performance on receptive vocabulary and found the latter more affected by levels of previous exposure than NWRT and SRT, with NWRT and SRT showing acceptable sensitivity levels. Quite recently, LITMUS NWRTs and SRTs have been studied as to their diagnostic accuracy in bilingual populations. Boerma et al. (2015) use a quasi-universal LITMUS NWRT and report excellent accuracy for their population of bilingual children with Dutch as L2. Armon-Lotem and Meir (2016) find good accuracy for their Hebrew LITMUS SRT in Russian–Hebrew bilingual children, whereas the accuracy for their NWRT, with word-like items incorporated, is described as fair. The arguably good diagnostic accuracy of NWRT and SRT in monolingual and bilingual populations (but see also Gutiérrez-Clellen and Simon-Cereidjido, 2010) led us to develop and investigate an SRT for German and to adopt the NWRT developed by Grimm et al. (2014) and investigate it with our bilingual population.

Research Questions and Aims of the Present Study

This study presents data from 54 bilingual children living in Germany with Arabic, Portuguese and Turkish as their home language, comparing them to 22 monolingual children. The overall aim of the study is to investigate two new LITMUS tools for German, a sentence repetition and a non-word repetition task developed according to the LITMUS principles (COST Action IS0804, Chiat, 2015; Marinis and Armon-Lotem, 2015). We want to know in particular whether they are able to identify SLI in bilinguals. For the NWRT, we want to know how accurate it is for our population, and we specifically investigate the German SRT as a new method and discuss its evaluation by different scoring procedures. As a first step, we therefore investigate the performance of monolingual children with and without SLI on these tasks. For evaluating the accuracy of the new tasks in bilingual children, groups of typically developing bilingual children and of bilingual children with SLI were defined. For this goal, mono- and bilingual children without any history of language problems were recruited in ordinary kindergartens and schools and children with a diagnosis of SLI (mono- and bilingual) were recruited in speech language centers or private practice. This initial grouping was verified and if necessary corrected by using norm-referenced L1 and L2 tests adjusting the norms as suggested by Thordardottir (2015) and described in more detail in Section "Participants and Procedure for Verification of Clinical Status." This procedure, as pointed out by Thordardottir (2015), is not unproblematic and will be discussed with respect to the status of the home languages as heritage languages and the different subgroups of SLI. We will then proceed to show that tests in the L2 can be very reliable, especially the LITMUS tasks. It will also emerge, however, that in most cases a combination of tests should be applied to achieve good diagnostic accuracy.

---

[3]See Fleckstein et al. (2016), Almeida et al. (2017), and dos Santos and Ferré (2016) for results on the French versions of the NWRT and the SRT.

METHODS AND PROCEDURES

Participants and Procedure for Verification of Clinical Status

We investigated bilingual children with Arabic, Portuguese and Turkish as home languages. These languages were chosen because there are substantial groups of Arabic, Portuguese and Turkish immigrants in Germany[4] and because the language communities differ from each other, so that comparisons can be made. Children were recruited in kindergartens, schools and in speech language therapy centers. The study was carried out in accordance 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). Written informed consent was obtained from all adult research participants as well as from the parents/legal guardians of all minors. Written informed consent was obtained from the parents both for the purposes of data collection through the Parental Questionnaire as well as for the purposes of their children's participation in this research. The protocol was approved by the "Kommission für Forschungsfolgenabschätzung und Ethik" of the Carl-von-Ossietzky University of Oldenburg.

The age range of the children was chosen as 5;5–9;4 years since this includes the last year of kindergarten and the crucial first 2 or 3 years in primary school. We recruited 22 monolingual children, 10 typically developing and 12 with a diagnosis of SLI. In addition, 38 typically developing bilingual children were selected in Germany as well as 16 bilingual children in SLT, see (**Table 1**). We included only bilingual children with an LoE of more than 24 months. Our group includes simultaneous and sequential bilinguals, where we define the latter as children who were systematically exposed to their L2 at the age of 36 months or later.

The status of all of these children as typical or language impaired was then verified by a battery of tools following part of the protocol suggested by Thordardottir (2012). We first tested for non-verbal cognition with the German version of Raven's colored progressive matrices (CPM), see Bulheller and Häcker (2002), excluding children who scored below percentile 9 (the cutoff for low-average

---

[4]This also holds for France, which makes cross-country comparisons possible in the project.

TABLE 1 | Bilingual children in Germany: children not in speech language therapy (SLT) and children in SLT.

	Not in SLT	In SLT
Bilingual: L1 Arabic	7	4
Bilingual: L1 Portuguese	19	1
Bilingual: L1 Turkish	12	11
Total no. participants	38	16
M age (SD)	82.5 (12.80)	84.68 (14.96)
Range	66–107	65–109
Gender	21 F, 17 M	6 F, 10 M

non-verbal intelligence, equivalent to an IQ-score ≤ 80 according to Wechsler's IQ scale). We also collected a narrative language sample in each of a child's languages. For the latter, we used the materials provided by the Multilingual Assessment Instrument for Narratives (MAIN), another LITMUS task perfected within Cost Action IS0804 (Gagarina et al., 2015), but did not evaluate the narratives according to the MAIN protocol. Instead, we used the material to (a) judge the expressive abilities of a child in each of her languages to confirm or disconfirm the status of a language as the weaker or the dominant language and (b) to scan the material for clinical markers of SLI such as SVA errors in German.[5] In our sample no child was excluded because of performance in CPM and all children in the bilingual groups had at least receptive command of two languages.

Following many researchers on SLI, see Leonard (2014), Tomblin et al. (1997), and also Thordardottir (2015), we classified a monolingual child as having SLI (MoSLI) whenever performance was below −1.25 SDs in two language domains in appropriate norm-referenced tests. Relevant language domains in this context are phonology (receptive and productive), receptive and productive vocabulary and comprehension and production of morphosyntax. For bilingual children, we followed Thordardottir (2015), who suggests the following norm adjustments for norm-referenced tests with monolingual norms: A bilingual child is considered SLI if she scores −1.5 SDs below mean scores of typical monolingual peers in her dominant language, −2.25 SD in her weaker language, and −1.75 SD in either language if she is a balanced bilingual. We are aware that these cutoffs were calculated for groups of simultaneous bilingual children.

We administered three norm-referenced L2 tests, the LiSe-DaZ, the WWT and the PLAKSS-II, covering morphosyntax, lexicon and phonology separately, and the ELO-L for Arabic, the PALPA-P and GOL-E for Portuguese, and the TEDIL for Turkish as L1 tests, see Section "Standardized L2 and L1 Tests" for details. The results were interpreted on the background information provided by the PaBiQ. In particular, the calculation of children's language dominance allowed the application of adjusted cutoffs. With the help of these adjustments for tests providing monolingual norms, we classified a bilingual child as language impaired only if the child performed below the respective cutoffs in two language domains in both of her languages. For the TEDIL, because it provides only two composite values, we used the suggested monolingual norm of −1.0 adjusting it according to dominance. For the LiSe-DaZ, which provides bilingual norms for sequential bilingual children (defined by the authors as AoO > 2) and also monolingual norms, we used a cutoff of −1.25 SD. Since expressive vocabulary is a notorious domain of difficulty for bilingual children in both languages, we decided to count the lexicon as a single domain and consider a bilingual child as typically developing in her lexicon if she scored above the appropriate cutoffs in receptive vocabulary. This leads to the classification of participants as shown in **Table 2**, which also shows our control groups, the monolingual children with and without SLI.

---

[5]Further evaluation of narrative micro- and macro-structure according to the MAIN protocol will be the next step.

TABLE 2 | Participants including monolingual children and final status of bilingual children as BiTD and BiSLI: age at testing (months), colored progressive matrices (CPM) scores (percentile ranks), and gender.

	MoTD ($n = 10$)	MoSLI ($n = 12$)	BiTD				BiSLI[a] ($n = 8$)
			BiTD-A ($n = 8$)	BiTD-P ($n = 19$)	BiTD-T ($n = 19$)	Total ($n = 46$)	
Simult./total						40/46	8/8
			Mean (SD) and range				
Age at testing	75.90 (8.99) 66–92	81.75 (13.41) 68–112	88.50 (12.11) 70–107	80.95 (13.51) 66–106	88.68 (14.32) 70–111	85.45 (13.87) 66–11	82.12 (16.4) 65–109
CPM	81.20 (13.98) 56–100	53.41 (23.39) 25–99	47.37 (30.91) 9–93	71.63 (20.59) 28–94	67.57 (22.39) 36–100	65.73 (24.39) 9–100	59.12 (26.58) 27–98
Gender	8 F, 2 M	5 F, 7 M	5 F, 3 M	9 F, 10 M	10 F, 9 M	25 F, 21 M	2 F, 6 M

[a]BiSLI group (n = 8): 3 L1 Arabic, 1 L1 Portuguese, and 4 L1 Turkish.

Comparing the initial groups from **Table 1** to the classification achieved by L1 and L2 testing in **Table 2**, it is striking that the bilingual population with language impairment has been cut in half. Our procedure, and testing in L1 in particular, has uncovered eight potential cases of overdiagnosis.

The four final groups (MoTD, MoSLI, BiTD, and BiSLI) were comparable concerning non-language variables such as age, non-verbal intelligence, and SES (see **Table 2**).[6] A Kruskal–Wallis non-parametric test[7] revealed no significant differences in terms of age at testing between the four groups of participants [$\chi^2(3, N = 76) = 4.061$, $p = 0.255$]. The age difference remains statistically insignificant even when the BiTD group is split by the children's home language into three subgroups (BiTD-A, BiTD-P, and BiTD-T) [$\chi^2(5, N = 76) = 7.782$ $p = 0.169$]. Although the Kruskal–Wallis test revealed a marginally significant difference with respect to the four groups' non-verbal intelligence [$\chi^2(3, N = 76) = 7.689$, $p = 0.053$], *post hoc* Mann–Whitney U test applying Bonferroni correction revealed only one significant comparison between the MoTD and MoSLI group ($U = 154$, $p = 0.036$, $r = 0.348$). Nevertheless, all of the children in the MoSLI group have normal non-verbal intelligence. We further checked whether the L1 Arabic, L1 Portuguese, and L1 Turkish typically developing children were comparable for SES as measured by years of mother's education. Since no significant differences were observed [$\chi^2(2, N = 46) = 0.181$, $p = 0.913$], the three subgroups were collapsed into one BiTD group. A Kruskal–Wallis test also revealed that the BiTD and BiSLI groups were similar with respect to SES.

