




ORIGINAL ARTICLE

# Language assessment tools for Arabic-speaking heritage and refugee children in Germany

Cornelia Hamann<sup>1\*</sup>, Solveig Chilla<sup>2</sup> , Lina Abed Ibrahim<sup>1</sup>  and István Fekete<sup>1</sup> 

<sup>1</sup>Carl von Ossietzky Universität Oldenburg, Germany and <sup>2</sup>Europa Universität Flensburg, Germany

\*Address for Correspondence: Cornelia Hamann, Institute of English and American Studies, University of Oldenburg, P.O. Box, D-26111 Oldenburg, Germany. E-mail: [cornelia.hamann@uni-oldenburg.de](mailto:cornelia.hamann@uni-oldenburg.de)

(Received 04 April 2019; accepted 28 June 2020; first published online 12 October 2020)

## Abstract

Though Germany has long provided education for children speaking a heritage language and received two recent waves of refugees, reliable assessment tools for diagnosis of language impairment or the progress in the acquisition of German as a second language (L2) by refugee children are still lacking. The few tools expressly targeting bilingual populations are normed for younger, early successive bilingual children. This study investigates 27 typically developing children with Arabic as first language (L1), comparing 15 school-age Syrian refugees (6;6–12;8), with 12 heritage speakers (6;0–12;9). We assess the L1 and L2 skills of these two groups with standardized tests, but crucially with an Arabic and a German sentence repetition (SRT) as well as a nonword (NWRT) repetition task (Grimm & Hübner, *in press*; Marinis & Armon-Lotem, 2015). Comparable scores emerged only for German LITMUS-NWRT and Arabic LITMUS-SRT. Refugee children had an advantage in L1 measures, for example, vocabulary and morphosyntactic production, whereas they performed poorly in the German LITMUS-SRT and other L2 tests involving morphosyntax and vocabulary even with 24 months of systematic exposure. This indicates that the acquisition of adequate vocabulary and complex syntax takes time. The paper explores factors influencing performance on the repetition tasks and relates the results to established diagnostic procedures and educational policies.

**Keywords:** educational policies; heritage children; language assessment; refugee children; repetition tasks

Germany has experienced continuous immigration for the last five decades so that bilingualism and multilingualism have been a focus of linguistic research which has investigated trajectories and outcomes of language acquisition in situations where a child either acquires two languages simultaneously (2L1), or successively at preschool (early child L2) or at school-age (late child L2). In heritage situations, that is, when a language is spoken at home, but not by the majority, 2L1 or early child L2 acquisition is prevalent and research has focused on the development of the majority as well as the heritage language in such bilinguals. In refugee children, that is, newcomers forced to leave their country because of war or other political circumstances,

© The Author(s), 2020. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

early or late child L2 acquisition is the rule, but their school-age second language (L2) and first language (L1) development is less well documented: the recent influx of refugees has revealed that tools for assessing the progress of late child L2 speakers, or for identifying typical development by excluding language disorders in such populations, are still lacking. For excluding language disorders, it is considered best practice to test a bilingual child in both her languages. It is, thus, necessary to investigate how heritage children and refugee children compare to each other for L2 and L1 assessments and, crucially, whether L2 assessment tools developed in heritage contexts, such as the LITMUS repetition tasks, can be used with school-age newcomers, who in this study are all Syrian refugees, called “refugees” in the following.

## Background

### ***2L1 and early child L2 bilinguals***

Due to sociopolitical circumstances, heritage children are well studied in Germany and elsewhere (see Meisel, 2009, for an overview). The age of first contact with the L2, the age of onset (AoO), was singled out as a decisive factor for success in L2 acquisition, even though it has been controversially discussed (Bialystok, 1997; Birdsong, 2018; Newport, 1990). Other (external) factors, such as L1 and L2 input and use, also socio-economic status (SES), particularly mothers’ education, have been shown to influence a child’s L1 and L2 performance (see Chondrogianni & Marinis, 2011), often leading to a certain unevenness in the development of the two languages such that a child can be more proficient in one language than the other.

The observation of unevenness becomes particularly striking when early successive bilinguals do not develop their L1 as their monolingual peers in the home country do once they experience systematic exposure to the L2 (Bedore, Peña, Griffin, & Hixon, 2016; Montrul, 2008). This phenomenon is often found in heritage speakers and is variously described as language attrition or “incomplete” L1 acquisition (Karayayla & Schmid, 2018; Köpke, 2007; Montrul, 2008; but see Rothman & Kupisch, 2018, for arguments that the term “incomplete” L1 acquisition is inappropriate). The AoO of the additional language has been identified as particularly important for L1 skills, which seem to be more affected the earlier the contact with the additional language occurs (see Albirini, 2018; Montrul, 2008). However, L1 input and use may interact with AoO in interesting ways, see Bedore *et al.* (2016). Rothman and Kupisch (2018), investigating the effect of L1 literacy and schooling on L1 maintenance, add that quality of L1 input is another crucial factor. Moreover, there are differential effects in L1 development if an L1 speaker lives in isolation in an L2 environment or if such a speaker is surrounded by a community using a variety of the L1 that has already evolved as a contact variety (see Köpke, 2007, and the situation of Immigrant Turkish, Chilla, *in press*; Chilla & Şan, 2017; Schroeder & Dollnick, 2013).

### ***Late child L2 acquisition by newcomers and refugee children***

The strengths and needs of heritage children in the context of early education are thus well understood. However, Germany has experienced two major waves of refugees, first from the Balkans and then from Syria. Consequently, several longitudinal studies have investigated appropriate schooling and language support

(e.g., Becker-Mrotzek, Henstchel, Hippmann, & Libbenabb, 2012; Gogolin, Lamge, Lengyel, & Scheippert, 2011) and models for language support and formative language assessment for refugees have been developed (Diehm & Radtke, 1999; von Dewitz, Terhart, & Massumi, 2018). The challenge for these studies on newcomers, that is, late child L2 learners, is the question whether a late AoO (at 6 or more years of age) will necessarily lead to more difficulties in L2 acquisition. The evidence in this respect is mixed. Clearly, cognitive and linguistic factors will interact at school-age, where working memory and other cognitive functions have developed, whereas specific discovery mechanisms may no longer be easily available (Meisel, 2009), and language is not only learned by naturalistic exposure but also crucially in a classroom environment. It has been shown that the acquisition of the lexicon (Goldberg, Paradis, & Crago, 2008) and of complex constructions, such as passive (Rothman et al., 2016), proceeds faster in older than in younger children due to their greater cognitive resources. Because of the greater variation described for adult L2 development (see Birdsong, 2018, for an overview) and also for late child L2, other factors such as length of exposure (LoE) or working memory capacities have been explored (Farnia & Geva, 2011; Paradis, 2011). As in the case of heritage children, L1 and L2 input and use, literacy, SES, mothers' use of L1 and proficiency in the L2, have been discussed as influencing L1 and L2 performance (Cobo-Lewis, Pearson, Eilers, & Umbel, 2002; Gutierrez-Clellen & Kreiter, 2003; Paradis, 2011; Paradis & Jia, 2016; Unsworth, 2016).

Taken together, (psycho-)linguistic research on the L1 and the L2 acquisition of bilingual children sketches them as a group of language learners that varies in many aspects. However, most studies up to the present base evaluations of proficiency or new assessment tools on data from heritage children (Chilla, 2008; Gagarina, 2017; Hamann & Abed Ibrahim, 2017; Montanari, Akinci, & Abel, 2019; Tuller et al., 2018). The two language assessments developed in Germany providing norms for bilingual (and monolingual) children, Russian SKRUK (Gagarina, Klassert, & Topaj, 2010, for Russian-German bilinguals) and LiSe-DaZ (Schulz & Tracy, 2011, for bilinguals with German as majority language), likewise are normed on heritage children, due to a certain stability in the resettlement of immigrants in Germany up until September 2015.

Studies with school-age refugee children remain a notable exception, and assessment tools developed for this group mostly focus on performance in the academic variety of German, with an emphasis on vocabulary, narrative production, writing, and literacy in the L2 (see Chilla, Krupp, & Wulff, 2019; Gantefort & Roth, 2008; Grieshaber, 2013). First results comparing heritage children at primary school age with a group of older refugees (10–17 years of age) on their vocabulary development and their ability to write a picture-based essay show a big gap in language skills between this older group of refugees and the younger group of heritage children, favoring the latter (Montanari, 2017). The study highlights the difficulties in choosing and applying language assessment tools in the absence of valid norms or of comparable L2 reference groups (see the recommendations in Geva & Wiener, 2015). Unfortunately, tests with age-appropriate norms are completely absent for learners with their first L2 exposure at the age of 6 years or older. Consequently, it is strikingly evident that systematic studies on the language development of refugees who enter the school system when they are older than 6 or 7 years are lacking.

### **Education for refugee children in Germany**

Due to the influx of almost a million refugees to Germany in September 2015, many thousands of children and adolescents, who differed in age of arrival, schooling experience, language backgrounds, asylum status, and transit itineraries, had to be integrated into the German school system (Hahn-Hobeck, 2017; Schroeder & Seukwa, 2017; von Maurice & Roßbach, 2017, and on the legal rights and obstacles for newly arrived refugee children). Thus, the challenges for educational policies are quite comparable with the situation in Germany after the Balkan wars in the 1990s. Populations, however, differ not only with respect to the typology of languages involved but also in their level of education and literacy, the effects of interrupted schooling, transit itineraries, and duration of stay in mass shelters.

Research on the academic success of bilinguals in German schools is based on heritage children and is focused on their L2 development (e.g., Montanari *et al.*, 2019). Based on the previous discussion on language development, it is clear that comparing and transferring these results to refugees who arrived at later ages is difficult. Even though many studies in educational science recommend including languages of origin for teaching (Roth, 2018) or using translanguaging as a method of language education (García, 2009; Gogolin, 1994; List & List, 2004), knowledge of the German language is the basic condition for academic or professional success. Since education policy rests with the 16 Federal States in Germany, several parallel models of schooling exist (Ahrenholz, Fuchs, & Birnbaum, 2016). Importantly, irrespective of such different models, the general design, established in the 1980s and based on 2L1 or early child L2 speakers, remains in effect: refugee children are expected to acquire oral and literary L2 skills within 6 to 18 months before they are to participate in regular classes. Furthermore, most schools establish separate language classes irrespective of the students' age or overall development.

Given the different models across Germany, (second) language education of refugees finds itself caught between general schooling and specific language support for refugees, ranging from exclusion of refugee students from regular classes to immediate integration into regular age-appropriate classes (see Figure 1). The schooling model assigned to one individual refugee child is therefore rarely based on age and/or overall development, but depends heavily on the schooling policies of the place of residence.

These programs clearly have different assumptions about language acquisition, academic success of immigrant children, and the contribution multilingual students bring to the learning environment of ordinary classes. Schooling in separate classes has been criticized for its lack of inclusion and integration and seems to presuppose that the refugees' language development will not suffice to enable general participation. In contrast, there are models based on the idea of fast and easy language acquisition. Clearly, joint schooling—most often found in primary schools—assumes that an L2 is learned from the input only and further language support is, at least at a young age, unnecessary. Models offering a certain amount of language support are scheduled for 6 to 18 months only, often restricted to 12 months due to limited resources. This time estimate—18 months suffice for building sufficient language knowledge to be able to participate in a regular class and to continue building-up L2 knowledge successfully—seems to be based on findings from studies of 2L1 and

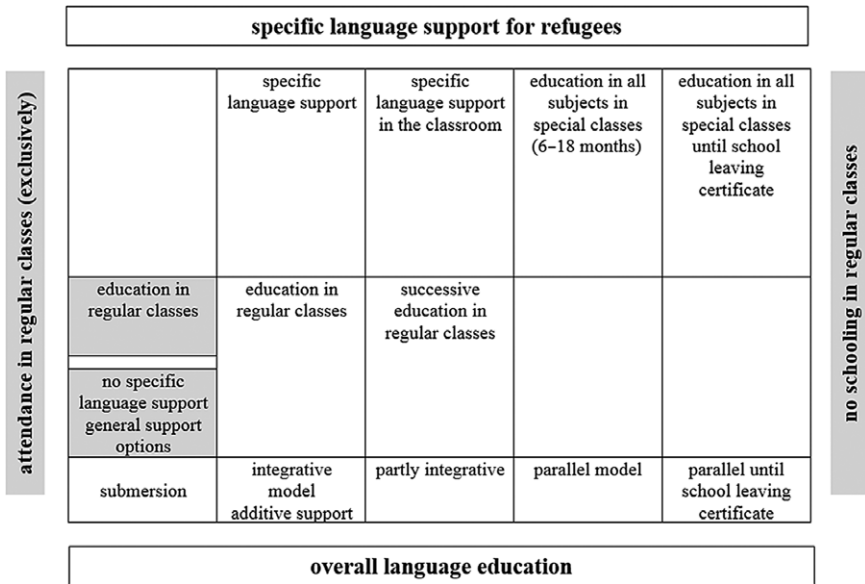


Figure 1. Models for schooling for refugees (“Seiteneinsteiger”) in Germany, adopted from Massumi and von Dewitz (2015, p. 45).

early child L2 reporting fast and sufficient language acquisition. However, more pessimistic estimates of the time needed to catch up in important areas of language are found in the recent literature (see Paradis & Jia, 2016; Schönenberger, Rothweiler, & Sterner, 2012) and schools report that they try to extend support as long as resources are available (Hertel, 2019).

The lack of assessment tools for this population is even more relevant as the German school system offers special support for children with language and learning difficulties, and thus requires language assessment of all at-risk children (Deml et al., 2018). Thus research is urgently needed not only for guiding the choice of assessment tools for measuring proficiency and identifying of language disorders but also for shaping expectations and models of participation in the education system.

**LITMUS Tools for Bilingual Language Assessment**

Recently developed LITMUS tools have been shown to identify developmental language disorder (DLD) in bilingual populations (Armon-Lotem, de Jong, & Meir, 2015). In previous work we focused on applying a quasi-universal LITMUS non-word repetition task (QU-NWRT; Grimm, Fèrre, dos Santos, & Chiat, 2014), allowing insights into the phonological abilities of a child, and a LITMUS sentence repetition task (SRT) in French and in German tapping morphosyntactic knowledge, and knowledge of computationally complex structures in particular (Fleckstein, Prévost, Tuller, Sizaret, & Zebib, 2018; Hamann, Chilla, Ruigendijk, & Abed Ibahim, 2013; Marinis & Armon-Lotem, 2015). Sentence and nonword repetition have a long tradition as tools for diagnosing DLD in monolinguals and have

been found to be highly accurate and reliable in this population (Conti-Ramsden, Botting, & Faragher, 2001). They were therefore selected for further cross-linguistic development (see Chiat, 2015; Marinis & Armon-Lotem, 2015), and several studies have since shown that well-constructed tasks have excellent diagnostic accuracy in bilingual populations, particularly when used in combination (Abed Ibrahim & Fekete, 2019; Armon-Lotem & Meir, 2016; Chiat & Polišenská, 2016; Hamann & Abed Ibrahim, 2017; Tuller *et al.*, 2018; Zebib, Tuller, Hamann, Abed Ibrahim, & Prévost, 2019). This is an encouraging result as the tasks are easy and fast to administer and could thus serve as a first assessment in many contexts, including schools.

Both repetition tasks have been associated with verbal working memory (VWM) and verbal short-term memory (VSTM) in so far as NWRTs usually are taken to measure VSTM (Archibald & Gathercole, 2007; Bishop, Adams, & Norbury, 2006) and VSTM has been argued to be the best predictor for performance in SRTs (Willis & Gathercole, 2001). Other authors have shown, however, that such tasks measure language skills, especially when they address phonological or morphosyntactic complexity (Gallon, Harris & van der Lely, 2007, for NWRT; Klem *et al.*, 2015; Meir, 2017; Polišenská, Chiat, & Roy, 2014, for SRT). Abed Ibrahim and Hamann (2017) confirmed that linguistic complexity has a decisive influence on performance in both the German LITMUS QU-NWRT and SRT. Moreover, investigation of bilingual children's performance on SRT in French has shown that differences of the typical group and the group with DLD can not only be explained by differences in VSTM or VWM but also crucially depend on syntactic complexity (Zebib *et al.*, 2019).

Previous studies have shown that the German LITMUS-NWRT and -SRT in particular, as well as the Arabic SRT (see Henry, Tuller, Prévost, & Zebib, *in press*), measure language abilities effectively. For NWRT, it has been shown that structures containing branching onsets or a coda are difficult for children with DLD (see dos Santos & Ferré, 2018). Such structures are included in the German QU-NWRT (here mostly referred to as NWRT), which otherwise uses vowels and consonants prevalent in many languages, resulting in the high accuracy of this task. As to the SRT, it incorporates morphosyntactic structures involving movement (*Wh*-questions), or embedding (finite complement clauses), or combinations of both (relative clauses). These structures have been found to be difficult cross-linguistically for children with language disorders (Friedmann & Novogrodsky, 2011; Hamann, Penner, & Lindner, 1998; Owen & Leonard, 2006; van der Lely, 1998). In addition, the German SRT contains structures identified as milestones in German L2 acquisition, such as the sentence bracket (where an auxiliary or modal in second position and a lexical verb in final position "bracket" one or more constituents) and topicalization (see Clahsen, 1986; Griefhaber, 2013; Haberzettl, 2005 or Wegener, 1992). Interestingly, topicalization or object relative clauses have also been associated with literacy so that their mastery can count as a relevant measure of language proficiency in typical school-age children. (For further details on the tasks, see Abed Ibrahim & Fekete, 2019; Hamann & Abed Ibrahim, 2017; Chilla *et al.*, *in press*).

In work by Tuller *et al.* (2018), performance in these two tasks was evaluated with respect to background information, gathered with the Questionnaire for

Parents of Bilingual Children (PaBiQ; Tuller, 2015). This LITMUS tool combines parts of the Alberta Language Development Questionnaire (Paradis, Emmerzael, & Duncan, 2010), with a focus on early language development and use, with parts of the Alberta Language Environment Questionnaire (Paradis, 2011), with a focus on current language input and use. Collecting background information about bilingual children concerning factors such as the age of the first words or first sentences in L1 (and L2), input quantity and quality, and language use (past and current) for both of a child's languages, is indispensable for evaluation of individual profiles of language development or for a diagnosis of a risk for language impairment (e.g., Restrepo, 1998). The information about input quantity, quality, and use is crucial also for calculating a (purely experiential) dominance index as a baseline for adjusting norms for L1 or L2 language tests standardized with monolingual norms (Thordardottir, 2015). In addition, the background information allows exploring factors such as early development, SES, AoO, and LoE, which have been discussed in the literature as influencing L1 and L2 performance. Tuller et al. (2018) established that early development, a risk factor for language disorders including age of first word and first sentence, and not language exposure and use, generally explained more of the variance in the performance in the German NWRT and SRT. SES, measured in years of mother's education, correlated with performance in the German SRT but was not further explored in the latter study.

Specifically for German, it was shown that the LITMUS-NWRT and -SRT have fair to good—in combination even excellent—diagnostic accuracy (Abed Ibrahim & Fekete, 2019; Chilla et al., *in press*; Hamann & Abed Ibrahim, 2017; Tuller et al., 2018). Cutoffs separating typical and atypical bilingual development were established for the SRT and the NWRT by various methods in the different studies arriving at very similar figures. However, due to the relative demographic stability in Germany up to September 2015, all these studies were based on heritage populations.

## The Present Study

In this investigation, we included a group of Syrian refugees with different educational backgrounds and compared their performance on the LITMUS repetition tasks mentioned above to that of Arabic heritage children. We expected our tasks to yield the following: (a) indicate if a child shows typical language development and can therefore be expected to encounter no extraordinary difficulties in her L2 development; and (b) ensure that, if the child has a language disorder, the child will be diagnosed and will have access to language therapy or the special support supplied by German schools.

In bilingual children, a reliable diagnosis is ideally based on assessment of both languages (Royal College of Speech and Language Therapists Specific Interest Group in Bilingualism, 2007; International Association of Logopedics and Phoniatrics, 2011), or at least on evaluation of the child's dominant language (Fredman, 2006). Therefore, we investigated both L2 and L1 performance in our groups of heritage and refugee children. In the case of recently arrived refugees,

the L1 should be the dominant language and thus, the presence of a language disorder should be identifiable in the L1 whereas a low performance in L2 would likely be due to short exposure. Heritage children, in contrast, might be at a disadvantage with L1 tests given the problems with L1 maintenance in heritage populations as discussed above. Comparing L1 development in refugee and heritage children will enable us to judge whether certain L1 tests are fair also to heritage children who might not be dominant in their L1 after extensive L2 exposure in preschool or school. As L1 measures may be crucial, we included both a standardized Arabic language assessment tool and as well as a LITMUS-SRT in Arabic (Zebib, Prévost, Tuller, & Henry, *in press*).

Standardized assessment tools for German are readily available for monolingual children, but either lack norms for bilinguals, do not cover the relevant age-range, or cannot be applied after short exposure to L2. This constitutes a well-known dilemma: as teachers or speech-language therapists capable of assessing the L1 in bilinguals are rare, in general German (L2) assessment tools have been used, and therefore the LITMUS tools were developed to improve this situation. As the German LITMUS SRT and QU-NWRT have shown promising results for heritage children, this study asks if the same tests also render reliable and fair assessments of language abilities for refugee children and explore how much language exposure is required to achieve fair results.

In addition, we want to know which factors influence individual variance in performance. Because it has been observed that performance on L2 repetition tasks can be influenced by age and input factors (AoO, LoE, SES, and current L1 and L2 input and use) as well as by working memory (verbal short term memory and verbal working memory) and language factors (lexicon and morphosyntax), we want to investigate the extent to which these factors influence performance in our participants (see Chiat & Polišínská, 2016, on properties of differently constructed NWRTs; and Meir, 2018, Tuller *et al.*, 2018; Zebib *et al.*, 2019, for detailed investigations of SRTs). In particular, we want to explore whether the tests allow fair assessment of emerging L2 abilities after an exposure of 18 months or more. Final points of investigation are in how far early and present language exposure/use in L1 or L2 determine language learning and whether L1 and/or L2 schooling leads to an advantage in L1 maintenance or L2 development (see Geva & Wiener, 2015; Rothman & Kupisch, 2018).

The present study addresses the following broad research question: can L1 and L2 LITMUS repetition tasks be used as measures for language ability with equal success with heritage and with refugee children? To answer this overall question, we asked the following more specific questions:

- a. How do heritage and refugee children compare in their Arabic skills on a standardized test and on a LITMUS-SRT for Arabic?
- b. How do heritage and refugee children compare in their performance on the German LITMUS-NWRT and -SRT?
- c. Which factors emerging from previous studies, such as AoO, LoE, early and current language input/use or quality of input, L1 or L2 schooling, SES, VSTM, VWM, or vocabulary and morphosyntactic knowledge can explain the variance of performance on the three LITMUS tasks? If LoE is a predictor,



we also ask how much exposure is necessary for the refugee group to perform in the range of typical heritage children.

## Materials and Methods

### Participants

Twenty-seven bilingual participants were recruited in kindergartens, schools, community associations, or places of worship from different federal states, representing a spectrum of schooling in the L2. Only children without risk for DLD and who scored above percentile rank 9 (IQ score  $\geq 80$  according to Wechsler's IQ scale) were included in the study (see the Procedures and Assessment Tools section). Twelve children (age range 6;0–12;9) with (Levantine) Arabic as heritage language were included. They were born in Germany, or came at a young age, and were systematically exposed to German upon preschool entry, that is, around age 3, resulting in simultaneous or early successive bilingualism. The heritage participants were compared to 15 refugee children from Syria (6;6–12;8), most of whom had their first exposure to German in special language classes with accompanying formal and literacy instruction in the L2 and had been attending such classes for at least 18 months (see Appendix A). The refugee children in our study represent individual examples of the diversity of L2 schooling programs since they were residing in different parts of Germany.

Background information on the participants was collected with the PaBiQ (Tuller, 2015; see below for details), which was augmented by questions about schooling, access to language courses, as well as transit itineraries, the means of transportation, and the past and present living conditions for the purposes of this study (see Appendix A). These additional questions were developed in close collaboration with J. Paradis and X. Cheng.

We provide some of this information here for our refugee children. Fifteen children from nine households are included in this sample, that is, there are six pairs of siblings, of which two pairs are twins. All families are legally recognized as refugees (see explanation in Appendix A), two with regular work permits now. The sample includes sufficient heterogeneity to mirror the situation of many learner groups: two of the children (siblings) count legally as unaccompanied minors since they came with family not including their parents. The families' trajectories differed widely: four families came on a regular flight and had been granted asylum before they boarded the plane; five families came by boat and subsequently spent up to 7 months in mass shelters. Even the children from families who came by plane and had private housing from the beginning varied with respect to schooling. The younger five children in our group started primary school in Germany, whereas nine of the older ones are "students with limited or interrupted formal education" (DeCapua & Marshall, 2011), with interruptions either in Syria or during transit. Eight of the nine children with previous schooling had literacy skills in Arabic and the majority of the children attended the age-appropriate grade in school at time of testing. One child, the unaccompanied minor, was rather old for first grade, however, and one child changed grades in school after 2 years.

### **Procedures and assessment tools**

All participants, heritage and refugee children, were tested with a comprehensive assessment procedure (see Appendix B), including standardized tests in the L1 and L2 respecting dominance effects on test performance, see Hamann & Abed Ibrahim, 2017; Tuller *et al.*, 2018, for particulars. Following procedures established by the above authors, a child from the heritage group was regarded as language impaired if she scored below adjusted norms in two language domains in each of the languages. Adjustment of monolingual norms was performed following Thordardottir's (2015) recommendations and by carefully establishing language dominance. For our refugee group we did not base clinical classification on L2 tests, however, but relied on a comprehensive, standardized L1 test. The German versions of the LITMUS-NWRT and -SRT, as well as the Arabic LITMUS-SRT were administered to all children. Cognitive and working-memory tasks as well as a parental questionnaires (PaBiQ) were also administered.

#### *Standardized L1 and L2 tests*

The lack of age appropriate, standardized tools posed a serious problem for background assessment of our participants, which include children up to the age of 12 years and is one of the reasons for developing new assessment methods. Nevertheless, the following standardized language tests were administered and provide potentially predictive language measures.

*Standardized L1 tests.* For assessing Arabic abilities and for determining typical development for the refugee children in particular, we used the Batterie d'Evaluation du Langage Oral chez l'enfant libanais (ELO-L; Zebib *et al.*, 2017; Appendix B). There are versions for younger (3;0–5;11) and older (6;0–7;11) children. Both versions were normed in Lebanon on large samples of children growing up in the kind of institutional bilingualism typical for Lebanon. The test was translated and adapted to other varieties of Arabic by linguistically trained native speakers of these varieties, and specific care with respect to vocabulary and phonological characteristics went into the creation of the Syrian Arabic version used here. As the age-range of the norming population shows, the test was well suited for our younger children (Appendix B). Norm-referencing was not applicable to a subset of our sample, however.

*Standardized L2 tests.* We encountered similar difficulties for some standardized L2 tests. Vocabulary assessment by the Wortschatz- und Wortfindungstest für 6- bis 10-Jährige (WWT; Glück, 2011) provides age-appropriate norms for monolinguals (5;6–10;11). In line with other researchers, such as Montanari (2017), we considered raw values for meaningful comparisons. For assessment of morphosyntax, we used the LiSe-DaZ (Schulz & Tracy, 2011) and in addition, for the group of refugees, the Test for Reception of Grammar–German (TROG-D; Fox, 2009, an adaptation of TROG; Bishop, 1989).

The TROG-D tests comprehension of words from different word classes and subsequently targets the comprehension of morphological (perfect or plural) or structural (passive, finite complement clauses, relative clauses, adverbial clauses, or topicalization) grammatical knowledge. Because of ceiling effects, test results can

only be reliably evaluated in the age range of 3;0–9;11. As there is no norming sample matching our oldest children, raw values/percentages were used in addition to comparisons to the oldest age range with available norms.

The LiSe-DaZ, a morphosyntactic test normed for early successive bilinguals as well as monolinguals, provides norms for early child L2 learners, for the ages 3;0 to 7;11, and for monolinguals from 3;0 through 6;11. The test addresses core areas of morphosyntax assessing comprehension of negation, constituent (*Wh*)-questions and verb semantics, as well as production of subject–verb agreement, complex sentences, case marking, and of selected lexical items. For the simultaneous bilinguals in the heritage group, monolingual or bilingual norms can be used according to a child’s dominance. Despite the possibility of norm extension for bilingual children older than 7;11 (Grimm & Schulz, 2014), norms cannot be adapted in any way for our population of refugees given the minimum LoE required for the norms of each age range. Therefore, we used norm extensions as far as possible and also compared raw scores of subtests such as “subject-verb-agreement” or “developmental level.”

#### *LITMUS-QU-NWRT and LITMUS- SRT*

*The German LITMUS-QU-NWRT.* The quasi-universal NWRT chosen for this study was constructed by Grimm et al. (2014) for use in Germany and France and relies on increasing phonological complexity, not increasing numbers of syllables (dos Santos & Ferré, 2018; Grimm & Hübner, *in press*). Using the most common vowels and consonants of the world’s languages, it presents one-, two- and three-syllable nonwords of different phonological complexity with complex onsets or codas and combinations thereof. There is a language independent part that uses consonant clusters common in most languages, and a language dependent part that integrates the combination of /s/ and an obstruent in word initial and final positions as a language specific property of German (see Abed Ibrahim & Fekete, 2019; Grimm & Hübner, *in press*; Hamann & Abed Ibrahim, 2017, for more detailed descriptions of the task’s properties). The items are presented to the child in an appealing power point presentation in pseudorandomized order through headphones. The task takes about 5–10 min to administer and is scored according to whole item accuracy. For this scoring method, minimally different vowels and, crucially, voicing of consonants are not counted as errors.