Standardized L2 and L1 Tests

For L1 and L2 assessment, we chose standardized tests in both languages that are commonly used in speech language therapy and are normed for the age range investigated here—or for which norms can be extended, see **Table 3** for an overview. An important decision for assessment in German was made in the choice of the LiSe-DaZ (Schulz and Tracy, 2011), which is the first German standardized test normed not only for

monolinguals but also for sequential bilingual children between 3;0–7;11. Comprehension of negation, of constituent questions, and of telic events is tested. The assessment of production targets SVA, sentence complexity, case marking, and word classes (prepositions, main verbs, auxiliaries, focus particles, and conjunctions). All subtasks except those for sentence complexity and SVA provide t values. The recommendation of the authors is to consider a child "at risk for language impairment if she performs more than 1 SD below $t = 50$ in two of the 9 subtests with t values" (Grimm and Schulz, 2014, p. 831). This procedure excludes an area of morphosyntax, SVA, which has been discussed as clinical marker for (bilingual) SLI in German (Rothweiler et al., 2012), and does not allow separate evaluation of performance in production and comprehension. We departed from the authors' own rating procedure by (a) setting the cutoff at −1.25 SD and (b) ignoring the results of the case task (see Lein et al., 2016; Abed Ibrahim et al., in press). The test does not offer norms for simultaneous bilingual children with an AoO < 24 months or bilingual children older than 8 years. For older children, however, a cutoff of −0.5 SD is suggested by the authors, and for simultaneous bilinguals monolingual norms can be applied whenever German is the dominant language. Since the LiSe-DaZ is an assessment of comprehension and production of morphosyntax only, other domains of language had to be evaluated with separate tests. We chose the WWT (Glück, 2007) for evaluation of lexical reception and production and the PLAKSS-II (Fox-Boyer, 2014) for evaluation of phonology. For classifying a child as BiSLI it was necessary that she performs below adjusted cutoffs in two domains of L1 and two domains of L2. For the L2 tests, this implies that she had to perform below cutoffs in two subtasks of the LiSe-DaZ (morphosyntax) combined with low performance in either the PLAKSS-II (phonology) or the receptive subtest of the WWT (vocabulary), or she had to perform below cutoffs in the receptive part of the WWT and in the PLAKSS-II.

Turning to the three different L1s, we chose the ELO-L for Arabic (Zebib et al., 2017). It uses word repetition for phonological abilities, picture naming and picture selection for lexical production and reception, sentence completion and picture selection for assessing morphosyntax. It exists in two versions, for younger (3;0–5;11) and older (6;0–7;11) children, is normed for both versions on a large and mixed population, and takes 30–45 min to

---

[6]Information on SES is only available for bilingual children in our data set.
[7]See Section "Data Analysis" for the choice of statistical tests, taking account of the unequal group sizes.

TABLE 3 | Standardized tests used for language assessment in Arabic, German, Portuguese, and Turkish: overview.

Language	Name of test	Language skill assessed					Scoring method	Norming population
		Phonology	Receptive vocabulary	Expressive vocabulary	Morphosyntax comprehension	Morphosyntax production		
Arabic	ELO-L[a]	Word repetition	Picture selection	Picture naming	Picture–sentence matching	Sentence completion	Individual subtest scores and global score	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 Bilinguals: 3;0–7;11
	PLAKSS-II[d]	Picture naming	–	–	–	–	Individual subtest scores	2;6–7;11
EP	PALPA-P[e]	Non-word repetition	Picture selection	Picture naming	Picture selection	Sentence repetition	Individual subtest scores	5;0–9;11 (with some missing norms for all tasks)
	GOL-E[f]	–	Word definition	Antonyms naming	–	Sentence construction	Individual subtest scores and global score	5;07–10;00
Turkish	TEDIL[g]	–	Picture selection	Picture naming	Picture selection	Sentence completion	2 Composite scores, 1 production and 1 comprehension	2;0–7;11

[a]*Zebib et al. (2017).*
[b]*Glück (2007).*
[c]*Schulz and Tracy (2011).*
[d]*Fox-Boyer (2014).*
[e]*Castro et al. (2007).*
[f]*Sua-Kay and Santos (2014).*
[g]*Topbaş and Güven (2013).*
EP, European Portuguese; TVJT, truth value judgment task.

administer. The test takes into account the bilingual situation in Lebanon and was translated by native speakers to other varieties of Arabic such as Algerian, Egyptian, Moroccan, Tunisian, Libyan, Palestinian, and recently Syrian.

The PALPA-P (Provas de Avaliação da Linguagem e da Afasia em Português) was adapted by Castro et al. (2007) from the *Psycholinguistic Assessments of Language Processing in Aphasia* by Kay et al. (1996) and provides a linguistically well-controlled instrument for the assessment of children with European Portuguese as L1. The test evaluates the domains of phonology, lexical production and reception as well as morphosyntactic production and comprehension. It is normed for children aged 5;0–9;0 (with certain gaps, especially in the lexical evaluation) and takes about 50 min to administer. Scoring is correct (1) or incorrect (0). Since there are age gaps in the norming population for the lexical tasks in the PALPA-P, we used the GOL-E (Sua-Kay and Santos, 2014) for lexical production and comprehension with norms for children between 5;7 and 10;0 years of age.

For Turkish, we chose the TEDIL by Topbaş and Güven (2013), an adaptation of the TELD-3, which has been normed for children aged 2;0–7;11. It exists in two different versions for younger and older children, and measures comprehension and production in morphosyntax, morphology and lexical semantics. The task does not specifically test for phonology, but has a subtask for lexical reception and two further receptive tasks on lexical relations. For morphosyntax, there is a comprehension and a production part in the form of a repetition task. Norms exist for composite scores of reception and expression only, not for individual subscores.

The LITMUS-PaBiQ

An important assessment tool for the evaluation of language abilities in bilingual children is a questionnaire that can provide the background for the interpretation of test results. Information about the child's language exposure and use, current and in her early years of development, is essential and allows determination of language dominance. For this purpose, we chose the Questionnaire for Parents of Bilingual Children (PaBiQ; Tuller, 2015), which was developed within COST Action IS0804 based on questionnaires developed in Paradis et al. (2010) and Paradis (2011). We used a German translation of the questionnaire as well as translations into Arabic, Portuguese, or Turkish so that parents could choose in which language the interview, by phone or in person, would be conducted.

Parental questionnaires, and the PaBiQ in particular, pay special attention to age of first systematic language exposure (AoO), LoE, quality and quantity of input at home, and other everyday situations and also provide information about parents' education, which can be taken as an indication of SES. Apart from these variables known to impact bilingual development, indicators for language impairment were also incorporated into the questionnaire. These include early language development (first words and first sentences) and family history of language difficulties. The

latter variables allow calculating a No-Risk Index, a reliable indicator for the French group of children investigated in the BiLaD project (Almeida et al., 2017) and currently under investigation for the whole group and the German bilinguals in particular.

Returning to the factors influencing bilingual development, they allowed us to determine an L2 Exposure Index and an L1 Exposure Index. These indexes were calculated by weighing factors such as AoO, LoE, language use, and richness at home, at school, in extracurricular activities, before and after the age of 4 years. The Language Dominance Index (LDI) can be calculated as the difference between the L2 and the L1 exposure indexes. Given the individual contributions of the factors in L1 and L2, the LDI ranges from −50 to +50. For the project, several cutoff points were explored and compared with impressions of bilingual investigators, specifically taking into account free conversation and the samples of spontaneous speech collected for each child in each language (see also Almeida et al., 2017). Following that procedure, we define bilingual children in Germany as balanced if they score between the values of −5 and +5 of the LDI ($-5 \leq LDI \leq +5$). Children with an LDI below −5 are considered L1 dominant whereas children with an LDI above +5 are classified as L2 dominant.

The New German LITMUS Repetition Tasks

The German LITMUS NWRT

Since the goal is to not disadvantage bilingual children when assessing their phonological abilities, the NWRT (see Grimm and Hübner, in press) was designed to include vowels and consonants common in most languages of the world, at the same time targeting complex phonological structures, i.e., consonant clusters, known to cause difficulty in children with SLI, see Chiat (2015) and Ferré et al. (2012). In particular, the NWRT contained a language-independent (LI) part and a language-dependent (LD), see Grimm and Hübner (in press)[8] and Abed Ibrahim and Hamann (2017) for a detailed description of the task. There were maximally three syllables in the non-words so that memory effects would only minimally influence performance. The 30 non-words of the LI part were built using phonemes and phonotactic properties well attested crosslinguistically (Maddieson et al., 2011). Differing from the universal NWRT discussed in Chiat and Polišenská (2016), the task does not only contain simple CV syllables but also syllables with branching onsets of the type "CCV" and a final consonant coda (coda, CVC#), which are nonetheless characterized by their crosslinguistic frequency (Maddieson, 2006). We expect monolingual and bilingual children with SLI to have difficulties with these phonologically complex structures whereas typical monolingual and bilingual children should master them. The LD part contains 36 non-words with two more additional German consonants /s, ʃ/ and more syllable types as shown in **Table 4**. Since sC# and #Cs sequences are not unique to German but violate the Sonority Sequencing Principle, they are difficult for

TABLE 4 | Overview of the German LITMUS non-word repetition task.

	Vowels	Consonants	Syllable types	Examples
Language-independent part 30 items 23 test items 7 control items (e.g., pilu, paf)	/a, i, u/	/p, k, f, l/	CV CCV CVC#	pilu flaplu kafip
Language-dependent part 36 items 32 test items 4 control items (e.g., kas, ʃaku)	/a, i, u/	/p, k, f, l/ plus /s/ /ʃ/	same syllable types plus #sCV #sCCV Cs#	skifapu ʃplaklu fikapuks
			internal /s/	Kufiski

children with SLI (dos Santos and Ferré, 2016) but should not be problematic for typically developing children.

The task, in the form of a PowerPoint presentation (PPT), is easy to administer and takes about 5–10 min. It is appealing to children since they are told that it is an alien who is trying to teach them his language. Items were presented in pseudo-randomized order through headphones. Scoring took into account whole item accuracy, disregarding systematic substitutions, e.g., /t/ for /k/, as well as errors in minimally different vowels or voicing of consonants. Following Grimm and Hübner (in press), we also disregarded substitution of extrametrical /ʃ/ by [s] since their substitution does not lead to a phonemic contrast in syllable initial position in German.

The German LITMUS SRT

The German SRT, first introduced by Hamann et al. (2013), was constructed in parallel to the French task (Fleckstein et al., 2016; Almeida et al., 2017) during COST Action IS0408 incorporating the LITMUS principles (Marinis and Armon-Lotem, 2015). It thus contains complex structures known to be difficult for children with SLI crosslinguistically, including object questions, subject and object relative clauses, finite complement clauses and passives, as well as structures identified as milestones in the acquisition of German word-order properties such as topicalization, and the sentence bracket, see examples (5) and (1). See Hamann et al. (2017) and Lein et al. (2016) for details on the German SRT and Hamann (2015) for an overview of SLI in German.