*The German and the Arabic LITMUS-SRT.* The German SRT was first introduced by Hamann et al. (2013) and the Arabic SRT by Henry et al. (*in press*) for application in Lebanon. Both SRTs originally targeted children between 5;6 and 9 years of age. This version of the Arabic SRT was adapted (lexically and phonologically) to Syrian Arabic by Syrian speakers and Arabic linguists at the University of Oldenburg. Both SRTs include complex structures described as difficult for children with language impairment in the cross-linguistic literature. In addition to simple sentences varying in agreement and/or tense, Marinis and Armon-Lotem (2015) recommend including object (which-) questions, passives, finite complement clauses, (object) relative clauses, and topicalizations, which are all included in the German version (for details and examples see Hamann & Abed Ibrahim, 2017; Tuller et al., 2018). Not all these structures are available in all languages, however, so that the version of the Arabic SRT used in this study includes simple perfective and imperfective

sentences, object which-questions, nonfinite and finite complement clauses, as well as subject and object relatives, but no passives or topicalizations (see Zebib *et al.*, [in press](#), for examples and a more detailed overview). The German SRT in addition includes the sentence bracket, a structure typical for German in which objects and adverbials occur after an auxiliary or modal and before the lexical verb in final position.

The current versions of the German and the Arabic SRT include 45 and 36 sentences, respectively (the latter in parallel to the French SRT). The stimuli are presented via a child friendly power point presentation in randomized order. Administration of the tasks takes 5 to 10 min. The tasks can be rated by “identical repetition” counting only exact repetition as correct, disregarding only phonological errors, or it can be rated by “target structure,” where mastery of the structure is counted as correct even if the child substituted lexical items or used the incorrect gender, or in some instances where it does not change the structure, incorrect case. Both are used in this study.

#### *Standardized tests for cognitive development and working memory measures*

For a rough estimate of cognitive development, we measured nonverbal intelligence with the German version of Raven’s Colored Progressive Matrices (Bulheller & Häcker, 2002). We also chose two working memory measures, the Forward Digit Span (FDS; Petermann & Petermann, 2011), which is associated with VSTM, and the Backward Digit Span (BDS; Petermann & Petermann, 2011), which involves storage and processing and assesses VWM.

#### *The parental questionnaire PaBiQ*

The LITMUS–Questionnaire for Parents of Bilingual Children (PaBiQ; Tuller 2015, see also the LITMUS Tools for Bilingual Assessment section) was used to elicit information about age and input variables. The following variables were considered for the current study: chronological age, SES (as measured by of maternal education in years), AoO, LoE, early L1 and L2 exposure (measured by frequency of early language use and exposure and diversity of exposure contexts before the age of 4), current L1 and L2 use (as measured by the proportion of L1 and L2 use within the family), the richness of the L1 and L2 language environment, and length of L1 and L2 schooling measured in months. The PaBiQ was administered orally during an interview with the parents/legal guardians of the participating children in their language of preference. The version administered to the refugee sample contained additional questions about schooling and the family’s migration history, access to language courses, as well as past and present living conditions. These additional questions were developed in close collaboration with J. Paradis and X. Cheng. Only the information about schooling is relevant here for further exploration.

#### **Data analysis**

All standardized tests, as well as the German LITMUS-QU-NWRT and the German and Arabic LITMUS-SRT and PaBiQ, were administered as per test instructions

(cf. Hamann & Abed Ibrahim, 2017; Tuller et al., 2018; Zebib et al., *in press*). As already indicated, norm-referencing was possible only for certain age groups in the different standardized tests. For L1, we expect ceiling effects on the ELO-L for our older refugee population showing their basic familiarity with Arabic whereas older heritage children may stagnate in their L1 due to intensified exposure to German in school. For L2, it can be expected that, due to inadequate exposure, tools normed for younger populations can still be challenging for older refugees. In both cases, we consider it meaningful if we find that a child scores lower than the adjusted norm in the oldest age range with an available norm. All the same, basic comparisons will be performed on raw values (percentage-correct scores) as well.

Children's SRT and NWRT responses were recorded with special dictaphones. Data were transcribed, verified, and coded offline independently by two linguistically trained raters (percentage of agreement: at least 90%). Data analysis was performed on the percentage of correct responses for each repetition measure. Null reactions counted as errors, unless they were caused by technical problems or errors of the investigators (Abed Ibrahim & Fekete, 2019, p.14).

IBM SPSS 24 (IBM Corp. Released, 2016) was used to conduct the statistical analyses. Nonparametric tests were used for group comparisons due to small sample sizes and the violation of the normality assumption of parametric tests, revealed by the Shapiro–Wilk test. Between-group pairwise comparisons were carried out using the Mann–Whitney *U* test. Friedmann and Wilcoxon signed rank test with Bonferroni-adjustment were used for within-group comparisons.

To determine which factors influenced performance on the LITMUS repetition tasks, we conducted hierarchical regression analyses. The reason for using hierarchical regression analysis was to investigate the influence of selected independent variables on performance after controlling for other variables. Changes in  $R^2$  were computed between two models, and the improvement in  $R^2$  was checked for significance. We report the standardized regression coefficients, the  $R^2$  values and the *p* values. Models with multiple independent variables were checked for multicollinearity using the variance inflation factor to ensure that they were all within accepted limits of a variance inflation factor value of 10 (Hair, Anderson, Tatham, & Black, 1995).

Age and input variables, as well as language/working memory variables, shown to influence performance on repetition tasks in previous research were tested for correlation with performance on the LITMUS repetition tasks (dependent variables). Only those yielding significant moderate to strong bivariate Spearman correlations with the dependent performance variables (see Table 1 and 2) were selected as potential predictors, which were later entered as independent variables into hierarchical regression models. Given the small sample size ( $N = 27$ ), the maximal number of blocks entered into the hierarchical regression models was limited to three with one independent variable per block. Recent research suggests that 2 subjects per independent variable are needed minimally in linear regression analyses (Austin & Steyerberg, 2015). Thus, having 27 subjects and 3 independent variables is considered to be appropriate in the light of recent statistical research. The order of entry into the blocks was based on the strength of the correlation (as measured by the correlation coefficient) between the independent variable and the dependent performance measure.

**Table 1.** Spearman correlations between PaBiQ age and input variables and performance in German/Arabic LITMUS-SRTs

| Background variables (PaBiQ) | German SRT     |                | Arabic SRT |            |
|------------------------------|----------------|----------------|------------|------------|
|                              | SRT Id         | SRT Tar        | SRT Ar Id  | SRT Ar Tar |
| Age                          | -.285          | -.364          | .240       | .283       |
| SES                          | <b>.433*</b>   | <b>.485*</b>   | .338       | .203       |
| AoO L2                       | <b>-.534**</b> | <b>-.513**</b> | -.201      | .204       |
| LoE L2                       | <b>.601***</b> | <b>.619***</b> | .276       | -.251      |
| Early L1 exposure            | -.136          | -.108          | -.098      | .343       |
| Early L2 exposure            | .151           | .217           | .213       | -.257      |
| Current L1 use               | <b>-.599**</b> | <b>-.536**</b> | -.377      | .075       |
| Current L2 use               | <b>.723**</b>  | <b>.609**</b>  | .281       | -.087      |
| L1 richness                  | -.155          | -.214          | -.074      | .170       |
| L2 richness                  | .211           | .261           | .042       | -.315      |
| Length of L1 schooling       | .172           | .089           | .153       | .338       |
| Length of L2 schooling       | <b>.482*</b>   | <b>.409*</b>   | .327       | .075       |

Notes: Heritage and refugee children were collapsed together for correlational analyses. SRT Id, German LITMUS-SRT “identical repetition.” SRT Tar, German LITMUS-SRT “target structure.” SRT Ar Id, Arabic LITMUS-SRT “identical repetition.” SRT Ar Tar, Arabic LITMUS-SRT “target structure.” \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

**Table 2.** Spearman correlations between German/Arabic LITMUS-SRTs, working memory and standardized L1 and L2 measures

|                                     | German SRT     |                | Arabic SRT    |               |
|-------------------------------------|----------------|----------------|---------------|---------------|
|                                     | SRT Id         | SRT Tar        | SRT Ar Id     | SRT Ar Tar    |
| WM variables                        |                |                |               |               |
| Forward Digit Span (FDS)            | <b>.580**</b>  | <b>.673***</b> | <b>.408*</b>  | .061          |
| Backward Digit Span (BDS)           | .267           | .190           | .262          | .317          |
| Linguistic variables (L2)           |                |                |               |               |
| WWT LexP (L2 expressive vocabulary) | <b>.725***</b> | <b>.746***</b> | n/a           | n/a           |
| WWT LexR (L2 receptive vocabulary)  | <b>.729***</b> | <b>.722***</b> | n/a           | n/a           |
| TROG-D (morphosyntax comp.)         | <b>.894***</b> | <b>.854***</b> | n/a           | n/a           |
| LexP ELO-L                          | n/a            | n/a            | .094          | <b>.424*</b>  |
| LexR ELO-L                          | n/a            | n/a            | .287          | .096          |
| MorphP ELO-L                        | n/a            | n/a            | .153          | <b>.444*</b>  |
| MorphR ELO-L                        | n/a            | n/a            | <b>.601**</b> | <b>.657**</b> |
| Phon ELO-L                          | n/a            | n/a            | <b>.442*</b>  | <b>.442*</b>  |

Notes: Heritage and refugee children were collapsed together for correlational analyses. SRT Id, German LITMUS-SRT “identical repetition.” SRT Tar, German LITMUS-SRT “target structure.” SRT Ar Id, Arabic LITMUS-SRT “identical repetition.” SRT Ar Tar, Arabic LITMUS-SRT “target structure.” \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

## Results

### Comparisons of background variables between the heritage and refugee groups

The information gathered from the PaBiQ allows comparisons of the two groups of participants. Table 3 gives a summary for such background values. From input and use information, we also calculated a language dominance index (see Abed Ibrahim & Fekete, 2019; Tuller et al., 2018) that influenced our norm adjustments where available.

We first compared the two bilingual groups in terms of nonlanguage variables (age and SES), bilingualism related variables (AoO, LoE, early L1/L2 exposure, current L1/L2 exposure and richness, and L1 schooling), and working memory measures (VSTM: FDS; VWM: BDS). Exposure and use as well as richness variables were preferred over language dominance because they contribute to the latter, but allow more fine-grained analysis when considered separately (Tuller et al., 2018). Between-group comparisons showed that both groups were comparable in terms of age ( $U = 89.0, p = .961$ ) and SES, as measured by maternal education in years ( $U = 75.5, p = .474$ ). Importantly, the refugee children started to acquire German later in childhood (AoO,  $U = 7.00, p < .001, r = .781$ ), had a shorter LoE to the L2 (LoE,  $U = 5.00, p < .001, r = .799$ ), and differed from the heritage children in the amount of early and current exposure to the L2 (early L2 exposure:  $U = 16.0, p < .001, r = .782$ ; current L2 use:  $U = 32.5, p = .004, r = .547$ ).

**Table 3.** Overview of background variables in the two participant groups

| Mean (SD)   | Heritage<br>( $n = 12$ ) | Refugees<br>( $n = 15$ ) |
|---|--------------------------|--------------------------|
| Age at testing (mo.)  | 114.5 (24.9)             | 114.27 (24.5)            |
| Age of onset (mo.)  | 37.91 (13.08)            | 87.67 (25.17)            |
| Length of exposure (mo.)  | 76.83 (30.01)            | 26.6 (7.2)               |
| Early exposure L1 (%)   | 86% (6.3)                | 100% (0)                 |
| Early exposure L2 (%)   | 46% (2.3)                | 0%                       |
| Language use L1 (/16)   | 9.00 (3.25)              | 12.1 (2.59)              |
| Language use L2 (/16)   | 6.33 (2.06)              | 3.07 (2.52)              |
| Current L1 richness (/14)   | 4.25 (2.09)              | 7.07 (1.79)              |
| Current L2 richness (/14)   | 9.42 (2.50)              | 7.2 (2.85)               |
| Length L1 schooling (mo.)   | 1.91 (2.11)              | 11.5 (11.4)              |
| Length L2 schooling (mo.)   | 23.6 (17.5)              | 14.3 (9.63)              |
| Socioeconomic status (SES)<br>(years of mother's education)<br>(maternal education in years)              | 13.3 (2.7)               | 14.3 (5.34)              |
| Language dominance index, LDI<br>( $< -5 = L1$ dominant; $-5$ – $+5 =$ balanced;<br>$> +5 = L2$ dominant) | -5.42 (8.67)             | -24.63 (8.24)            |

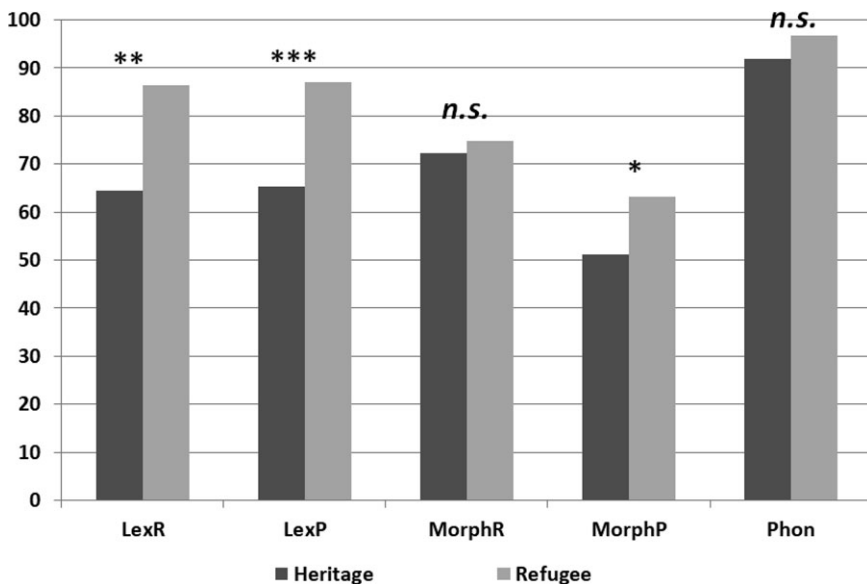
The heritage group, in contrast, had less early L1 exposure than their refugee peers ( $U = 22.5, p < .0001, r = .782$ ), used Arabic to a lesser extent in everyday communication ( $U = 44.0, p = .023, r = .437$ ), and had a less enriched input in the L1 ( $U = 26.5, p = .002, r = .605$ ). These differences in AoO, LoE, as well as in early and current patterns of exposure to L1/L2 lead to a significant difference in the degree of language dominance between the two groups ( $U = 11.0, p < .001, r = .743$ ), where the refugee group was much more dominant in the L1.

No significant differences emerged between the heritage and refugee groups with respect to the amount of L1 schooling (refugee:  $M = 11.5, SD = 11.4$ ; heritage:  $M = 2.58, SD = 2.64, U = 75.0, p = .232$ ) or L2 schooling (refugee:  $M = 14.3, SD = 9.6$ ; heritage:  $M = 23.6, SD = 17.5, U = 52.5, p = .163$ ). Concerning working memory measures, a significant between-group difference was observed for VSTM (FDS:  $U = 43.0, p = .035, r = .405$ ) but not for VWM (BDS:  $U = 76.5, p = .760$ ).

### **Comparisons of standardized measures in the L1 and L2 between the heritage and refugee groups**

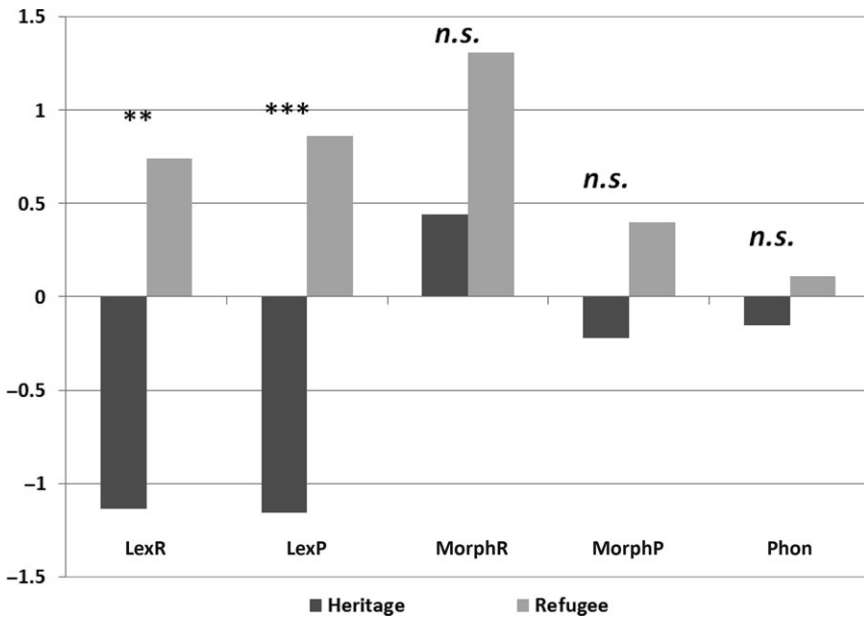
#### *Standardized L1 test (ELO-L)*

For the L1, the refugee group showed superior lexical skills, both for receptive (raw scores: refugee:  $M = 87.3, SD = 10.6$ ; heritage:  $M = 64.5, SD = 17.3, U = 31.0, p = .004, r = .557$ ; Z scores: refugee:  $M = 0.811, SD = 0.933$ ; heritage:  $M = -1.14, SD = 1.46, U = 30.5, p = .004, r = .561$ ), and expressive vocabulary



**Figure 2.** Performance in the subdomains of the standardized Arabic test ELO-L (raw scores, percentage correct). LexR, receptive vocabulary. LexP, expressive vocabulary. MorphR, morphosyntax comprehension. MorphP, morphosyntax production. Phon, phonology.





**Figure 3.** Performance in the subdomains of the standardized Arabic test ELO-L (Z scores). LexR, receptive vocabulary. LexP, expressive vocabulary. MorphR, morphosyntax, comprehension. MorphP, morphosyntax production. Phon, phonology.

(raw scores: refugee:  $M = 87.6$ ,  $SD = 5.44$ ; heritage:  $M = 65.24$ ,  $SD = 20.1$ ,  $U = 20.5$ ,  $p < .001$ ,  $r = .653$ ; Z scores: refugee:  $M = 0.924$ ,  $SD = 0.557$ ; heritage:  $M = -1.16$ ,  $SD = 1.99$ ,  $U = 20.0$ ,  $p = .001$ ,  $r = .658$ ). Furthermore, the refugee children outperformed their heritage peers on morphosyntax production, when raw scores were considered (raw scores: refugee:  $M = 63.1$ ,  $SD = 8.77$ ; heritage:  $M = 51.1$ ,  $SD = 12.1$ ,  $U = 42.0$ ,  $p = .019$ ,  $r = .453$ ; Z scores: refugee:  $M = 0.380$ ,  $SD = 0.782$ ; heritage:  $M = -0.219$ ,  $SD = 1.64$ ,  $U = 67.0$ ,  $p = .261$ ). In contrast, no significant between-group comparisons were found for morphosyntax comprehension (raw scores: refugee:  $M = 73.3$ ,  $SD = 12.5$ ; heritage:  $M = 72.2$ ,  $SD = 13.9$ ,  $U = 76.0$ ,  $p = .492$ ; Z scores: refugee:  $M = 1.20$ ,  $SD = 4.13$ ; heritage:  $M = 0.440$ ,  $SD = 0.762$ ,  $U = 70.5$ ,  $p = .340$ ), or in phonology (raw scores: refugee:  $M = 96.4$ ,  $SD = 5.23$ ; heritage:  $M = 91.0$ ,  $SD = 8.89$ ,  $U = 79.0$ ,  $p = .562$ ; Z scores: refugee:  $M = 0.09$ ,  $SD = 1.01$ ; heritage:  $M = -0.154$ ,  $SD = 1.70$ ,  $U = 69.0$ ,  $p = .295$ ). Mean raw scores (%correct) and Z scores for the ELO-L subtests are given in Figures 2 and 3.

#### Standardized L2 tests

With regard to L2 lexical abilities, both groups performed poorly in both expressive (raw scores: refugee:  $M = 18.9$ ,  $SD = 14.9$ ; heritage:  $M = 26.7$ ,  $SD = 19.1$ ,  $U = 53.5$ ,  $p = .115$ ; PR: refugee:  $M = 4.43$ ,  $SD = 7.94$ ; heritage:  $M = 6.33$ ,  $SD = 13.8$ ,  $U = 80.5$ ,  $p = .841$ ) and receptive subparts of the vocabulary test (WWT; raw scores: refugee:  $M = 72.7$ ,  $SD = 22.6$ ; heritage:  $M = 76.46$ ,  $SD = 15.4$ ,  $U = 62.0$ ,  $p = .256$ ; PR: refugee:  $M = 11.4$ ,  $SD = 14.2$ ; heritage:  $M = 9.1$ ,  $SD = 15.8$ ,  $U = 76.5$ ,  $p = .691$ ). Most

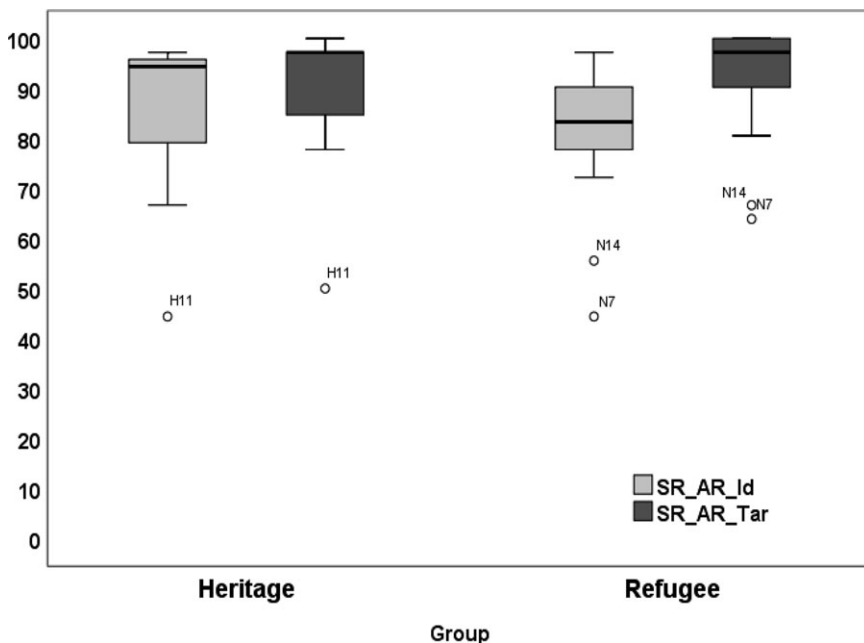
children performed within the language impaired range in L2 vocabulary if individual dominance-adjusted standardized scores were considered (Thordardottir, 2015), even on the receptive measure.

Considering morphosyntax, none of the heritage children would be considered as having DLD (using norm extensions, see Grimm & Schulz, 2014) and all showed age-appropriate developmental stages in the LiSe-DaZ subtest for production of sentence complexity; in contrast, more than a third of the refugee children (6/15), namely, all those with less than 24 months of exposure, would be classified as at risk for DLD even if compared to the norms valid for younger early successive bilingual children. Similar results were observed for performance of the refugee group on TROG-D applying dominance-adjusted cutoffs: all children with an LoE <24 months (6/15) would be identified as at risk of DLD using this test.

### **LITMUS repetition tasks**

#### *Arabic LITMUS-SRT*

As can be seen in Figure 4, heritage children were not disadvantaged by the Arabic LITMUS-SRT, unlike in the standardized L1 measures of the ELO-L. Both groups generally scored better when the measure “target structure” was applied. Although the heritage group displayed more variance on “identical repetition” (refugee:  $M = 80.8$ ,  $SD = 14.2$ ; heritage:  $M = 84.8$ ,  $SD = 16.7$ ) than on “target structure” (refugee:  $M = 92.0$ ,  $SD = 12.2$ ; heritage:  $M = 89.1$ ,  $SD = 15.2$ ), no significant between-



**Figure 4.** Arabic LITMUS-SRT: percentage correct identical repetition and target structure. SRT AR Id, Arabic LITMUS-SRT “identical repetition.” SRT AR Tar, Arabic LITMUS-SRT “target structure.”

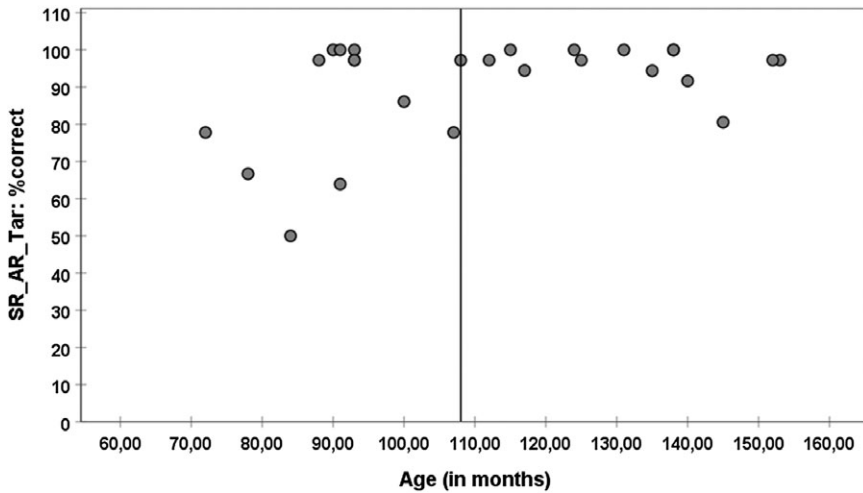


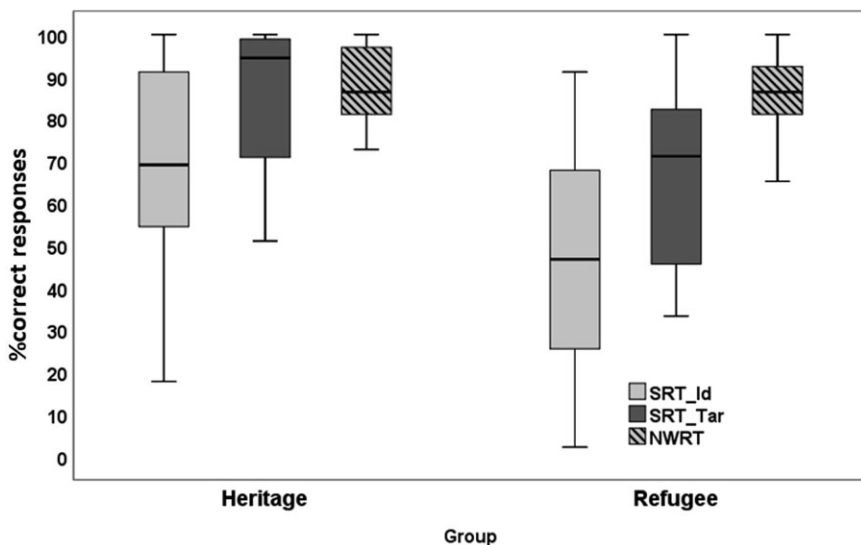
Figure 5. Performance in Arabic LITMUS-SRT: percentage correct target structure plotted against chronological age. SRT AR Tar, Arabic LITMUS-SRT “target structure.”

group differences were found for either “identical repetition” ( $U = 62.5, p = .178$ ) or “target structure” ( $U = 70.0, p = .320$ ). Ceiling effects were observed for “target structure”: 20/27 children performed above 90% correct and 12 of these were older than 9 years (see Figure 5).

#### German LITMUS-QU-NWRT and LITMUS-SRT

Comparing bilingual heritage and refugee children for global performance in German LITMUS-NWRT and -SRT (Figure 6), little variance with no significant group differences were observed, for LITMUS-NWRT (heritage:  $M = 89.9, SD = 9.78$ ; refugee:  $M = 86.5, SD = 9.32$ ;  $U = 70.5, p = .340$ ). As for the German SRT, the heritage group performed better and showed less variance, especially on “target structure” (SRT Id (identical repetition: heritage:  $M = 68.0, SD = 28.4$ ; refugee:  $M = 47.6, SD = 28.9, U = 50.0, p = .051, r = .376$ ; SRT Tar (target structure): heritage:  $M = 84.3, SD = 19.6$ ; refugee:  $M = 66.1, SD = 22.8, U = 45.5, p = .029, r = .419$ ).

With regard to within-group comparisons on performance in “identical repetition” and “target structure” in the SRT and in NWRT, the Friedman test results were significant for both groups, heritage:  $\chi^2(2) = 38.17, p = .017$ ; refugee:  $\chi^2(2) = 27.8, p < .001$ . However, pairwise comparisons (Wilcoxon signed ranks tests applying Bonferroni correction) of the performance of the heritage group revealed a significant difference between NWRT and “identical repetition” in the SRT ( $Z: -2.51, p = .036, r = .724$ ) but not between NWRT and SRT “target structure” ( $Z: -0.157, p > .90$ ). Refugee children, in contrast, performed significantly better on the NWRT compared to both “identical repetition” ( $Z: -3.41, p = .003, r = .879$ ) and “target structure” in the SRT ( $Z: -3.17, p = .006, r = .818$ ).



**Figure 6.** Performance in German LITMUS-NWRT and -SRT (SRT Id and SRT Tar): percentage correct responses. NWRT, German LITMUS-nonword repetition task. SRT Id, German LITMUS-SRT “identical repetition.” SRT\_Tar, German LITMUS-SRT “target structure.”

### **Predictors of performance on LITMUS repetition tasks**

Given the small linguistic load in the NWRT compared to the SRTs and the results from previous studies about robustness of this measure against exposure variables (Abed Ibrahim & Fekete, 2019; Tuller *et al.*, 2018), analysis of sources of individual variation in the present study was limited to the Arabic and German LITMUS-SRTs.