The version[9] of the German LITMUS SRT investigated in this study contains 45 sentences with three levels of increasing complexity controlled for number of syllables in each level (five conditions per level and three items per condition). Stimuli are presented in randomized order *via* a child friendly PPT. The levels arise through adding factors of complexity such as Wh-movement, embedding, intervention and the fact that two propositions are presented. Thus level 1 contains simple declaratives and assesses SVA, tense and the sentence bracket, (1). Level 2 includes object

---

[8]We particularly thank Angela Grimm for sharing the task with us.

[9]The original long version of the German LITMUS-SRT was shortened to meet the needs of the age range investigated in the BiLaD project.

questions with an intervening lexical NP subject. Following Rizzi (2004), these are Which-NP questions, where the interrogative constituent contains a lexical NP as restriction, which has moved over a lexical subject as in "Welchen *Clown* umarmt der *Wikinger* <welchen *Clown*>—which *clown* does the *viking* hug <which *clown*>." These are contrasted with questions where the question constituent does not carry a lexical restriction (*wen–whom*, bare Wh) and therefore there is no intervention. All questions ask for masculine persons with unambiguous case marking, see (2a) and (2b). The task also contains finite (3), and non-finite complement clauses contrasting with coordinate structures. Level 3 contains long passives, subject relatives, object relatives with, (4), and without a lexical intervener, as well as topicalizations (5). **Table 5** gives a summary.

(1)   Sentence bracket:

Die	Köchin	**hat**	den	Cowboy	**geweckt**
The/*nom.*	cook	has	the/*acc.*	cowboy	woken up

"The cook woke the cowboy up"

(2a)   Bare WH

Wen	umarmt der			Pinguin	heute?
Who/*acc.*	hugs	the/*nom.*		penguin	today?

"Whom does the penguin hug today?"

(2b)   Which-NP

Welchen	**Clown**	besucht	der Zauberer?
Which/*acc.*	Clown	visits	the/*nom.* Magician?

"Which clown does the magician visit?"

(3)   Finite complement clause:

Der	Prinz	will,	**dass**	der	Ritter	die	Affen	**jagt**
The/*nom.*	Prince	wants,	that	the/*nom.*	knight	the/*acc.*	monkeys	hunts

"the prince wants that the knight hunts the monkeys"

(4) Object relative with intervention:

Ich	sehe	den	Clown,	**den**	**der**	**Wikinger**	umarmt
I	see	the/*acc.*	Clown	who/*acc.*	the/*nom.*	viking	hugs

"I see the clown who(m) the viking hugs"

TABLE 5 | Overview of the German LITMUS sentence repetition task.

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.)

(5)   Topicalization

Den	**Koch**	besucht	der	Zauberer	als ersten
The/*acc.*	cook	visits	the/*nom.*	magician	first

"The cook, the magician visits first."

The task takes about 10 min to administer. Items are scored as 0/1 using different criteria for this rating. "Identical repetition" only disregards phonological errors and is the fastest and easiest way of scoring. Since lexical substitutions and omissions are counted as errors in this scoring method, difficulties that bilingual children have with vocabulary will clearly show in this measure. An alternative method is "target structure" which aims to ascertain that a child masters certain complex structures in principle. It compensates for L2 errors such as lexical substitutions and systematic recurrent case errors as well as gender errors that do not affect the realization of the targeted structure see the examples in (6) to (8). Errors not affecting the realization of the target structure in the examples are given in bold print:

(6)   Target structure: (sentence bracket)

Die	Köchin	hat	den	Cowboy	geweckt
The/*nom.*	cook	has	the/*acc.*	cowboy	woken up

"The cook woke the cowboy up"

Child repetition:

Die	Köchin	hat	den	**Clown**	geweckt
The/*nom.*	cook	has	the/*acc.*	**clown**	woken up

lexical substitution error, target structure score: 1

(7)   Target structure: (long passive)

Die	Oma	wird	von	dem	großen	Cowboy	geärgert
The/*nom.*	grandma	is	by	the/*dat.*	tall	cowboy	annoyed

"The grandma is annoyed by the tall cowboy"

Child repetition:

Die	Oma	wird	von	**den**	*großen*	Cowboy	geärgert
The/*nom.*	grandma	is	by	**the/acc**.	Tall	cowboy	annoyed

systematic case error in passives, *von dem* rendered das *von den*, target structure score:1

(8)   Target structure: (SVA, third, sg)

Der	kleine	Hund	bringt	die	Zeitung
The/*nom.*	small	dog	brings	the/*fem.-acc.*	newspaper

"The small dog brings the newspaper"

Child repetition:

Der	kleine	Hund	bringt	**den**	Zeitung.
The/*nom.*	small	dog	brings	**the/masc.-acc**.	newspaper

gender error, target structure score: 1

Using this method might miss measuring the total effect of linguistic complexity. Quite often, several errors occur in complex structures, not necessarily however on the specific marker of the structure itself. To give an example: When a finite complement clause is targeted, the structural difficulty might be manifest in an omission of the complementizer (*dass—that*) and a simple juxtaposition of clauses. This would clearly be 0 for scoring as "target structure." The difficulty could surface in lexical substitution, however, or there could be additional errors unrelated to the complementizer. Since "target structure" is a measure that does not penalize bilingual children and can establish whether structures such as finite complement clauses are acquired or not, we nevertheless use this measure for scoring German SRTs in addition to the measure of identical repetition.

Research Questions Concerning the German LITMUS Repetition Tasks

As stated in Section "Research Questions and Aims of the Present Study," we want to know whether the German LITMUS SRT and NWRT are able to identify language impairment in bilingual settings. For this purpose, we first ask whether the tasks successfully identify SLI in monolingual German children. We also want to know in how far our tasks can be used as a first evaluation, i.e., we calculate cutoffs and accuracy of the new LITMUS tasks based on the identification of our clinical population by the use of norm-adjusted L1 and L2 tests. In particular, we want to know if the German SRT with the score of "target structure" can successfully identify bilingual children with SLI.

Data Analysis

The children's NWRT and SRT responses were recorded with special audio recorders. They were transcribed offline, verified and scored by two independent linguistically trained research and student assistants.

IBM SPSS 22 (2013) was used for all statistical analyses. Due to unequal group sizes and since explorative statistics revealed a violation of the assumption of normality in our data set, non-parametric statistical tests were used for group comparisons throughout the study. To measure the diagnostic accuracy of the LITMUS NWRT and SRT, sensitivity (the proportion of children with SLI identified as such by the task) and specificity (proportion of children with typical language development identified as such by the task) were calculated for each task upon an established cutoff score. The optimal cutoff score on a test is the performance score yielding the highest specificity and sensitivity ratios. Sensitivity and/or specificity rates ≥90% are considered good, whereas rates between 80 and 89% are considered fair (Plante and Vance, 1994). In addition, likelihood ratios were calculated for the established sensitivity and specificity levels because they are less likely to be affected by variations in the sample's characteristics (see Dollaghan, 2004). A positive likelihood ratio (LR+) indicates the likelihood of scores below a cutoff criterion to occur in children with language impairment and is calculated as follows: LR+ = sensitivity/(1 − specificity). The negative likelihood ratio (LR−), on the other hand, indicates

the likelihood of a child performing above the cutoff point to be typically developing and is calculated with the following formula: LR− = (1 − sensitivity)/specificity. LR+ values ≥10 are considered to be clinically informative (highly indicative) of the presence of an impairment, and LR− values ≤0.10 are viewed as highly indicative of the absence of impairment. LRs+ ≥ 3.0 and LRs− ≤ 0.3 are viewed as "clinically suggestive," whereas LRs+ < 3.0 and LRs− > 0.3 are considered to be clinically uninformative (e.g., Dollaghan, 2007).

Receiver operating characteristic (ROC)[10] curve analysis (Dunn, 2011) is widely used to estimate the discriminatory power and optimal cutoff criterion of a task. The optimal cutoff point is the score associated with the highest diagnostic accuracy of a task and is generated by plotting "the true positive rate (sensitivity) against the false positive rate (1 − specificity)" (Gutiérrez-Clellen and Simon-Cereidjido, 2010). One of the important drawbacks of the ROC analysis is that it uses the dichotomous variable "clinical group membership" as dependent variable to predict sensitivity and specificity for different thresholds. Thus, sensitivity and specificity ratios obtained by this procedure could be influenced by how well the participants were assigned to the SLI and TD groups. Since the clinical status of the bilingual children was determined using norm-referenced L1 and L2 tests standardized on monolingual children with adapted bilingual cutoffs, one cannot fully rule out the possibility of false group assignment especially in cases of selective impairments. For the aforementioned reasons, ROC curve analysis was performed only for our monolingual data. In case of bilinguals, we opted for an alternative measure that does not rely on the assignment procedure. We use k-means cluster analysis, which is one of the simplest clustering algorithms, to partition data into k clusters (MacQueen, 1967). The k-means clustering algorithm attempts to show which cluster each observation belongs to. In our case, the algorithm classified our observations into two clusters using the test variables as dependent measures. Crucially, such clusters are extracted based on the mathematical characteristics of the data independently from clinical status in an unsupervised manner, that is, assigned clinical status is not taken into consideration in the clustering procedure.[11] We ran k-means cluster analyses on each of the LITMUS tasks separately entering just one dependent measure into the clustering procedure at each run.

Our premise was that the two clusters would cut across the clinical status, since our test variables (LITMUS NWRT and SRT) have been proposed to be sensitive to the presence or absence of language impairment, see Section "The Language Impairment Testing in Multilingual Settings (LITMUS) Tools for Crosslinguistic Research." The cutoff is a reference value ascertained after the cluster memberships are determined. Since we have uni-dimensional data (using just one variable per cluster analysis), the cutoff is on the same scale as the score of the dependent measure. The cutoff is an imaginary line separating the two clusters. It is calculated as the mean of the maximum score in

---

[10]A ROC analysis is currently being prepared for the performance of the bilingual groups on NWRT and in SRT-Id.

[11]We thank Istvan Fekete for drawing our attention to this method and his support with statistics in the following analysis.

TABLE 6 | Summary of bilingualism factors in the bilingual groups [mean (SD) and range].[a]

	BiTD				BiSLI (*n* = 8)
	BiTD-A (*n* = 8)	BiTD-P (*n* = 19)	BiTD-T (*n* = 19)	Total (*n* = 46)	
Age of onset	29.88 (13.18)	15.74 (17.46)	22.58 (16.25)	21.02 (16.77)	19.50 (16.59)
	0–42	0–39	0–48	0–48	0–36
Length of exposure	58.62 (19.17)	67.10 (22.48)	66.10 (21.48)	65.21 (21.30)	62.00 (22.77)
	36–99	34–101	34–111	34–111	30–88
Early L1 exposure (%)	83% (19.9)	70% (24.3)	67% (19.6)	71% (22.0)	71% (11.97)
Early L2 exposure (%)	36% (30.9)	57% (21.2)	45% (27.0)	48% (26.1)	71% (10.68)
Current L1 richness (/14)	3.5 (1.41)	4.0 (1.5)	4.47 (2.41)	4.11 (1.94)	4.0 (0.0)
Current L2 richness (/14)	9.63 (1.76)	9.32 (1.76)	10.16 (2.41)	9.72 (2.05)	9.0 (0.0)
Language Dominance Index (−50 … +50)	−3.63 (11.2)	1.63 (12.66)	−4.97 (12.53)	−2.01 (12.5)	1.13 (13.15)
	−26 to +11	−17 to +23	−28 to +18	−28 to +23	−21 to +24

[a]*When applicable.*

the "lower" cluster and the minimum score of the "higher" cluster. Individual data points (here scores) allotted to the participants can then be ordered by group, which in turn allows calculation of sensitivity and specificity of the test.

RESULTS

Background Comparisons on Bilingualism Measures

In Section "Participants and Procedure for Verification of Clinical Status," we established that the bilingual groups were comparable in terms of the LI variables "age, non-verbal intelligence and SES." We further compared the bilingual groups for language background information obtained *via* the PaBiQ as displayed in **Table 6**. Group comparisons using a non-parametric Kruskal–Wallis test revealed no significant differences between the bilingual typically developing children according to L1 group on AoO, LoE, early L1 exposure, early L2 exposure, current L1 richness, and current L2 richness as well as the degree of L2 dominance as indicated by the LDI. Likewise, there were no significant differences between the BiSLI and BiTD groups on the aforementioned bilingualism measures.

Following the procedure and using the calculations described in Section "The LITMUS-PaBiQ," we established language dominance in our groups of bilingual participants. **Table 7** summarizes these classifications by L1 and by final status. Note that in the Turkish/German typical children we find the highest rate of L1-dominant children. Among the BiSLI children, balanced or German dominant children are the majority. This might be a reflex of the traditional advice given to parents of bilingual children with language difficulties that they should use the majority language at home or with the child.

Overall Results on the LITMUS NWRT and SRT

We first ran omnibus Kruskal–Wallis tests using scores on NWRT, SRT "identical repetition," henceforth SRT_Id, and SRT "target structure," henceforth SRT_Tar, as dependent variables to determine if clinical group has an effect. All three tests yielded significant results [$\chi^2(3, N = 76) = 33.394$, $p < 0.001$ for NWRT, $\chi^2(3, N = 76) = 38.926$, $p < 0.001$ for SRT_Id, and

TABLE 7 | Language dominance in bilingual children per L1.

	Final status	Dominant L1	Balanced	Dominant L2	Total
Arabic	BiTD	3	3	2	8
	BiSLI	0	2	1	3
Portuguese	BiTD	8	3	8	19
	BiSLI	0	1	0	1
Turkish	BiTD	9	5	5	19
	BiSLI	2	1	1	4
Total	BiTD	20	11	15	46
	BiSLI	2	4	2	8

$\chi^2(3, N = 76) = 38.126$, $p < 0.001$ for SRT_Tar]. In a next step, *post hoc* Mann–Whitney *U* comparisons were carried out on the dependent measures applying Bonferroni-adjustment of *p*-values to reduce Type I error that can arise due to multiple comparisons.

The overall performance of the different groups defined in **Table 2** in the NWRT and SRT is given in **Figure 1**. The NWRT significantly distinguishes the MoSLIs from the MoTDs ($U = 5.5$, $p < 0.001$, $r = 0.767$) and the BiSLIs from the BiTDs ($U = 24.5$, $p < 0.001$, $r = 0.528$). Moreover, BiTDs perform significantly different from the MoSLIs in the NWRT ($U = 38.0$, $p < 0.001$, $r = 0.600$). This means that the LITMUS NWRT can identify SLI across populations. In addition, performance in NWRT does not statistically differ in BiTDs and MoTDs.

Figure 1 further shows that the SRT can well discriminate SLI from TD children in monolingual and bilingual populations with both scoring methods. The score of SRT_Id distinguishes the MoSLIs from the MoTDs ($U = 0.000$, $p < 0.001$, $r = 0.846$) and the BiSLIs from the BiTDs ($U = 36.00$, $p < 0.001$, $r = 0.490$). Here as well, BiTDs perform significantly better than the MoSLIs ($U = 46.5$, $p < 0.001$, $r = 0.578$). If SRT is rated with the measure of SRT_Tar, bilingual children perform better. Again, MoTDs are significantly different from MoSLIs ($U = 0.000$, $p < 0.001$, $r = 0.844$), BiTDs perform significantly better than BiSLIs ($U = 32.5$, $p < 0.001$, $r = 0.539$), and also BiTDs perform significantly better than MoSLIs ($U = 40.0$, $p < 0.001$, $r = 0.595$). However, the MoTDs and BiTDs do not perform alike in the SRT by score SRT_Id: ($U = 76.5$, $p = 0.006$, $r = 0.438$) and SRT_Tar: ($U = 102.5$, $p = 0.036$, $r = 0.364$). Outliers in the SRT scored by
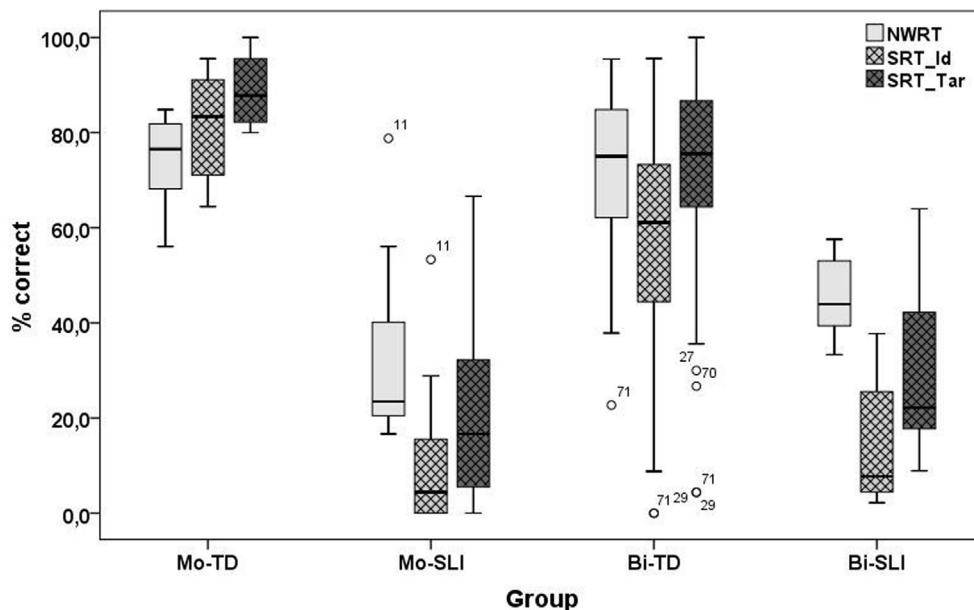
FIGURE 1 | Non-word repetition task (NWRT), SRT_Id, and SRT_Tar: % of correct identical repetition split by group.

SRT_Id are **29**[12] and **71**, where the latter is also the outlier in the NWRT. These two children perform within or even below the BiSLI range. The outlier in the MoSLI group, **11**, is an older child (9;4). Bilingual children performing below the group range in the mastery of SRT_Tar are **29** and **71**, but also **27** and **70**.[13]

For further analyses, we first present results from NWRT and SRT_Id and single out SRT_Tar for closer analysis. We first run a ROC curve analysis on MoTD and MoSLI to determine the optimal monolingual cutoff score and the diagnostic accuracy for each of the tasks. As can be seen in **Table 8**, both LITMUS tests have excellent diagnostic accuracy in monolingual children. When looking at the individual scores of the monolingual children, it emerges that a cutoff of 59.85% for the NWRT and 63.33% on SRT_Id sharply group the children with 100% sensitivity and specificity for SRT_Id and 91.7% sensitivity and 90% specificity for NWRT. Applying the measure SRT_Tar to the monolingual data allows a cutoff of 77.78%, still with 100% sensitivity and specificity.

Comparison of the bilingual groups with the monolingual groups points to the fact that other factors than language impairment could lead to poor performance. To address this problem, we performed a k-means cluster analysis of the performance of all bilinguals on NWRT, SRT_Id and SRT_Tar. The k-clustering, unbiased as to any given classification of participants, renders two clusters, participants who are performing well on the task (cluster A "higher cluster") and those performing poorly (cluster B "lower cluster"), the cutoff line between the two clusters was determined for each of the measures as outlined in Section "Data Analysis."

TABLE 8 | Diagnostic accuracy of the German Language Impairment Testing in Multilingual Settings sentence repetition task (SRT) and non-word repetition task (NWRT) among monolingual children.

	AUC	Cutoff (%)	Sensitivity	Specificity	LR+	LR−
NWRT	0.954	59.85	91.7% (11/12)	90% (9/10)	9.17	0.09
SRT_Id	1.000	63.33	100% (12/12)	100% (10/10)	Undefined[a]	0.0
SRT_Tar	0.996	77.78	100% (12/12)	100% (10/10)	Undefined	0.0

[a]If specificity = 100% then LR+ ratios are undefined.

For the NWRT, the k-means cluster analysis rendered two clusters separated by a k-means cutoff of 63.5%: 34 children performing above cutoff and 20 children scoring below. In the SRT_Id, the analysis rendered a 41.25% cutoff separating the two clusters. On this measure, 35 children, cluster A, scored above the cutoff, whereas 19 children, cluster B, performed below cutoff score. To complete the analysis and calculate the sensitivity and specificity levels, the individual values in the clusters for each task were identified as scores of individual BiTD or BiSLI children. **Figures 2** and **3** depict the performance of cluster A and cluster B in NWRT and SRT_Id, respectively. All of the eight children assigned to the BiSLI group based on standardized test procedures belonged to the lower cluster on both measures, which yields a sensitivity of 100% with an LR− of 0.0. However, the specificity levels and the corresponding LR+ values for ruling out language impairment were only suggestive as can be seen in **Table 9**. This is ascribed to the fact that 12 children with the final status BiTD scored below cutoff on NWRT and 11 children scored below cutoff on SRT_Id, and thus belonged to cluster B.