As outlined in the Data Analysis section, hierarchical linear regression modeling was used to examine which age and input, working memory (FDS and BDS) and linguistic variables (i.e., performance in the standardized L1/L2 tests), can predict performance in the Arabic and German LITMUS-SRTs investigated in this study. Modeling was done with the percentages of correct “identical repetition” and correct “target structure” for each of the German and Arabic LITMUS-SRTs as the dependent variables.

Due to their significant moderate to strong correlations with the dependent measures (see Table 1 and 2), the following factors were considered for regression analysis as potential predictor variables explaining the variance in performance: AoO L2, LoE L2, current L1/L2 use, L2 schooling, and SES as age and input variables; FDS as the only working memory variable; and scores of expressive and receptive vocabulary, morphosyntax, and phonology on standardized L1 and L2 tests as linguistic variables. Given the small sample size and as combined effects of age/input and linguistic/working memory variables on performance are not under investigation in this study, separate regression models were created to examine the influence of only age/input variables and only linguistic/working memory variables on performance in the repetition tasks.

*Predictors of performance on the Arabic LITMUS-SRT*

*Age and input factors.* Regarding performance on the Arabic LITMUS-SRT, no significant correlations emerged between performance in Arabic LITMUS-SRT and any of the age and input variables AoO L2, LoE L2, early L1/L2 exposure, current L1/L2 language use, and L1/L2 linguistic richness, SES and L1 schooling (see Table 1). However, visual inspection of Figure 5 indicates that age effects across groups are responsible for the observed ceiling performance on Arabic LITMUS-SRT when scored by correct target structure.

*Linguistic and working memory variables.* As to working memory and linguistic variables, significant correlations were found between Arabic LITMUS-SRT (identical repetition) and L1 receptive morphosyntax, L1 phonology, as well as FDS. Importantly, only L1 receptive morphosyntax emerged as a significant predictor for performance on “identical repetition” in the Arabic SRT, accounting for 39.6% of the variance (see Table 4). Both FDS and L1 phonology were not significant in the final step of the model.

As can be seen in Table 2, the measure “target structure” of the Arabic SRT was significantly correlated with L1 receptive morphosyntax, L1 morphosyntax production, L1 phonology, and L1 expressive vocabulary. Two different three-step hierarchical regression models were built to explain performance variance for the measure “target structure” in the Arabic SRT. The two independent variables yielding the strongest correlations, that is, L1 receptive morphosyntax and L1 morphosyntax production, were entered into Blocks 1 and 2 (in the order of strength of correlation) and were kept constant in both models. The other variables with significant correlations, that is, L1 phonology and L1 expressive vocabulary, were alternatively entered into the

**Table 4.** Summary of hierarchical regression analysis for working memory and linguistic variables predicting performance in Arabic LITMUS-SRT scored by identical repetition

|                           | <i>b</i> | <i>SE b</i> | $\beta$ |
|---------------------------|----------|-------------|---------|
| <i>Step 1</i>             |          |             |         |
| Constant                  | 36.67    | 12.36       |         |
| L1 receptive morphosyntax | 0.643    | 0.166       | .629*** |
| <i>Step 2</i>             |          |             |         |
| Constant                  | -22.84   | 31.89       |         |
| L1 receptive morphosyntax | 0.613    | 0.156       | .600*** |
| L1 phonology              | 0.646    | 0.322       | .307    |
| <i>Step 3</i>             |          |             |         |
| Constant                  | -25.25   | 30.95       |         |
| L1 receptive morphosyntax | 0.538    | 0.159       | .527**  |
| L1 phonology              | 0.598    | 0.314       | .284    |
| FDS                       | 2.33     | 1.49        | .243    |

Note:  $R^2 = .396$  for Step 1;  $\Delta R^2 = .093$  for Step 2 ( $p = .058$ );  $\Delta R^2 = .053$  for Step 3 ( $p = .135$ ). \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

**Table 5.** Summary of hierarchical regression analysis for working memory and linguistic variables predicting performance in Arabic LITMUS-SRT scored by correct target structure

|                          | Model 1  |             |         | Model 2                  |             |         |         |
|--------------------------|----------|-------------|---------|--------------------------|-------------|---------|---------|
|                          | <i>b</i> | <i>SE b</i> | $\beta$ | <i>b</i>                 | <i>SE b</i> | $\beta$ |         |
| <i>Step 1</i>            |          |             |         |                          |             |         |         |
| Constant                 | 38.55    | 10.82       |         | Constant                 | 38.55       | 10.82   |         |
| L1 receptive morphosyn.  | 0.717    | 0.146       | .707*** | L1 receptive morphosyn.  | 0.717       | 0.146   | .707*** |
| <i>Step 2</i>            |          |             |         |                          |             |         |         |
| Constant                 | 24.09    | 10.96       |         | Constant                 | 24.09       | 10.96   |         |
| L1 receptive morphosyn.  | 0.567    | 0.141       | .560*** | L1 receptive morphosyn.  | 0.567       | 0.141   | .560*** |
| L1 morphosyn. production | 0.435    | 0.159       | .380*   | L1 morphosyn. production | 0.435       | 0.159   | .380*   |
| <i>Step 3</i>            |          |             |         |                          |             |         |         |
| Constant                 | -23.27   | 22.11       |         | Constant                 | 19.67       | 11.06   |         |
| L1 receptive morphosyn.  | 0.561    | 0.128       | .554**  | L1 receptive morphosyn.  | 0.560       | 0.137   | .552*** |
| L1 morphosyn. production | 0.254    | 0.163       | .222    | L1 morphosyn. production | 0.244       | 0.200   | .213    |
| L1 phonology             | 0.615    | 0.256       | .322*   | L1 expressive vocabulary | 0.204       | 0.134   | .254    |

Note: Model 1:  $R^2 = .500$  for Step 1:  $\Delta R^2 = .122$  for Step 2 ( $p = .012$ ):  $\Delta R^2 = .078$  for Step 3 ( $p = .025$ ). Model 2:  $R^2 = .500$  for Step 1:  $\Delta R^2 = .122$  for Step 2 ( $p = .012$ ):  $\Delta R^2 = .036$  for Step 3 ( $p = .143$ ). \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

third block of each of the hierarchical regression models. As can be seen in Table 5, a sizable proportion of the variance in performance (50%) was accounted for by L1 receptive morphosyntax. L1 morphosyntax production explained an additional 12.2% of the variation in Step 2; however, adding L1 phonology or L1 expressive vocabulary in the final step of Models 1 and 2 rendered L1 morphosyntax production not significant. While L1 phonology accounted for an additional 7.8% of the variance in the first model, no significant effects were observed for expressive L1 vocabulary in the final step of the second regression model.

#### *Predictors of performance on the German SRT*

*Age and input factors.* In order to determine which of the significantly correlated age and input variables predicted performance in “identical repetition” and “target structure” (see Table 1), we built four different three-step hierarchical regression models for each of the dependent variables “identical repetition” and “target structure”. Since current L2 use and LoE L2 yielded the strongest correlations with performance on “identical repetition” and “target structure,” they were entered into Blocks 1 and 2 (in the order of strength of correlation) and were kept constant in all four models. The remaining four variables with significant moderate

**Table 6.** Summary of hierarchical regression analysis for age and input factors predicting performance in German LITMUS-SRT scored by identical repetition

|                | Model 1  |             |         | Model 2        |             |         |         |
|----------------|----------|-------------|---------|----------------|-------------|---------|---------|
|                | <i>b</i> | <i>SE b</i> | $\beta$ | <i>b</i>       | <i>SE b</i> | $\beta$ |         |
| <i>Step 1</i>  |          |             |         |                |             |         |         |
| Constant       | 23.24    | 8.12        |         | Constant       | 23.24       | 8.12    |         |
| Current L2 use | 7.27     | 1.51        | .694*** | Current L2 use | 7.27        | 1.51    | .694*** |
| <i>Step 2</i>  |          |             |         |                |             |         |         |
| Constant       | 17.59    | 8.33        |         | Constant       | 17.59       | 8.33    |         |
| Current L2 use | 5.35     | 1.77        | .510**  | Current L2 use | 5.35        | 1.77    | .510**  |
| LoE L2         | 0.311    | 0.168       | .314    | LoE L2         | 0.311       | 0.168   | .314    |
| <i>Step 3</i>  |          |             |         |                |             |         |         |
| Constant       | -10.00   | 33.10       |         | Constant       | -27.94      | 24.10   |         |
| Current L2 use | 7.14     | 2.74        | .882*   | Current L2 use | 7.46        | 1.98    | .712*** |
| LoE L2         | 0.300    | 0.170       | .302    | LoE L2         | 0.508       | 0.187   | .512*   |
| Current L1 use | 1.86     | 2.16        | .204    | AoO L2         | 0.407       | 0.204   | .440    |
|                | Model 3  |             |         | Model 4        |             |         |         |
|                | <i>b</i> | <i>SE b</i> | $\beta$ | <i>b</i>       | <i>SE b</i> | $\beta$ |         |
| <i>Step 1</i>  |          |             |         |                |             |         |         |
| Constant       | 23.24    | 8.12        |         | Constant       | 23.24       | 8.12    |         |
| Current L2 use | 7.27     | 1.51        | .694*** | Current L2 use | 7.27        | 1.51    | .694*** |
| <i>Step 2</i>  |          |             |         |                |             |         |         |
| Constant       | 17.59    | 8.33        |         | Constant       | 17.59       | 8.33    |         |
| Current L2 use | 5.35     | 1.77        | .510**  | Current L2 use | 5.35        | 1.77    | .510**  |
| LoE L2         | 9.311    | 0.168       | .314    | LoE L2         | 0.311       | 0.168   | .314    |
| <i>Step 3</i>  |          |             |         |                |             |         |         |
| Constant       | -23.72   | 12.38       |         | Constant       | 14.62       | 8.50    |         |
| Current L2 use | 3.14     | 1.51        | .300*   | Current L2 use | 6.06        | 1.82    | .578    |
| LoE L2         | 0.467    | 0.138       | .470**  | LoE L2         | 0.077       | 0.242   | .077    |
| SES            | 3.19     | 0.812       | .462**  | L2 schooling   | 0.575       | 0.433   | .273    |

Note: Model 1:  $R^2 = .481$  for Step 1:  $\Delta R^2 = .065$  for Step 2 ( $p = .076$ ):  $\Delta R^2 = .014$  for Step 3 ( $p = .398$ ). Model 2:  $R^2 = .481$  for Step 1:  $\Delta R^2 = .065$  for Step 2 ( $p = .076$ ):  $\Delta R^2 = .067$  for Step 3 ( $p = .058$ ). Model 3:  $R^2 = .481$  for Step 1:  $\Delta R^2 = .065$  for Step 2 ( $p = .076$ ):  $\Delta R^2 = .183$  for Step 3 ( $p = .001$ ). Model 4:  $R^2 = .481$  for Step 1:  $\Delta R^2 = .065$  for Step 2 ( $p = .076$ ):  $\Delta R^2 = .032$  for Step 3 ( $p = .196$ ). \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

correlations (current L1 use, AoO L2, SES, and length of L2 schooling) were entered into the third block of each of the hierarchical regression models (one per model).

The regression results for “identical repetition” and “target structure” measures of the German SRT are summarized in Tables 6 and 7, respectively. As can be seen

**Table 7.** Summary of hierarchical regression analysis for age and input factors predicting performance in German LITMUS-SRT scored by correct target structure

|                | Model 1  |             |         | Model 2        |             |         |         |
|----------------|----------|-------------|---------|----------------|-------------|---------|---------|
|                | <i>b</i> | <i>SE b</i> | $\beta$ | <i>b</i>       | <i>SE b</i> | $\beta$ |         |
| <i>Step 1</i>  |          |             |         |                |             |         |         |
| Constant       | 52.79    | 6.66        |         | Constant       | 52.79       | 6.66    |         |
| LoE L2         | 0.461    | 0.121       | .606*** | LoE L2         | 0.461       | 0.121   | .606*** |
| <i>Step 2</i>  |          |             |         |                |             |         |         |
| Constant       | 45.87    | 6.79        |         | Constant       | 45.87       | 6.79    |         |
| LoE L2         | 0.272    | 0.137       | .358    | LoE L2         | 0.272       | 0.137   | .358    |
| Current L2 use | 3.41     | 1.44        | .425*   | Current L2 use | 3.41        | 1.44    | .425*   |
| <i>Step 3</i>  |          |             |         |                |             |         |         |
| Constant       | 50.25    | 27.39       |         | Constant       | 19.26       | 20.46   |         |
| LoE L2         | 0.274    | 0.140       | .361    | LoE L2         | 0.387       | 0.158   | .509*   |
| Current L2 use | 3.12     | 2.27        | .389    | Current L2 use | 4.64        | 1.68    | .578*   |
| Current L1 use | -.296    | 1.79        | -.042   | AoO L2         | 0.238       | 0.173   | .336    |
|                | Model 3  |             |         | Model 4        |             |         |         |
|                | <i>b</i> | <i>SE b</i> | $\beta$ | <i>b</i>       | <i>SE b</i> | $\beta$ |         |
| <i>Step 1</i>  |          |             |         |                |             |         |         |
| Constant       | 52.79    | 6.66        |         | Constant       | 52.79       | 6.66    |         |
| LoE L2         | 0.461    | 0.121       | .606*** | LoE L2         | 0.461       | 0.121   | .606*** |
| <i>Step 2</i>  |          |             |         |                |             |         |         |
| Constant       | 45.87    | 6.79        |         | Constant       | 45.87       | 6.79    |         |
| LoE L2         | 0.272    | 0.137       | .358    | LoE L2         | 0.272       | 0.137   | .358    |
| Current L2 use | 3.41     | 1.44        | .425*   | Current L2 use | 3.41        | 1.44    | .425*   |
| <i>Step 3</i>  |          |             |         |                |             |         |         |
| Constant       | 10.80    | 9.80        |         | Constant       | 44.03       | 7.04    |         |
| LoE L2         | 0.404    | 0.110       | .532*** | LoE L2         | 0.127       | 0.200   | .167    |
| Current L2 use | 1.53     | 1.19        | .191    | Current L2 use | 3.85        | 1.51    | .480*   |
| SES            | 2.71     | .643        | .511*** | L2 schooling   | 0.357       | 0.358   | .220    |

Note: Model 1:  $R^2 = .367$  for Step 1:  $\Delta R^2 = .119$  for Step 2 ( $p = .027$ ):  $\Delta R^2 = .001$  for Step 3 ( $p = .870$ ). Model 2:  $R^2 = .367$  for Step 1:  $\Delta R^2 = .119$  for Step 2 ( $p = .027$ ):  $\Delta R^2 = .039$  for Step 3 ( $p = .182$ ). Model 3:  $R^2 = .367$  for Step 1:  $\Delta R^2 = .119$  for Step 2 ( $p = .027$ ):  $\Delta R^2 = .224$  for Step 3 ( $p < .001$ ). Model 4:  $R^2 = .367$  for Step 1:  $\Delta R^2 = .119$  for Step 2 ( $p = .027$ ):  $\Delta R^2 = .021$  for Step 3 ( $p = .330$ ). \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .



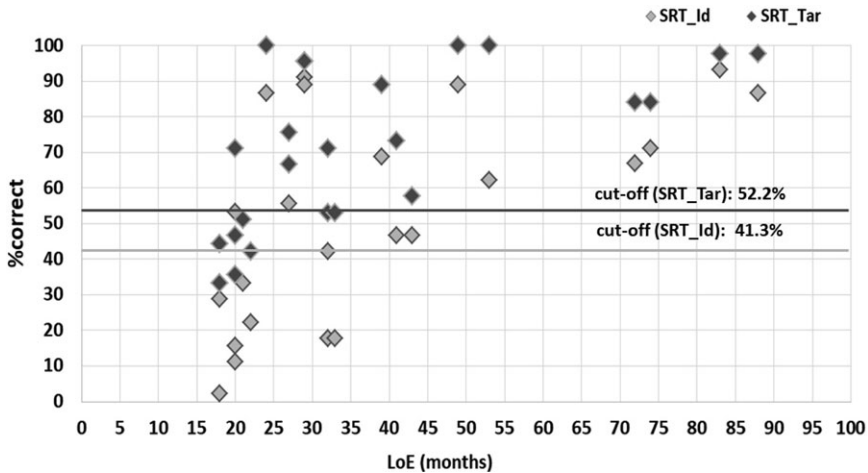


Figure 7. German LITMUS-SRT: percentage correct identical repetition and target structure versus length of exposure (LoE L2). Cutoff scores (Hamann & Abed Ibrahim, 2017). SRT Id, German LITMUS-SRT “identical repetition.” SRT Tar, German LITMUS-SRT “target structure.”

in the model summaries for “identical repetition” (Table 6), the major predictor for performance on “identical repetition” was current L2 use accounting for 48.1% of the variance in Step 1. The addition of LoE to L2 did not explain any additional variance in Step 2 of either model. Current L1 use, AoO, and L2 schooling did not explain any additional variance in the final step of Models 1, 2, and 4. In contrast, introducing SES in Step 3 explained an additional 18.3% of variance (see Table 6, Model 3). In the final step of regression Model 3, all predictors, including LoE to L2, were significant.

As for the measure “target structure,” the hierarchical regression analyses (Table 7) revealed that at Step 1, LoE to L2 accounted for 36.7% of variance. The addition of current L2 use to the model at Step 2 explained further 11.9% of the variance. Similar to “identical repetition,” adding AoO, current L1 use, or L2 schooling at Step 3 did not account for any additional variance on “target structure” in Models 1, 2, and 4, whereas including SES in the last step of Model 3 explained an additional 22.4% of the variance. In the final step of Model 3, only LoE to L2 and SES were significant predictors.

As the analyses above showed that LoE to the L2 was significantly related to the children’s performance on German LITMUS-SRT and in line with Research Question d, we wanted to explore whether less than 24 months of exposure would allow a fair assessment by the task. To that end, we plotted children’s performance on both SRT measures against LoE to the L2. A visual inspection of Figure 7 shows that the majority of refugee children with less than 24 months of exposure perform below the cutoff scores separating typically developing from language impaired children in Hamann and Abed Ibrahim (2017). Note that these cutoffs were determined for a population of younger bilinguals (5;6–9;0) including a subset of the present heritage sample.

*Linguistic and working memory variables.* Hierarchical regression modeling was conducted with performance scores in “identical repetition” and “target structure” as the dependent variables and L2 morphosyntax as measured by TROG-D, expressive and receptive vocabulary, as well as FDS as independent variables given their strong to moderate correlations with the two dependent measures (see Table 2). To avoid multicollinearity between L2 expressive and receptive vocabulary scores, both factors were run in separate models (Dormann *et al.*, 2013) entering either receptive or expressive vocabulary at Stage 2, and FDS at Stage 3.

As can be seen in the model summaries in Table 8, L2 morphosyntax emerged as the factor explaining most of the variance (36.9%) in “identical repetition.” Expressive vocabulary accounted for an additional 26.9% and FDS for further 18.6%. In the second model, L2 morphosyntax accounted for 36.9% of the variance, an additional 29.7% of the variance was explained by receptive vocabulary with FDS adding 10.6% more in the final step.

For “target structure,” L2 morphosyntax explained 35.5% of the variance, with expressive vocabulary and FDS weighing in to account for an additional 22.5% and 15.9%, in Steps 2 and 3, respectively (Table 9, Model 1). Importantly, adding FDS in Step 3 did not significantly contribute toward explaining variance when receptive vocabulary is entered at Stage 2, where L2 morphosyntax explained 35.5% of the variance and receptive vocabulary an additional 30.2%. (Table 9, Model 2).

In order to determine whether a certain developmental level is necessary for performance above the cutoff for impairment, especially for the score “target structure,”

**Table 8.** Summary of hierarchical regression analysis for working memory and linguistic variables predicting performance in German LITMUS-SRT scored by identical repetition

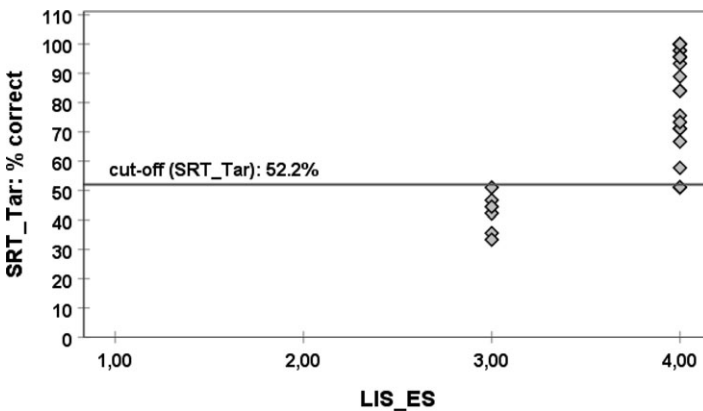
|               | Model 1  |             |         | Model 2  |             |         |        |
|---------------|----------|-------------|---------|----------|-------------|---------|--------|
|               | <i>b</i> | <i>SE b</i> | $\beta$ | <i>b</i> | <i>SE b</i> | $\beta$ |        |
| <i>Step 1</i> |          |             |         |          |             |         |        |
| Constant      | 40.20    | 9.01        |         | Constant | 40.20       | 9.01    |        |
| TROG-D        | 0.576    | 0.188       | .607**  | TROG-D   | 0.576       | 0.188   | .607** |
| <i>Step 2</i> |          |             |         |          |             |         |        |
| Constant      | 24.23    | 8.51        |         | Constant | -26.63      | 19.54   |        |
| TROG-D        | 0.325    | 0.165       | .343*   | TROG-D   | 0.208       | 0.174   | .219*  |
| WWT LexP      | 0.975    | 0.292       | .582**  | WWT LexR | 1.023       | 0.280   | .669** |
| <i>Step 3</i> |          |             |         |          |             |         |        |
| Constant      | -18.54   | 12.68       |         | Constant | -41.00      | 17.66   |        |
| TROG-D        | 0.221    | 0.122       | .233*   | TROG-D   | 0.205       | 0.149   | .216*  |
| WWT LexP      | 0.717    | 0.212       | .428**  | WWT LexR | 0.628       | 0.286   | .410*  |
| FDS           | 9.85     | 2.55        | .487**  | FDS      | 8.41        | 3.31    | .417*  |

Note: Model 1:  $R^2 = .369$  for Step 1:  $\Delta R^2 = .269$  for Step 2 ( $p = .004$ ):  $\Delta R^2 = .186$  for Step 3 ( $p = .002$ ). Model 2:  $R^2 = .369$  for Step 1:  $\Delta R^2 = .297$  for Step 2 ( $p = .002$ ):  $\Delta R^2 = .106$  for Step 3 ( $p = .023$ ). \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

**Table 9.** Summary of hierarchical regression analysis for working memory and linguistic variables predicting performance in German LITMUS-SRT scored by correct target structure

|               | Model 1  |             |         | Model 2  |             |         |        |
|---------------|----------|-------------|---------|----------|-------------|---------|--------|
|               | <i>b</i> | <i>SE b</i> | $\beta$ | <i>b</i> | <i>SE b</i> | $\beta$ |        |
| <i>Step 1</i> |          |             |         |          |             |         |        |
| Constant      | 60.91    | 6.83        |         | Constant | 60.91       | 6.83    |        |
| TROG-D        | 0.424    | 0.143       | .596**  | TROG-D   | 0.424       | 0.143   | .596** |
| <i>Step 2</i> |          |             |         |          |             |         |        |
| Constant      | 49.95    | 6.88        |         | Constant | 10.32       | 14.85   |        |
| TROG-D        | 0.252    | 0.134       | .354*   | TROG-D   | 0.146       | 0.132   | .205*  |
| WWT LexP      | 0.670    | 0.236       | .533*   | WWT LexR | 0.774       | 0.213   | .674** |
| <i>Step 3</i> |          |             |         |          |             |         |        |
| Constant      | 20.25    | 11.58       |         | Constant | 1.38        | 14.4    |        |
| TROG-D        | 0.180    | 0.112       | .253*   | TROG-D   | 0.144       | 0.121   | .202*  |
| WWT LexP      | 0.490    | 0.202       | .390*   | WWT LexR | 0.528       | 0.233   | .460*  |
| FDS           | 6.83     | 2.33        | .451*   | FDS      | 5.23        | 2.70    | .345   |

Note: Model 1:  $R^2 = .355$  for Step 1;  $\Delta R^2 = .225$  for Step 2 ( $p = .012$ );  $\Delta R^2 = .159$  for Step 3 ( $p = .011$ ). Model 2:  $R^2 = .355$  for Step 1;  $\Delta R^2 = .302$  for Step 2 ( $p = .002$ );  $\Delta R^2 = .073$  for Step 3 ( $p = .073$ ). \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .



**Figure 8.** Performance in German LITMUS-SRT scored by correct target structure plotted against syntactic developmental level (LiSe DaZ). Cutoff scores (Hamann & Abed Ibrahim, 2017). LIS ES, LiSe DaZ Entwicklungsstufe (ES), that is, developmental level. SRT Tar, German LITMUS-SRT “target structure.”

performance was plotted against the developmental level the children attained in the LiSe-DaZ subscale assessing sentence complexity in elicited production. Visual inspection of the scatterplot in Figure 8 indicates that only children who reached developmental Level 4, that is, produced embedded clauses in this subtask, were capable of performing above the cutoffs for impairment established for younger bilinguals in Hamann and Abed Ibrahim (2017).

## Discussion

The point of departure for our study was the lack of assessment tools for refugee children who are late child L2 learners, arriving in Germany at school age and receiving language support only for 18 to 24 months. We therefore investigated Arabic L1 and German L2 LITMUS repetition tasks with a view to their use as language assessments for refugee and heritage children, Research Question a.

We therefore asked first how the groups compared on an Arabic LITMUS-SRT and on a standardized Arabic test, Question b. Turning to L2 assessments, we then asked how our two groups performed on the German LITMUS NWRT and SRT, Question c. Finally, we wanted to know which factors influence performance on the repetition tasks, with special attention to language measures and exposure to L2, Question d. The 12 heritage children were born in Germany or came at a very early age. The refugee group,  $N = 15$ , had diverse transit itineraries to Germany, housing conditions in Germany, and enrollment in a variety of German schooling programs. This heterogeneity makes the group typical for the German situation. Our findings showed that the Arabic SRT and the German NWRT rendered comparable results for both groups, whereas the German SRT (as well as the standardized L2 measures we employed) was more difficult for the refugees than for the heritage children, even after 24 months of exposure to German.

### ***Standardized L1 assessments may be difficult for heritage children; L2 assessments are difficult for refugee children***

For these comparisons it was important to ensure beforehand that all children had typical development, so that low performance on a task could not be attributed to language impairment. Nevertheless, the performance of the refugee group on standardized L2 tests for vocabulary (WWT) and morphosyntax (LiSe-DaZ & TROG-D) was low and in the DLD range (see the Standardized L2-Tests section). Therefore, L1 assessment is crucial to achieving unbiased assessment for this group. The members of the heritage group all had previously been classified as typically developing bilinguals, but findings on L1 maintenance/attrition (Lein, Rothweiler, & Hamman, 2017; Montrul, 2008) led us to expect that they would show lower scores on L1 tests compared to the refugee children, who have used Arabic exclusively up to their fleeing of the home country. This prediction is born out for ELO-L, the standardized L1 test used here. Figures 2 and 3 show that heritage children scored lower than the refugee children in all subtasks, except for comprehension of morphosyntax and phonology where both groups performed alike. Vocabulary was particularly affected in the heritage group, which confirms previous results on vocabulary size in bilinguals (Bialystok, Craik, Greeb, & Gollan, 2010; Chondrogianni & Marinis, 2011; Goldberg, Paradis, & Crago, 2008; Oller, Pearson, & Cobo-Lewis, 2007).

### ***Equal performance of both groups on Arabic LITMUS-SRT and predictive factors***

We next asked how the two groups would perform on the Arabic LITMUS-SRT, a test that can be applied and scored in about ten minutes. Both groups were not

significantly different in performance. The only linguistic measure predicting performance on both SRT scores, identical repetition and target structure, was the score for comprehension of morphosyntax on the ELO-L. The fact that performance was on par for this measure in both groups implies that the Arabic SRT could constitute a fair assessment tool, not only for refugee children, but also for heritage children at school age (see also results on DLD in French–Arabic bilingual children; Abboud, Tuller, Henry & Saad, 2013). Our analyses also showed that neither working memory, SES and L1 schooling, nor age and input variables related to bilingualism (AoO, LoE, input, and use) predicted performance on the Arabic SRT. The latter results indicate that heritage children would not be disadvantaged and that the Arabic LITMUS-SRT could be administered with both groups. However, it might not be challenging enough for older children, as ceiling performance was observed. The Arabic SRT, even considering the issue with older children, could thus be used to provide evidence that heritage and refugee children have typical L1 development, which, in turn, would imply that any difficulties in the L2 cannot be due to language impairment.

### ***Performance on the German LITMUS repetition tasks***

For Question c, we asked how the refugee children compared to the heritage group in their performance on the German LITMUS tasks, NWRT, and SRT. Previous research has shown that both tasks have good accuracy for identifying typical development in heritage children (see Abed Ibrahim & Fekete, 2019). Comparing our heritage group to the refugee children should, therefore, give indications of whether these tasks would be fair to the latter or whether they would disadvantage them.

#### *Good performance in German QU-NWRT by both refugee and heritage children*

Surprisingly, the refugee group, including children with the least L2 exposure, performed equal to the heritage group on the NWRT. First explanations might be that the scoring method allows disregarding voicing and vowel errors, which were abundant in the refugee group. The result is in line with results from dos Santos and Ferré (2018) on a similar NWRT, which show that it can be reliably applied after very short exposure to an L2. As a tentative explanation, we suggest that language development is crucially influenced by the development of the phonological loop in the first year of life (see Pierce, Genesee, Decenserie, & Morgan, 2017, for an overview). If this is true, then L1 phonological development may be just as important for performance on quasi-universal NWRTs as exposure to L2 or current L2 use. Studies with more children and language combinations and careful factor analysis will be necessary to decide these issues, however.