For the measure of SRT_Tar, k-clustering resulted in two clusters separated by a 52.2% cutoff: 39 children, cluster A, performing above cutoff and 15 children, cluster B, performing
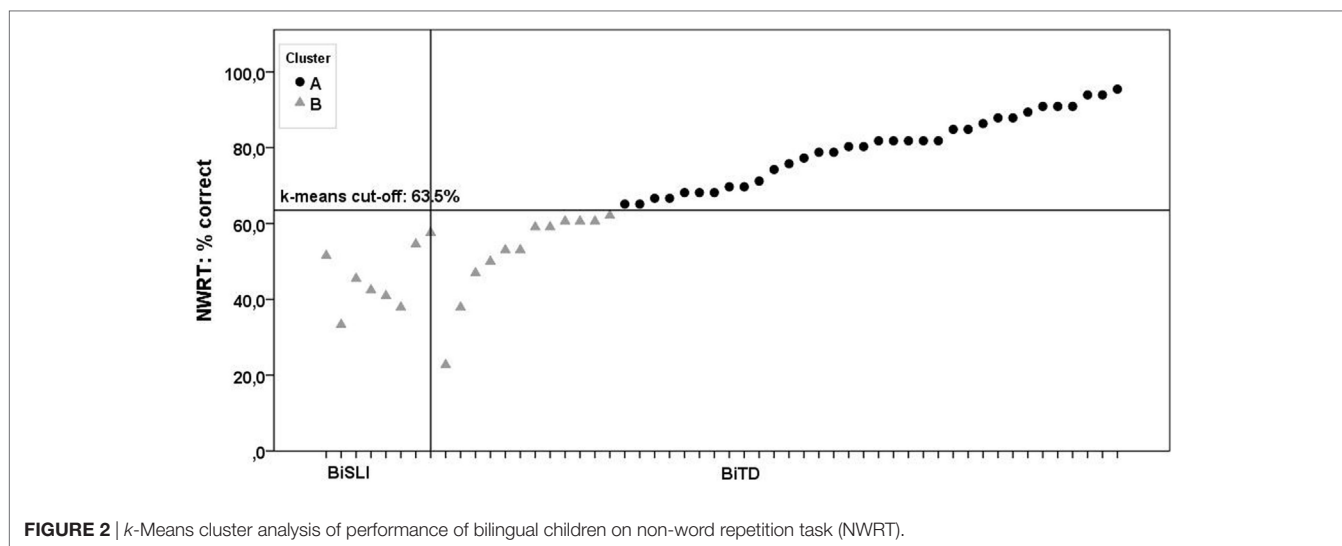
FIGURE 2 | *k*-Means cluster analysis of performance of bilingual children on non-word repetition task (NWRT).
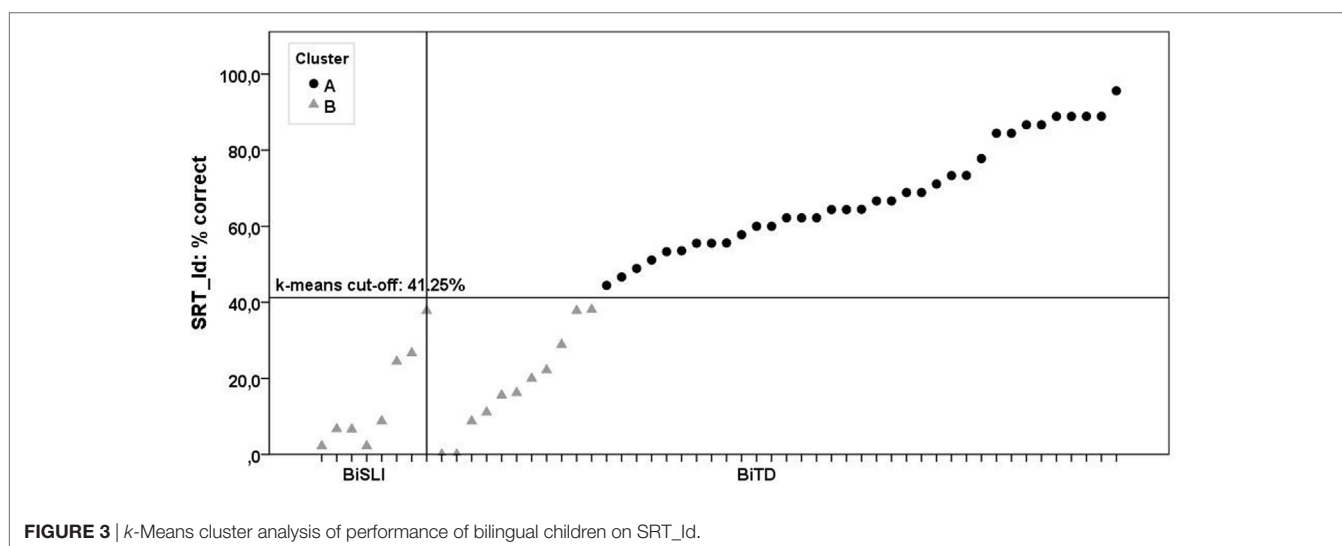

FIGURE 3 | *k*-Means cluster analysis of performance of bilingual children on SRT_Id.

TABLE 9 | Diagnostic accuracy of the German Language Impairment Testing in Multilingual Settings sentence repetition task (SRT) and non-word repetition task (NWRT) among bilingual children for individual measures and for test combinations.

	Cutoff (%)	Sensitivity	Specificity	LR+	LR−
NWRT	63.5	100% (8/8)	73.91% (34/46)	3.83	0.0
SRT_Id	41.25	100% (8/8)	76.1% (35/46)	4.18	0.0
SRT_Tar	52.2	87.5% (7/8)	82.6% (38/46)	5.03	0.15
NWRT + SRT_Id		100% (8/8)	86.95% (40/46)	7.66	0.0
NWRT + SRT_Tar		87.5% (7/8)	86.95% (40/46)	6.70	0.14

below this cutoff. **Figure 4** shows the individual performance of members of cluster A and cluster B in the above measure. The children classified as BiSLI all belonged to cluster B, except **26** who is 9;1 years old and does not seem to be impaired in German morphosyntax. However, eight children in cluster B, some with extremely low scores, had received the final status of BiTD. These children are in particular: **70**, **71**, **44**, **45**, **27**, **76**, **28**,

and **29**. Interestingly, most of these children except for **44** and **45** performed below *k*-means cutoff on both SRT_Id and NWRT. It remains to be investigated why these children scored low.

The measure SRT_Tar, which gives more weight to mastery of syntactically complex structures than to lexical abilities, gave lower sensitivity but better specificity levels than SRT_Id or NWRT, see **Table 9**. We also investigated whether combining the NWRT with SRT raises diagnostic accuracy. The results in **Table 9** indicate that a combination of NWRT and SRT_Id or NWRT and SRT_Tar indeed results in better specificity and thus overall diagnostic accuracy.

Dominance As a Factor for the Performance of Bilingual Typically Developing Children

To examine whether language dominance affects the performance of bilingual children without language impairment, we plotted the children's individual scores on SRT_Tar and NWRT against
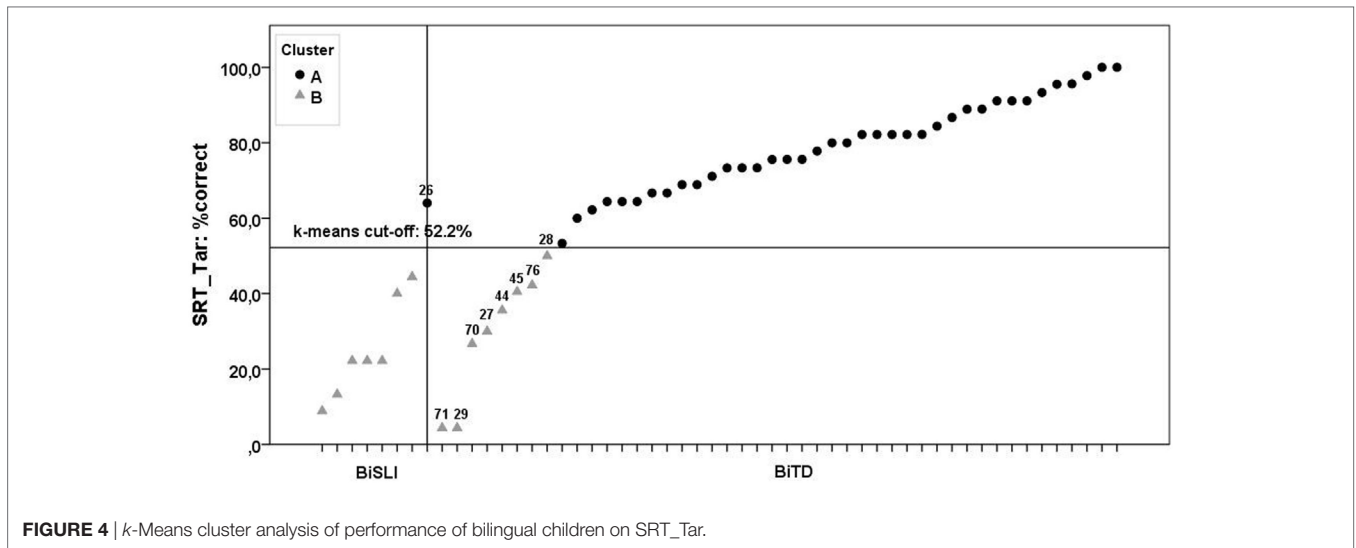
FIGURE 4 | *k*-Means cluster analysis of performance of bilingual children on SRT_Tar.
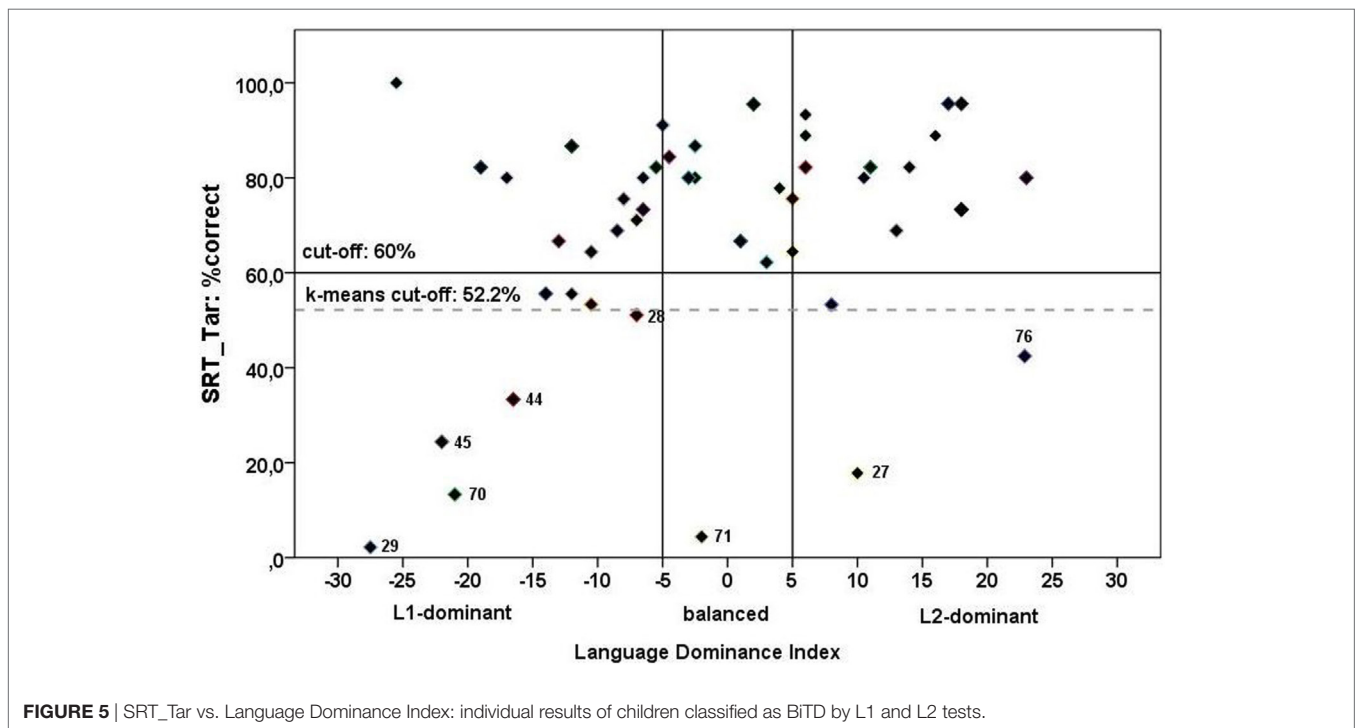

FIGURE 5 | SRT_Tar vs. Language Dominance Index: individual results of children classified as BiTD by L1 and L2 tests.
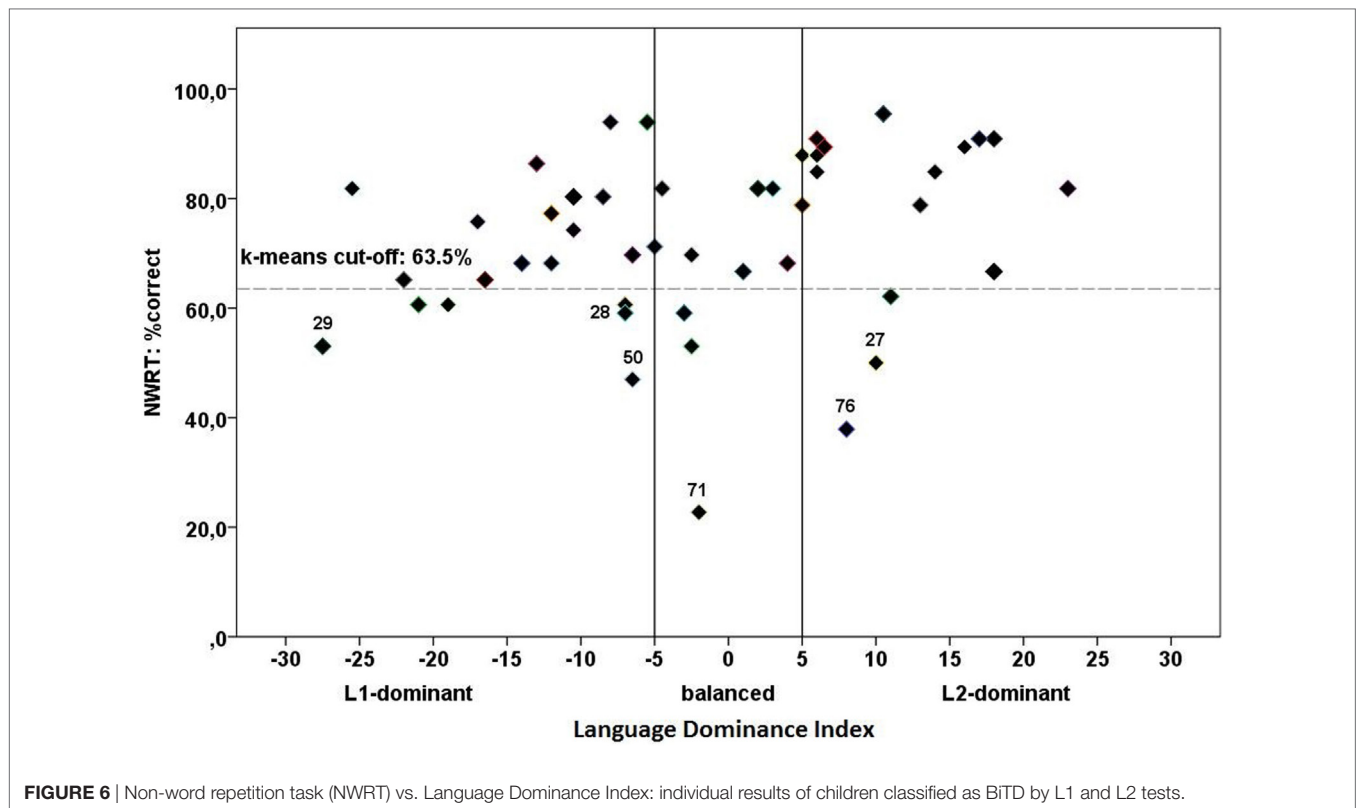
their LDI.[14] As illustrated in **Figure 5**, it emerges that language dominance strongly influences performance of the typical bilingual children in SRT_Tar. On the other hand, just as Almeida et al. (2017) show for the French SRT, among the 20 L1-dominant children the majority, here 70% (14/20) score over 60% correct in SRT-Tar and 75% (15/20) score over 52.2% (see **Figure 5**). The five L1-dominant children who perform below a 52.2% cutoff are identified as **70**, **44**, **45**, **28**, and **29**.

At first glance (see **Figure 6**), language dominance may seem to influence performance on NWRT to the same extent as in

SRT_Tar: 6 out of 20 L1-dominant children perform below the *k*-means cutoff score of 63.5%. However, unlike in the SRT_Tar, four of the latter six children perform almost at cutoff (≥61% correct) and all children perform above cutoff on the LD part of the task.[15] **29**, who performs below cutoff on NWRT, scores above cutoff in the LD part. Only two L1-dominant children **50** and **28** had an overall score on NWRT < 61% due to poor performance on the LD part of the NWRT (**50**: 27.78% correct, **28**: 44%

---

[14]We chose only the measure with higher specificity for the SRT.

[15]In this study, the results of both LI and LD parts of the NWRT are collapsed together. However, in cases of L1-dominant children, we verified that their scores on the LD part were above cutoff to exclude the potential effect of L1 dominance.

FIGURE 6 | Non-word repetition task (NWRT) vs. Language Dominance Index: individual results of children classified as BiTD by L1 and L2 tests.

correct). This allows the conclusion that performance on the NWRT is less independent of language dominance than the SRT. Note also that among the balanced and German dominant children, only three score below the cutoff both in SRT_Tar and in NWRT and these children are **71**, **76**, and **27**, whose status might have to be reanalyzed as will be discussed in Section "Discussion."

DISCUSSION

Summary

This study investigated the accuracy of two German LITMUS tasks, an NWRT and an SRT in the identification of language impairment in bilingual children. Both NWRT and SRT prove to have good sensitivity and specificity in monolinguals: NWRT (sensitivity = 91.7%, specificity = 90%) and SRT show 100% sensitivity and specificity for both scoring methods SRT_Id and SRT_Tar. The results for monolinguals clearly show that the tasks are well constructed and reliably identify SLI. The same can be said for bilingual settings. Especially, the fact that the results for the NWRT are more or less independent of language dominance makes it a valuable new tool for language assessment. The reduced specificity of the SRT for bilinguals is due to several factors that will emerge more clearly in a detailed discussion of the individual cases we highlighted in Sections "Overall Results on the LITMUS NWRT and SRT" and "Dominance As a Factor for the Performance of Bilingual Typically Developing Children." It was noteworthy that the same children were identified in several types of analyses as either being "underdiagnosed" by the SRT (**26**) or of being

"overdiagnosed" (**71**, **27**, and **76**) by both SRT and NWRT. Moreover, L1-dominant children such as **70**, **44**, **45**, **28**, and **29** also performed under cutoff 52.2% in the SRT_Tar.

In Section "Dominance As a Factor for the Performance of Bilingual Typically Developing Children," we already identified one factor of possible misdiagnosis in L2 tasks, namely, L1 dominance, see **Figure 5**. L1 dominance will then interact with other factors, which we discuss in the following. One possibility for a reduced diagnostic accuracy is that our final status assignment might have been too strict, see also Bossuyt et al. (2015) on the impact of clinical group definition on accuracy measures. Note that Armon-Lotem and Meir (2016) use L1 and L2 tests with global scores, but additionally rely on parental or teacher concern. Boerma et al. (2015) and Boerma and Blom (2017) rely on clinical referral, i.e., on L2 testing exclusively. Thordardottir (2015) recommends including measures from samples of spontaneous production in addition to norm-referenced L1 and L2 tests. Given these different methods for identifying the clinical population, we will discuss cases of possible misclassification by our procedure, drawing also on impressions from the samples of narratives we have at our disposal. Alternatively, and given that the SRT, and SRT_Tar in particular, targets morphosyntactic skills, misclassification could arise because our procedure did not take into account selective impairments such as grammatical/syntactic SLI. This would mean that an individual child has been classified as BiTD, but is syntactically impaired, which arguably leads to poor performance in SRT_Tar. Children who show impairments in phonology and lexicon, but not in morphosyntax, would have been classified as BiSLI, but will not

necessarily perform poorly in the SRT. Finally, misclassification could arise if standardized tests are not reliable in certain constellations of bilingualism, such as heritage situations.[16] In the discussion, we specifically address the problems arising from our strict procedure and the (non)-applicability of standardized L1 tests in heritage situations.

Subgroups of SLI

Since our classifying procedure did not isolate subgroups of SLI, but clearly aimed at a broader definition, we first address this problem by discussing the cases revealed by the clustering for SRT_Tar, see **Figure 4**. The BiSLI child in cluster A, **26**, was classified as BiSLI because of her scores in the L1 test, ELO-L, and because her lexical and phonological abilities were below norm in L2. Note that she was 9;1 years at the time of testing but she performed well below the norms for younger children (7;11) in the L1 test, in which her sentence production showed a slight impairment whereas her phonological production showed great deficits. For L2 testing, the lexical test is normed till 9;11, for phonology she was below norm of younger children and the LiSe-DaZ norms could be age adjusted as described in Section "Standardized L2 and L1 Tests." Her spontaneous L2-language sample did not evidence any of the characteristic markers of SLI. This indicates that she might not be syntactically impaired. This seems to be confirmed by her good score in SRT_Tar. Not surprisingly, her score in SRT_Id was below 41.25%, despite her age, since it involves recollection of vocabulary.