#### *Heritage children perform better on the German LITMUS-SRT than refugee children: predictors*

In contrast to the NWRT, results of the German LITMUS-SRT showed that it was very difficult for the refugee children; the heritage children generally performed significantly better, especially when scored by “target structure.” Moreover, the results

in Figure 7 also indicate that only after 24–27 months of exposure can the SRT be administered as an L2 test for refugees.

Factors influencing performance (Tables 6 and 7) were current L2 use (48.1%) and SES (18.3%) for the measure “identical repetition,” and a combination of LoE (36.7%) current L2 use (11.9%) and SES (22.4%) for “target structure.” This corresponds to results from other studies supporting the influence of current L2 use (Meir, 2018; Tuller *et al.*, 2018; Unsworth, 2016). Contrary to expectation, L2 schooling did not significantly contribute to explaining variance. The influence of SES, in line with findings from Cobo-Lewis *et al.* (2002), can best be explained when linguistic variables are considered in conjunction with it (see below).

When linguistic and working memory measures were combined (Tables 8 and 9), morphosyntactic competence as measured by the TROG-D explained most of the variance in “identical repetition,” followed by receptive vocabulary and FDS. FDS explained more of the variance if it was TROG-D and expressive vocabulary that were considered. This is not surprising as good repetition might depend either on good vocabulary alleviating memory load or on good VSTM memory so that difficulties with expressive vocabulary may be compensated. For “target structure” (Table 9), L2 morphosyntax explained 35.5% of the variance, with expressive vocabulary and FDS accounting for an additional 22.5% and 15.9%, respectively. However, FDS did not significantly contribute toward explaining variance when receptive vocabulary was entered at Stage 2, where L2 morphosyntax explained 35.5% of the variance and receptive vocabulary an additional 30.2%. This confirms that the German LITMUS-SRT, when scored with the measure “target structure,” is a measure of linguistic competence, measuring mainly syntactic competence but also receptive vocabulary (Hamann & Abed Ibrahim, 2017; Tuller *et al.*, 2018).

When other morphosyntactic measures were investigated, it turned out that a good predictor was “developmental level” from the LiSe-DaZ. Only children able to produce subordinate clauses were capable of performing reasonably well on the German SRT (see Figure 8). This aligns with previous studies indicating that early L2 learners acquire the German sentence structure in main clauses fast, but take longer to acquire verb placement in subordinate clauses (object–verb), which is particularly relevant for late learners with an L1 with verb–object structures (Haberzettl, 2005). These findings demonstrate that refugee children, during the first 24 months of exposure, will have difficulties with tests that include such L1 sensitive milestones of L2 acquisition. This is unfortunate as assessing such structures allows individual and detailed diagnosis and support, and the problem concerns not only this SRT but also the LiSe-DaZ and TROG-D: the participants performing in the DLD range on TROG-D also performed poorly on SRT. This, again, is no surprise, since the TROG-D, like the SRT, requires comprehension of *Wh*-, subordination, relative clauses, and passives.

These results on the influence of measures of vocabulary and complex syntax can be tied in with the influence of SES on performance on the German SRT. Immigrant families with higher education usually have a positive attitude toward academic achievement and language learning of their children, use more differentiated vocabulary and more complex structures in everyday conversations (Scheele, Leseman, & Mayo, 2010) and in our sample they have reasonable L2 skills allowing interaction with majority speakers, for example, educators. It is thus not surprising that SES

influences performance in the SRT, given that SES strongly influences vocabulary knowledge and complex morphosyntax (Chondrogianni & Marinis, 2011; Roy, Chiat, & Dodd, 2014).

It thus emerges from our analyses (Figures 7 and 8 and Tables 6–9) that good performance on the SRT depends on current L2 use, LoE and SES on the one hand, and on vocabulary knowledge as well as morphosyntactic knowledge on the other hand, and that at least 24 months of exposure are needed to acquire these language skills, probably even more. This might appear as a disappointing result on the SRT. At the same time, however, it confirms that performance on this SRT depends on syntactic complexity and vocabulary and, therefore, is a measure of language competence and development.

### **Conclusion**

Because language assessment tools for newcomers and refugee children are urgently needed for decisions on language support and schooling programs, we investigated whether certain L1 and L2 LITMUS repetition tasks can be used as assessments for typical development in bilingual populations including refugee children. The study confirms previous research showing that the German LITMUS-QU-NWRT is a reliable assessment tool for all groups of bilingual children (Chilla, Hamman, et al., in press; Grimm & Hübner, in press; Hamann & Abed Ibrahim, 2017). In contrast, the German LITMUS-SRT assesses heritage children (see also Abed Ibrahim & Fekete, 2019) very reliably but can, at best, be used to measure language abilities of refugee children only after 24 months of L2 exposure. Standardized L2 assessment tools measuring performance in German morphosyntax (reception or comprehension) confirmed previous findings on L2 children, also on spontaneous speech data, in that refugee children performed in the range of children with DLD (Chilla, 2008; Håkansson & Nettelbladt, 1993; Schöler, Fromm, & Kany, 1998); thus, the SRT, as a morphosyntactic measure, is no exception here.

These findings are complemented by the results on L1 assessments. Here, the refugee children had an advantage over the heritage children in most subtests of the standardized test. They performed on par with the heritage children on the Arabic LITMUS-SRT, however. This is encouraging as to the language competence not only of the refugee but also of the heritage children, even though it seems to demonstrate that reliable language assessment of bilinguals in the domain of morphosyntax would have to include L1 measures, particularly in refugee populations. Applying L2 and L1 measures for bilinguals has long been the best practice in speech–language pathology but is usually not feasible in schools. Using the German LITMUS-NWRT for a first assessment may thus provide an easy and welcome alternative.

Combinations of the tests will therefore allow a fast estimate of typical development. The German LITMUS-SRT demonstrably measures morphosyntactic competence, which, unfortunately, does not develop as fast as many schooling and integration models presuppose. This, in particular, is a result that needs to be discussed widely and that should be used to correct the often too optimistic expectations of teachers and educators.

**Acknowledgments.** The authors are grateful to the children, parents and teachers who participated in this study. We also thank Angela Grimm and Rasha Zebib for the permission to use the LITMUS-NWRT and the Lebanese version of the Arabic LITMUS-SRT and Laurice Tuller and the BiLaD-Team from Tour for the fruitful collaboration we had. This research was carried out within the BiLaD- and BilisAT- projects, the latter financed by DFG Grants to CH 1112/4-1S (to Chilla) and HA 2335/7-1C (to Hamann), the former financed by a bilateral grant French ANR-12-FRAL-0014-01 and German DFG: HA 2335/6-1, CH 1112/2-1, and RO 923/3-1.

## References

- Aboud, L., Tuller, L., Henry, G., & Saad, S.** (2013). *Sentence repetition as a feasible assessment tool for identifying children with SLI in the Lebanese and French bilingual contexts*. Poster presented at the COST meeting in Krakow, May 2013.
- Abed Ibrahim, L., & Fekete, I.** (2019). What machine learning can tell us about the role of language dominance in the diagnostic accuracy of German LITMUS non-word and sentence repetition tasks. *Frontiers in Psychology*, *9*, 27–57.
- Abed Ibrahim, L., & Hamann, C.** (2017). Bilingual Arabic–German & Turkish–German children with and without specific language impairment: Comparing performance in sentence and nonword repetition tasks. In M. LaMendola & J. Scott (Eds.), *Proceedings of BUCLD 41* (pp. 1–17). Somerville, MA: Cascadilla Press.
- Ahrenholz, B., Fuchs, I., & Birnbaum, T.** (2016). Dann haben wir natürlich gemerkt der Übergang ist der knackpunkt. *Modelle der Beschulung von Seiteneinsteigern in der Praxis. BiSS Journal*, *5*.
- Albirini, A.** (2018). The role of age of exposure to English in the development of Arabic as a heritage language in the United States. *Language Acquisition*, *25*, 178–196.
- Archibald, L. M. D., & Gathercole, S. E.** (2007). Short-term and working memory in specific language impairment. *International Journal of Language & Communication Disorders*, *41*, 675–693.
- Armon-Lotem, S., de Jong, J., & Meir, N.** (Eds.) (2015). *Assessing multilingual children: Disentangling bilingualism from language impairment*. Bristol: Multilingual Matters.
- Armon-Lotem, S., & Meir, N.** (2016). Diagnostic accuracy of repetition tasks for the identification of specific language impairment (SLI) in bilingual children: Evidence from Russian and Hebrew. *International Journal of Language & Communication Disorders*, *51*, 715–731.
- Austin, P. C. & Steyerberg, E. W.** (2015). The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, *68*, 627–636. doi: [10.1016/j.jclinepi.2014.12.014](https://doi.org/10.1016/j.jclinepi.2014.12.014)
- Becker-Mrotzek, M., Hentschel, B., Hippmann, K., & Linnemann, M.** (2012). *Mercator Institut für Sprachförderung und Deutsch als Zweitsprache. Experten*. Retrieved from <https://www.mercator-institut-sprachfoerderung.de/de/publikationen/studien-des-mercator-instituts/>
- Bedore, L. M., Peña, E. D., Griffin, Z. M., & Hixon, J. G.** (2016). Effects of age of English exposure, current input/output, and grade on bilingual language performance. *Journal of Child Language*, *43*, 687–706.
- Bialystok, E.** (1997). The structure of age: In search of barriers to second language acquisition. *Second Language Research*, *13*, 116–137.
- Bialystok, E., Craik, F. I. M., Green, D. W., & Gollan, T. H.** (2010). Bilingual minds. *Psychological Science in the Public Interest*, *10*, 89–129.
- Birdsong, D.** (2018). Plasticity, variability and age in second language acquisition and bilingualism. *Frontiers in Psychology*, *9*. doi: [10.3389/fpsyg.2018.00081](https://doi.org/10.3389/fpsyg.2018.00081)
- Bishop, D. V. M.** (1989). *Test for reception of grammar (TROG)*. London: Medical Research Council.
- Bishop, D. V. M., Adams, C. V., & Norbury, C. F.** (2006). Distinct genetic influences on grammar and phonological short-term memory deficits: Evidence from 6-year-old twins. *Genes, Brain, and Behavior*, *5*, 158–169.
- Bulheller, S., & Häcker, H.** (2002). *Coloured progressive matrices—Deutsche Bearbeitung und Normierung* [German adaptation and norms]. Frankfurt: Swets & Zeitlinger.
- Chiat, S.** (2015). Nonword repetition. In S. Armon-Lotem, J. De Jong, & N. Meir (Eds.), *Methods for assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 125–151). Bristol: Multilingual Matters.



- Chiat, S., & Polišíenská, K.** (2016). A framework for crosslinguistic nonword repetition tests: effects of bilingualism and socioeconomic status on children's performance. *Journal of Speech, Language Hearing Research*, *59*, 1179–1189.
- Chilla, S.** (2008). *Erstsprache, Zweitsprache, Spezifische Sprachentwicklungsstörung? Eine Untersuchung der deutschen Hauptsatzstruktur durch sukzessiv-bilinguale Kinder mit türkischer Erstsprache*. Hamburg: Dr. Kovač.
- Chilla, S.** (in press). Assessment of developmental language disorders in bilinguals: Immigrant Turkish as a bilectal challenge in Germany. In E. Saiegh-Haddad, L. Laks, & C. McBride (Eds.), *Handbook of literacy in diglossia and dialectal contexts—Psycholinguistic and Educational Perspectives*. New York: Springer.
- Chilla, S., Hamann, C., Prévost, P., Abed Ibrahim, L., Ferré, S., dos Santos, C., . . . Tuller, L.** (in press). The influence of different first languages on L2 LITMUS-NWR and L2 LITMUS-SRT in French and German: A crosslinguistic approach. In K. Grohmann & S. Armon-Lotem (Eds.), *LITMUS in action: Comparative studies across Europe, TILAR*. Amsterdam: Benjamins.
- Chilla, S., Krupp, S., & Wulff, N.** (2019). Konzepte für den Deutscherwerb neu zugewanderter Jugendlicher im deutschen Bildungssystem: Lehrwerke oder adaptive Baukästen—Welchen Anforderungen müssen Lehr- und Lernmaterialien genügen? In K. Kovács, & A. F. Dombi (Eds.), *National and international tendencies of education and training* (pp. 325–344). Szeged: Szegedi Egyetemi Kiadó Juhász Gyula Felsőoktatási Kiadó.
- Chilla, S., & Şan, H.** (2017). Möglichkeiten und Grenzen der Diagnostik erstsprachlicher Fähigkeiten: Türkisch-deutsche und türkisch-französische Kinder im Vergleich. *Sprachen 2016. Russisch und Türkisch im Fokus* (pp. 175–205). Berlin: Berlin Interdisciplinary Association for Multilingualism.
- Chondrogianni, V., & Marinis, T.** (2011). Differential effects of internal and external factors on the development of vocabulary, tense morphology and morphosyntax in successive bilingual children. *Linguistic Approaches to Bilingualism*, *1*, 318–342.
- Clahsen, H.** (1986). *Die Profilanalyse. Ein linguistisches Verfahren für die Sprachdiagnose im Vorschulalter*. Berlin: Marhold.
- Cobo-Lewis, A., Pearson, B., Eilers, R., & Umbel, V.** (2002). Effects of bilingualism and bilingual education on oral and written English skills: A multifactor study of standardized test outcomes. In D. K. Oller & R. Eilers (Eds.), *Language and literacy in bilingual children* (pp. 64–97). Clevedon: Multilingual Matters.
- Conti-Ramsden, G., Botting, N., & Faragher, B.** (2001). Psycholinguistic markers for specific language impairment. *Journal of Child Psychology and Psychiatry*, *42*, 741–748.
- DeCapua, A., & Marshall, H. W.** (2011). Reaching ELLs at risk: Instruction for students with limited or interrupted formal education. *Preventing School Failure: Alternative Education for Children and Youth*, *55*, 35–41.
- Deml, I., Binder, K., Schulte, M., Merkert, A., Bien-Miller, L., Wildemann, A., & Schilcher.** (2018). Eva-Prim. Evaluation im Primarbereich: Sprachförderung in alltäglichen und fachlichen Kontexten im Rahmen der Bund-Länder-Initiative BiSS. In S. Henschel, S. Gentrup, L. Beck, & P. Stanat (Eds.), *Projektatlas evaluation: Erste ergebnisse aus den BiSS-Evaluationsprojekten* (pp. 32–34). Berlin: BiSS-Trägerkonsortium.
- Diehm, I., & Radtke, F.-O.** (1999). *Erziehung & migration: Eine einföhrung*. Stuttgart: Kohlhammer.
- Dormann, C., Elith, J., Bacher, S., Bucmann, C., Carl, G., Carré, G., . . . S. Lautenbach.** (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*, 27–46. doi: [10.1111/j.1600-0587.2012.07348.x](https://doi.org/10.1111/j.1600-0587.2012.07348.x)
- dos Santos, C., & Ferré, S.** (2018). A nonword repetition task to assess bilingual children's phonology. *Language Acquisition*, *25*, 58–71.
- Farnia, F., & Geva, E.** (2011). Cognitive correlates of vocabulary growth in English language learners. *Applied Psycholinguistics*, *32*, 711–738.
- Fleckstein, A., Prévost, P., Tuller, L., Sizaret, E., & Zebib, R.** (2018). How to identify SLI in bilingual children: A study on sentence repetition in French. *Language Acquisition*, *25*, 1–17. doi: [10.1080/10489223.2016.1192635](https://doi.org/10.1080/10489223.2016.1192635)
- Fox, A.** (2009). *TROG-D. Test zur Überprüfung des Grammatikverständnisses*. Idstein: Schulz-Kirchner.
- Fredman, M.** (2006). Recommendations for working with bilingual children—Prepared by the multilingual affairs committee of IALP. *Folia Phoniater. Logop*, *58*, 458–464.
- Friedmann, N., & Novogrodsky, R.** (2011). Which questions are most difficult to understand? The comprehension of Wh questions in three subtypes of SLI. *Lingua*, *121*, 367–382.