The same problem, namely, that the impairment might be selective, is exemplified by **27**, who is BiTD because of reasonable scores in L2 lexicon and phonology, but has clear problems in some areas of morphosyntax identified by the LiSe-DaZ, among them SVA. Since this domain does not receive a *t* value in the test, it is not included in our final evaluation procedure. **27** is a simultaneous bilingual, L2 dominant, clearly impaired in L1, and her spontaneous production in both languages confirms problems with morphosyntax. **27** performs low in the SRT in both measures as well as in the NWRT and hence may be selectively impaired, i.e., the final status allotted may be misleading.

To see whether we might have missed bilingual children with grammatical SLI, we first consider children who show poor L2 performance only in the LiSe-DaZ. These are **24**, **27**, and **28**. **27** was already discussed as a possible case of grammatical SLI. **28** is below cutoff in SRT_Tar and in the NWRT. She is also language impaired by her L1 status and her parents voiced concern. In other words she would be a BiSLI child if we had included selective impairments. **24** does not show problems in any of the experimental tasks and is not L1 impaired.

To summarize: **27** and **28** might be cases of grammatical SLI who show poor performance in LiSe-DaZ, but also in the SRT. On the other hand, **26** might be a case of lexical/phonological impairment with good performance in the LiSe-DaZ and in SRT_Tar.

Reevaluating L1-dominant children below cutoff (52.2%) by taking into account the possibility of selective impairments, leads to the following picture: **70**, **45**, and **29**[17] have a very high L1-language index and score as typical children in their L1 tests. **45** also performs well in the NWRT, and **70** performs almost at cutoff on the task. This implies that L1 dominance explains the performance of **70** and **45**. **28** and **44**, however, score as language impaired in their L1 and are classified as impaired in one of the L2 tests applied here: **28** has morphosyntactic problems, **44** performs low in the lexical assessment. These children might therefore be selectively impaired.[18] However, only **28** also performs low in the NWRT, whereas **44** performs above cutoff. **29** remains a problem since she performs low in all L2 tests as well as the experimental tasks whereas her L1 test puts her firmly among the typical children.

Interestingly, most of the children discussed above perform below cutoff not only in SRT, but also in NWRT: **71**, **76**, and **27** among the balanced and L2-dominant children, and **28** and **29** among the L1-dominant children.

L1 Assessment in Heritage Situations

It is not surprising that some children who perform below cutoffs in the standardized L2 tasks (and also in the SRT and NWRT), nevertheless have a final status as BiTD because they did well in the L1 tests. Five of the 46 typical bilinguals would be diagnosed as language impaired by the L2 tests but are doing well in L1. Especially for L1-dominant children this might be expected. **29** is a case in point: all three German tests classified the child as impaired, so would NWRT and SRT. The child did perfectly well in the TEDIL, however. **71**, who is balanced according to the PaBiQ, seems to be a similar case but turns out to be a child whose final status might be reconsidered: **71** scored above the norms in lexical reception in the GOL-E, but would have been below the norm in the assessment of the lexicon provided by the PALPA-P, which exists for her age range. Recall that we decided to use GOL-E as lexical assessment for all Portuguese children because in this area there were age gaps in the norms for the PALPA-P, which does not apply to **71**. More surprising is the fact that even German dominant children who scored as impaired in the German tests sometimes do well in the L1 tasks as is the case of **76**. However, **76** performed only minimally above cutoff in the TEDIL. Both **76** and **71** would be classified as language impaired if their (L1 and L2) samples of spontaneous production had been included in the initial decision about final status.

If the results of the L1 tests are examined more closely, it is rather striking that 16 of the 46 bilingual typical children have an L1 diagnosis of impairment, whereas only 5 are so diagnosed by the L2 tests. Examining these numbers by dominance we see that among the 20 L1-dominant children, 6 would have been diagnosed as impaired by the L1 test (1 Arabic child, 3 Portuguese children, and 2 Turkish children). Among the 11 balanced

---

children, 2 would have been diagnosed as L1 impaired. Among the 15 German dominant children, 8 would have been diagnosed as L1 impaired (2 Arabic children, 5 Portuguese children, and 1 Turkish child). These figures point to problems with the applicability of the L1 tests, which, in turn, call in question the final BiSLI status.

There are multiple reasons for this situation. Heritage speakers growing up as simultaneous bilinguals often show differences to monolingual speakers in their adult performance. This seems to concern morphosyntax and lexicon more than phonology (Montrul, 2010; Rinke and Flores, 2014). Reasons for this situation have been sought in the fact that children who have been exposed to their L2 early or are simultaneous bilinguals are often subject to language attrition in their L1 or could be claimed to suffer from incomplete acquisition (Köpke et al., 2004; Montrul, 2008, 2010; Benmamoun et al., 2013). Moreover, the language of children growing up as Turkish/German bilinguals in Germany is special from several perspectives: They are often third generation heritage speakers and Immigrant Turkish in Germany has features (Schroeder and Dollnick, 2013) which count as clinical markers for SLI in Standard Turkish (see also Chilla and San, 2017). Finally, the L1 tests we chose might have other inherent problems: The TEDIL only has two global scores, which do not allow identifying language domains as specifically problematic. The version of the Portuguese PALPA-P that we used has been normed with only few children for some ages and subtests. It is a linguistically well-controlled test but the lack of norms in receptive vocabulary in crucial age ranges made it necessary to use a different test for assessment of the lexicon, the GOL-E. Some of the misdiagnosis may therefore be due to the specific language tests chosen here. The more fundamental problems seem to be the heritage situation and language attrition of L1 which has been shown to be particularly noticeable when L2 exposure is early, see Lein et al. (2017) for an analysis of heritage effects in the Portuguese bilinguals also investigated in this study.

If the 15 children dominant in their L2 German are considered, more than half of them (8) would have been classified as SLI if only tested in their L1, which is not surprising. In contrast, overdiagnosis due to the L2 tests did not occur. German dominant children with a final classification of BiTD (also considering L1) were all correctly classified as BiTD by the combination of the three norm-referenced tests used for classification. Of the 46 children with a final classification of BiTD there are only three children who would have been BiSLI if only the L2 had been considered. This seems to indicate that in heritage situations L2 tests are more reliable than L1 tests, which may have multiple reasons: the contact situation and the existence of immigrant varieties, language attrition, and possibly properties of the L1 tests. Incidentally, **29**, who remains a problematic case also after we reconsidered the status of the bilingual children, might highlight the problems with the Turkish test, see also Almeida et al. (2017).

Reconsidering the Status of the Bilingual Children

Following the argumentation about (a) selective impairments and (b) possible problems with L1 tests in heritage situations,

we suggest different criteria for the identification of the bilingual clinical group: We consider children as BiSLI if they have a selective L2 impairment, and score below norms in their L1 tests or show poor spontaneous production in both languages.[19] These criteria still require an impairment in both languages of a bilingual child but would classify **71**, **76**, **27**, **28**, and **44** as BiSLI. Incidentally, two of these children had been in SLT when recruited (**27** and **28**) and the remaining three might be cases of underdiagnosis. Given the clustering shown by **Figures 2–4**, and the foregoing discussion of these particular cases, such a regrouping would clearly raise diagnostic accuracy for all measures (SRT_Id, SRT_Tar and NWRT).

Given that the grouping of children we presented in **Table 2** takes into account language dominance by adjusting the norms in standardized L1 and L2 tests, the discussion points raised above show that especially selective impairments should be taken into account when deciding on the status of a bilingual child and when considering the accuracy of a particular test, which might be targeting one language domain more than others. The heritage situation adds to the difficulty and the cases discussed suggest that norm adjustments for L1 might have to be reconsidered for heritage speakers.

CONCLUSION

Our investigation of the German LITMUS NWRT and SRT has shown that both are well suited as tools for the identification of SLI in bilinguals. Both tasks clearly identify SLI in German monolinguals demonstrating that they target crucial phonological and syntactic areas and structures. In addition, both tasks can identify SLI in bilingual contexts. Since the construction of both tasks was guided by linguistic notions such as phonological or syntactic complexity and neither task primarily measures WM, this is a result relevant on the theoretical and the practical level.

Both tasks clearly measure linguistic abilities, the NWRT on the phonological side, the SRT in morphosyntax. The SRT was scored in two different ways: SRT_Id as a measure includes all morphosyntactic but also all lexical errors, not cumulating them. SRT_Tar scores only morphosyntax and, by concentrating on syntactic structures and not counting morphological errors such as case or gender if they do not change the structure aimed at, does not penalize bilingual children and seems a good measure of (morpho)syntactic abilities. From the practical point of view, the possibility of using both or one of these scoring methods allows fine-grained diagnosis of the impaired domains. Concentrating on certain structures such as those involving Wh-movement with and without embeddings or intervention (see The German LITMUS SRT) would give an even more detailed picture but was not the focus of this study.

Our evaluation of the LITMUS tools started with rather strict criteria as to the status of a bilingual child as typical or language impaired. We classified a child as BiSLI only if the

---

[19]We do not apply any formal measure here but judge production by certain markers: correct SVA, sentence bracket or V2 and presence of embeddings.

child scored below (adjusted) norms in two domains of both her L1 and her L2. For this categorization, and also for further evaluation of our results, see Section "Dominance As a Factor for the Performance of Bilingual Typically Developing Children," the parental questionnaire, the PaBiQ, was an indispensable tool. We concentrated on the language dominance value, which allowed adjusting the norms for standardized tests and helped us in the interpretation of our results on the performance in the LITMUS NWRT and SRT. It emerged that performance in the NWRT is largely independent of language dominance, whereas it influences performance in the SRT. However, 75% of the L1-dominant children performed above the cutoff in the SRT when scored as SRT_Tar, so that accuracy remains satisfactory. Similar findings are reported in Almeida et al. (2017) for the corresponding French tasks and in Grimm and Hübner (in press) for the German NWRT. Interestingly, LoE does not influence performance in the NWRT either, as reported in Grimm and Hübner (in press).[20]

Considering individual cases and their performance in these new tasks revealed that the grouping we chose on the basis of standardized L1 and L2 tests might have missed cases of language impairment, which would not be surprising giving that this classification cut in half the group of children in SLT. We attributed such missed cases to either the problems of using and interpreting L1 tests (even with adjusted norms) in heritage contexts or to cases of selective language impairments. Clearly, interpretation of individual results for bilinguals is impossible without background information as provided by parental questionnaires.

On the practical level, this leads us to conclude that the LITMUS NWRT and SRT are indeed reliable tools that can be used as a first evaluation of a child's language abilities, singly, but better in combination. Since their administration takes only a fraction of the time that has to be invested for standardized tests, this is a good overall result. On the theoretical level, we have shown that L2 tasks, if linguistically well controlled and targeting complex structures, clearly identify language impairment in bilingual contexts.

---

[20]Since this factor contributes to our dominance calculation, we did not consider it separately.

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: Cascadilla Press), 1–17.

Abed Ibrahim, L., Lein, T., Rothweiler, M., and Hamann, C. (in press). "SLI in bilinguals: comparing LITMUS-tasks to a standardized test with bilingual norms," in *Bilingualism and Specific Language Impairment, Bi-SLI 2015*, eds C. dos Santos and L. de Almeida (Amsterdam: Benjamins).

Almeida, L., Ferrré, S., Morin, E., Prévost, P., dos Santos, C., Tuller, L., et al. (2017). Identification of bilingual children with specific language impairment in France. *Linguist. ApproachesBiling.* 7, 3/4. doi:10.1075/lab.15019.alm

Archibald, L., and Gathercole, S. (2006). Short-term and working memory in specific language impairment. *Int. J. Lang. Commun. Disord.* 41, 675–693. doi:10.1080/13682820500442602

ETHICS STATEMENT

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

FUNDING

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

Armon-Lotem, S., de Jong, J., and Meir, N. (2015). *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment.* Bristol: Multilingual Matters.

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

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., Leseman, 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. *JSLHR* 58, 1747–1760. doi:10.1044/2015_JSLHR-L-15-0058

Bossuyt, P., Reitsma, J., Bruns, D., Gatsonis, C., Glasziou, P., Irwig, L., et al. (2015). STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology* 277, 826–832. doi:10.1148/radiol.2015151516

Bulheller, S., and Häcker, H. (2002). *Coloured Progressive Matrices*. Frankfurt: Pearson Assessment.

Castro, S. L., Caló, S., Gomes, I., Kay, J., Lesser, R., and Coltheart, M. (2007). *PALPA-P Provas de Avaliacao da Linguagem e da Afasia em Português*. Lisbon: CEGOC.

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., and Polišenská, K. (2016). A framework for crosslinguistic nonword repetition tests: effects of bilingualism and socioeconomic status on children's performance. *JSLHR* 59, 1179–1189. doi:10.1044/2016_JSLHR-L-15-0293

Chilla, S., and San, 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, N. Topaj, R. Thomas, and I. Gülzow (Berlin: Peter Lang Verlag), 175–205.

Clahsen, H. (1991). *Child Language and Developmental Dysphasia: Linguistic Studies of the Acquisition of German*. Amsterdam: John Benjamins.

Cobo-Lewis, A., Pearson, B., Eilers, R., and Umbel, V. (2002). "Effects of bilingualism and bilingual educations on oral and written English skills: a multifactor study of standardized test outcomes," in *Language and Literacy in Bilingual Children*, eds D. K. Oller and R. E. Eilers (Clevedon: Multilingual Matters), 43–63.

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

Cornips, L., and Hulk, A. (2008). Factors of success and failure in the acquisition of grammatical gender in Dutch. *Second Lang. Res.* 24, 267–297. doi:10.1177/0267658308090182

Dollaghan, C. (2004). Evidence-based practice in communication disorders: what do we know, and when do we know it? *J. Commun. Disord.* 37, 391–400. doi:10.1016/j.jcomdis.2004.04.002

Dollaghan, C. (2007). *The Handbook for Evidence-Based Practice in Communication Disorders*. Baltimore, MD: Paul H. Brookes.

dos Santos, C., and Ferré, S. (2016). A non-word repetition task to assess bilingual children's phonology. *Lang. Acquis.* 1–14. doi:10.1080/10489223.2016.1243692

Dunn, G. (2011). "Statistics in psychiatry," in *International Encyclopedia of Statistical Science*, ed. M. Lovric (Berlin: Springer), 1136–1138.

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.

Fleckstein, A., Prévost, P., Tuller, L., Sizaret, E., and Zebib, R. (2016). How to identify SLI in bilingual children: a study on sentence repetition in French. *Lang. Acquis.* 1–17. doi:10.1080/10489223.2016.1192635

Fox-Boyer, A. V. (2014). *Psycholinguistischer Analyse kindlicher Aussprachestörungen-II: PLAKSS II*. Frankfurt am Main: Pearson.

Fredman, M. (2006). Recommendations for working with bilingual children – prepared by the multilingual affairs committee of IALP. *Folia Phoniatr. Logop.* 58, 458–464. doi:10.1159/000096570

Friedman, 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.

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.

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

Genesee, F., Paradis, J., and Crago, M. (2004). *Dual Language Development and Disorders: A Handbook on Bilingualism and Second Language Learning*. Baltimore: Brookes Publishing.

Glück, C. (2007). *WWT 6-11. Wortschatz- und Wortfindungstest für 6-10jährige Kinder*. Munich: Elsevier.

Goldberg, H., Paradis, J., and Crago, M. (2008). Lexical acquisition over time in minority first language children learning English as a second language. *Appl. Psycholinguist.* 29, 41–65. doi:10.1017/S014271640808003X

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, July 13–19, 2014* (Amsterdam: IASCL).

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).

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

Gutiérrez-Clellen, V., and Simon-Cereidjido, G. (2010). Using nonword repetition tasks for the identification of language impairment in Spanish-English-Speaking children: does the language of assessment matter? *Learn. Disabil.* 25, 48–58. doi:10.1111/j.1540-5826.2009.00300.x

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 (Mahwah, NJ: 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. (2015). "Specific language impairment in German speaking children," in *Specific Language Impairment. Current Trends in Research*, ed. S. Stavrakaki (Amsterdam: Benjamins), 215–251.

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 (New Castle: 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*, Oldenburg.

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 di Grammatica Generativa* 36, 146–181.

IBM SPSS 22. (2013). *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY: IBM Corp.

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:10.1207/s15327817la0702-4_3

Kay, J., Lesser, R., and Coltheart, M. (1996). *PALPA: Psycholinguistic Assessments of Language Processing in Aphasia*. Hove: Lawrence Erlbaum Associates.

Köpke, B., Schmid, M. S., Keijzer, R., and Weilemar, L. (2004). "First language attrition: the next phase," in *First Language Attrition: Interdisciplinary Perspectives on Methodological Issues*, eds M. Schmid and S. Köpke (Amsterdam: John Benjamins), 1–43.

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 *GASLA Proceedings 2015*, eds D. Stringer, J. Garrett, B. Halloran, and S. Mossman (Somerville, MA: Cascadilla Proceedings), 124–135.

Lein, T., Rothweiler, M., and Hamann, C. (2017). "Factors affecting the performance in child heritage Portuguese in Germany," in *Proceedings of GALA 2015*, ed. H. Demirdache (Newcastle: Cambridge Scholars Publishing), 146–175.

Leonard, L. (1998). *Children with Specific Language Impairment*. Cambridge: MIT Press.

Leonard, L. (2014). *Children with Specific Language Impairment*, 2nd Edn. Cambridge: MIT Press.

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1 (California: University of California Press), 281–297.

Maddieson, I. (2006). Correlating phonological complexity: data and validation. *Linguis. Typol.* 10, 106–123. doi:10.1515/LINGTY.2006.017

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/

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*, Vol. 7, eds J. Rothman and S. Unsworth (Amsterdam: John Benjamins), 265–276.

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

Meisel, J. (2009). Second language acquisition in early childhood. *Zeitschrift für Sprachwissenschaft* 28, 5–34. doi:10.1515/ZFSW.2009.002

Montrul, S. (2008). *Incomplete Acquisition in Bilingualism. Re-Examining the Age Factor*. Amsterdam: Benjamins.

Montrul, S. (2010). Current issues in heritage language acquisition. *Annu. Rev. Appl. Linguis.* 30, 3–23. doi:10.1017/S0267190510000103

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. doi:10.1017/S0142716409990373

Paradis, J. (2011). Individual differences in child English second-language acquisition: comparing child-internal and child-external factors. *Linguis. Approaches Biling.* 1, 213–237. doi:10.1044/jslhr.4304.834

Paradis, J., Emmerzael, 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., Martha, C., 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., 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

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. (2014). Sentence repetition: what does the task measure? *Int. J. Lang. Commun. Disord.* 50, 106–118. doi:10.1111/1460-6984.12126

Rice, M., and Wexler, K. (1996). "Tense over time: the persistence of optional infinitives in English in children with SLI," in *Proceedings of the 20th Annual Boston University Conference on Language Development*, eds A. Stringfellow, D. Cahana-Amitay, E. Hughes, and A. Zukowski (Somerville: Cascadilla Press), 610–621.

Rinke, E., and Flores, C. (2014). Morpho-syntactic knowledge of clitics by Portuguese heritage children. *Biling. Lang. Cognit.* 17, 681–699. doi:10.1017/S136672891300076X

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.

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. Cognit.* 15, 39–57. doi:10.1017/S136672891100037X

Schönenberger, M., Rothweiler, M., and Sterner, F. (2012). "Case marking in child L1 and early child L2 German," in *Multilingual Individuals and Multilingual Societies. Hamburg Studies on Multilingualism*, eds K. Braunmüller and C. Gabriel (Amsterdam: Benjamins), 3–22.

Schroeder, C., and Dollnick, M. (2013). "Mehrsprachige Gymnasiasten mit türkischem Hintergrund schreiben auf Türkisch," in *Mehrsprachig in Wissenschaft und Gesellschaft. Tagungsband*, eds C. Riemer, H. Brandl, E. Arslan, and E. Langelahn (Open-Access Publication Publikationsplattform BieColl of the University of Bielefeld), 101–114. doi:10.2390/biecoll-mehrspr2013_0

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: Hogrefe Verlag.

Sua-Kay, E., and Santos, M. E. (2014). *Grelha de Avalicacao da Linguagem – nível escolar (GOL-E)*, 2nd edição revista Edn. Alcoitao: Escola Superior de Saúde do Alcoitao.

Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. *Int. J. Biling.* 14, 426–445. doi:10.1177/1367006911403202

Thordardottir, E. (2012). *Proposed Diagnostic Procedures and Criteria for Bi-SLI COST Action Studies. COST Action IS080*. 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.

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

Tomblin, B. J., Records, N. L., Buckwalter, P. R., Zhang, X., Smith, E., and O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *JSLHR* 40, 1245–1260.

Topbaş, S., and Güven, S. (2013). *Turkce Erken Dil Gelisim Testi (TEDİL) (Test of Early Language Development: TELD-3, Turkish Version)*. Ancara: 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.

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

Zebib, R., Henri, G., Khomsi, A., Messara, C., and Hreich, E. (2017). *Batterie d'Evaluation du Langage Oral chez l'enfant libanais (ELO-L)*. Liban: LTE.