- Gagarina, N.** (2017). Monolingualer und bilingualer Erstspracherwerb des Russischen: Ein Überblick. In N. Wulff, N., & K. Witzlack-Makarevich (Eds.), *Handbuch des Russischen in Deutschland: Migration—Mehrsprachigkeit—Spracherwerb* (pp. 393–410). Berlin: Frank & Timme.
- Gagarina, N., Klässert, A., & Topaj, N.** (2010). Sprachstandstest Russisch für mehrsprachige Kinder. *ZAS Papers in Linguistics*, *54*.
- Gallon, N., Harris, J., & van der Lely, H.** (2007). Non-word repetition: An investigation of phonological complexity in children with Grammatical SLI. *Clinical Linguistics & Phonetics*, *21*, 445–455.
- Gantefort, C., & Roth, H.-J.** (2008). Ein Sturz und seine Folgen. Zur Evaluation von Textkompetenz im narrativen Schreiben mit dem FÖRMIG-Instrument “Tulpenbeet.” In T. Klinger, K. Schwippert, & B. Leiblein (Eds.), *Evaluation im modellprogramm FÖRMIG. Planung und Realisierung eines Evaluationskonzepts* (pp. 29–50). Münster: Waxmann.
- Garcia, O.** (2009). Education, multilingualism and translanguaging in the 21st century. In A. Mohanty, M. Panda, R. Phillipson, & T. Skutnabb-Kangas (Eds.), *Multilingual education for social justice: Globalising the local* (pp. 128–145). New Delhi: Orient Blackswan.
- Geva, E., & Wiener, J.** (2015). *Psychological assessment of culturally and linguistically diverse children and adolescents*. New York: Springer.
- Glück, C.** (2011). *Wortschatz- und Wortfindungstest für 6- bis 10-Jährige: WWT 6–10* (2nd ed.) Munich: Elsevier, Urban & Fischer.
- Goldberg, H., Paradis, J., & Crago, M.** (2008). Lexical acquisition over time in minority first language children learning English as a second language. *Applied Psycholinguistics*, *29*, 1–25.
- Gogolin, I.** (1994). *Der monolinguale Habitus der multilingualen Schule*. Münster: Waxmann.
- Gogolin, I., Lamge, I., Lengyel, D., & Schwippert, K.** (2011). *FörMig. Kompetenzzentrum Förderung von Kindern und Jugendlichen mit Migrationshintergrund*. Hamburg: Universität Hamburg. Retrieved from <https://www.foermig.uni-hamburg.de/>
- Golberg, H., Paradis, J., & Crago, M.** (2008). Lexical acquisition over time in minority first language children learning English as a second language. *Applied Psycholinguistics*, *29*, 41–65.
- Griefhaber, W.** (2013). Die Profilanalyse für Deutsch als Diagnoseinstrument zur Sprachförderung. *Stiftung Mercator: proDaZ*. Universität Essen.
- Grimm, A., Ferré, S., dos Santos, C., & Chiat, S.** (2014). Can nonwords be language-independent? Cross-linguistic evidence from monolingual and bilingual acquisition of French, German, and Lebanese. In *Symposium on Language Impairment Testing in Multilingual Settings (LITMUS): Disentangling bilingualism and SLI*, Amsterdam, July 13–19, 2014.
- Grimm, A., & Hübner, J.** (in press). Nonword repetition by bilingual learners of German: The role of language-specific complexity. In C. dos Santos & L. de Almeida, *Bilingualism and Specific Language Impairment, Bi-SLI 2015*. Amsterdam: Benjamins.
- Grimm, A., & Schulz, P.** (2014). Specific language impairment and early second language acquisition: The risk of over- and underdiagnosis. *Child Indicators Research*, *7*, 821–841. doi: [10.1007/s12187-013-9230-6](https://doi.org/10.1007/s12187-013-9230-6)
- Gutierrez-Clellen, V., & Kreiter, J.** (2003). Understanding child bilingual acquisition using parent and teacher reports. *Applied Psycholinguistics*, *24*, 267–288.
- Haberzettl, S.** (2005). *Der Erwerb der Verstellungsregeln in der Zweitsprache Deutsch durch Kinder mit russischer und türkischer Muttersprache*. Tübingen: Max Niemeyer.
- Hahn-Hobek, N.** (2017). Unaccompanied minors in the German context. In A. Korntheuer, P. Pritchard, & D. Maehler (Eds.), *Structural context of refugee integration in Canada and Germany* (pp. 133–141). GESIS, Volume 15. Köln: Leibniz Institute for Social Sciences.
- Hair Jr., J. F., Anderson, R. E., Tatham, R. L., & Black, W. C.** (1995). *Multivariate data analysis* (3rd ed.). New York: Macmillan.
- Håkansson, G., & Nettelbladt, U.** (1993). Developmental sequences in L1 (normal and impaired) and L2 acquisition of Swedish. *International Journal of Applied Linguistics*, *3*, 131–157.
- Hamann, C., & Abed Ibrahim, L.** (2017). Methods for identifying specific language impairment in bilingual populations in Germany. *Frontiers in Communication*, *2*, 1–19. doi: [10.3389/fcomm.2017.00016](https://doi.org/10.3389/fcomm.2017.00016)
- Hamann, C., Penner, Z., & Linder, K.** (1998). German impaired grammar: The clause structure revisited. *Language Acquisition*, *7*, 193–245.
- Hamann, C., Chilla, S., Ruigendijk, E., & Abed Ibrahim, L.** (2013). *A German sentence repetition task: Testing bilingual Russian/German children*. Poster presented at the COST meeting in Krakow, May 2013.

- Henry, G., Tuller, L., Prévost, P., & Zebib, R. (in press). Les outils LITMUS-libanais: Synthèse et regard croisé. In R. Zebib (Ed.), *Plurilinguisme et Troubles Spécifiques du Langage au Liban*. Beirut: Presses universitaires de l'Université Saint Joseph.
- Hertel, I. (2019). *Refugee children in Germany: Introducing a new population to education system and linguistic research*. Master thesis, University of Oldenburg.
- IBM Corp. Released (2016). *IBM SPSS Statistics for Windows, Version 24.0*. Armonk, NY: Author.
- International Association of Logopedics and Phoniatrics. (2011). *Recommendations for working with bilingual children*. Available at: [http://www.specchioriflesso.net/media/162083/linee\\_guida\\_bilingui\\_ialp\\_may\\_2011.pdf](http://www.specchioriflesso.net/media/162083/linee_guida_bilingui_ialp_may_2011.pdf)
- Karayayla, T., & Schmid, M. (2018). First language attrition as a function of age of onset of bilingualism: First language attainment of Turkish-English bilinguals in the United Kingdom. *Language Learning*, *69*, 106–142.
- Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. A., Gustafsson, J. E., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science*, *18*, 146–154.
- Köpke, B. (2007). Language attrition at the cross-roads of brain, mind, society. In B. Köpke, M. S. Schmid, M. Keijzer, & S. Dostert (Eds.), *Language and attrition. Theoretical perspectives* (pp. 9–37). Amsterdam: Benjamins.
- Lein, T., Rothweiler, M., & Hamann, C. (2017). Factors affecting the performance in child heritage Portuguese in Germany. In J. Choi, H. Demirdache, O. Lungu, & L. Voeltzel (Eds.), *Proceedings of GALA 2015* (pp. 146–175). Newcastle: Cambridge Scholars Publishing.
- List, G., & List, G. (2004). Sprachliche Heterogenität, Quersprachigkeit und sprachliches Lernen. In J. Quetz, & G. Solmecke (Eds.), *Brücken schlagen: Fächer – Sprachen – Institutionen. Dokumentation zum 20. Kongress für Fremdsprachendidaktik* (pp. 89–104). Berlin: Pädagogischer Zeitschriftenverlag.
- Marinis, T., & Armon-Lotem, S. (2015). Sentence repetition. In S. Armon-Lotem, J. de Jong, & M. Neir (Eds.), *Methods for assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 95–125). Clevedon: Multilingual Matters.
- Massumi, M., & von Dewitz, N. (2015). *Neu zugewanderte Kinder und Jugendliche im deutschen Schulsystem*. Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache und vom Zentrum für LehrerInnenbildung der Universität zu Köln. Köln. Retrieved from [https://www.mercator-institut-sprachfoerderung.de/fileadmin/Redaktion/PDF/Publicationen/MI\\_ZfL\\_Studie\\_Zugewanderte\\_im\\_deutschen\\_Schulsystem\\_final\\_screen.pdf](https://www.mercator-institut-sprachfoerderung.de/fileadmin/Redaktion/PDF/Publicationen/MI_ZfL_Studie_Zugewanderte_im_deutschen_Schulsystem_final_screen.pdf)
- Meir, N. (2017). Effects of specific language impairment (SLI) and bilingualism on verbal short-term memory. *Linguistic Approaches to Bilingualism*, *7*, 301–330. doi.org/10.1075/lab.15033.mei
- Meir, N. (2018). Morpho-syntactic abilities of unbalanced bilingual children: A closer look at the weaker language. *Frontiers in Psychology*, *9*, 1318. doi: 10.3389/fpsyg.2018.01318
- Meisel, J. (2009). Second language acquisition in early childhood. *Zeitschrift für Sprachwissenschaft*, *28*, 5–34.
- Ministerium für Bildung, Wissenschaft und Kultur des Landes Schleswig-Holstein. (2018). Erlass über die Aufgaben der Ansprechpersonen für DaZ in den allgemein bildenden Schulen. Erlass des Ministeriums für Bildung, Wissenschaft und Kultur vom 15. November 2018 – III 21. *NBl.MBWK.Schl.-H.*, 523.
- Montanari, E. (2017). *Beschulung von neu in das niedersächsische Bildungssystem zugewanderten Schülerinnen und Schülern der Sekundarstufe 1* (pp. 1–40). Stiftung Universität Hildesheim, Zentrum für Bildungsintegration.
- Montanari, E., & Akinci, M., & Abel, R. (2019). Balance and dominance in the vocabulary of German-Turkish primary schoolchildren. *European Journal of Applied Linguistics*, *7*, 113–144. doi: 10.1515/eujal-2018-0003
- Montrul, S. (2008). *Incomplete acquisition in bilingualism*. Amsterdam: Benjamins.
- Newport, E. M. (1990). Maturation constraints on language learning. *Cognitive Science*, *14*, 11–28.
- Oller, D. K., Pearson, B., & Cobo-Lewis, A. (2007). Profile effects in early bilingual language and literacy. *Applied Psycholinguistics*, *28*, 191–230.

- Owen, A. J., & Leonard, L. B. (2006). The production of finite and nonfinite complement clauses by children with specific language impairment and their typically developing peers. *Journal of Speech, Listening Hearing Research*, *49*, 548–571.
- Paradis, J. (2011). Individual differences in child English second-language acquisition: Comparing child-internal and child-external factors. *Linguistic Approaches to Bilingualism*, *1*. doi: [10.1044/jslhr.4304.834](https://doi.org/10.1044/jslhr.4304.834)
- Paradis, J., Emmerzael, K., & Sorenson Duncan, T. (2010). Assessment of English language learners: Using parent report on first language development. *Journal of Communication Disorders*, *43*, 474–497.
- Paradis, J., & Jia, R. (2016). Bilingual children's long-term outcomes in English as a second language: Language environment factors shape individual differences in catching up with monolinguals. *Developmental Science*, *20*, 1–15. doi: [10.1111/desc.12433](https://doi.org/10.1111/desc.12433).
- Petermann, F., & Petermann, U. (2011). *Wechsler Intelligence Scale for Children*® (4th ed.). Frankfurt a. M.: Pearson Assessment.
- Pierce, L. J., Genesee, F., Delcenserie, A., & Morgan, G. (2017). Variations in phonological working memory: Linking early language experiences and language learning outcomes. *Applied Psycholinguistics*, *38*, 1265–1300.
- Poljšenská, K., Chiat, S., & Roy, P. (2014). Sentence repetition: What does the task measure? *International Journal of Language and Communication Disorders*, *50*, 106–118.
- Restrepo, M. A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language Hearing Research*, *41*, 1398–1411.
- Roth, H. (2018). Sprachliche Bildung und Neuzuwanderung: Auf dem Weg zu einer Didaktik des Deutschen als Zweitsprache im Kontext von Mehrsprachigkeit. In N. Dewitz, H. Terhart, & M. Massumi (Eds.), *Neuzuwanderung und Bildung. Eine interdisziplinäre Perspektive auf Übergänge in das deutsche Bildungssystem* (pp. 196–289). Weinheim, Basel: Beltz.
- Rothman, J., & Kupisch, T. (2018). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *International Journal of Bilingualism*, *22*, 564–582.
- Rothman, J., Long, D., Iverson, M., Judy, T., Chakravarty, T., & Lingwall, A. (2016). Older age of onset in child L2 acquisition can be facilitative: Evidence from the acquisition of English passives by Spanish natives. *Journal of Child Language*, *43*, 662–686. doi: [10.1017/S0305000915000549](https://doi.org/10.1017/S0305000915000549).
- Rothweiler, M. (2006). The acquisition of V2 and subordinate clauses in early successive acquisition of German. In C. Lleó (Ed.), *Interfaces in multilingualism. Acquisition and representation. Hamburg Studies on Multilingualism 4* (pp. 91–113). Amsterdam: Benjamins.
- Roy, P., Chiat, S., & Dodd, B. (2014). *Language and socioeconomic disadvantage: From research to practice*. London: City University London.
- Royal College of Speech and Language Therapists Specific Interest Group in Bilingualism. (2007). *Good practice for speech and language therapists working with clients from linguistic minority communities*. London: Author.
- Scheele, A., Leseman, P., & Mayo, A. (2010). The home language environment of monolingual and bilingual children and their language proficiency. *Applied Psycholinguistics*, *31*, 117–140.
- Schöler, H., Fromm, W., & Kany, W. (Eds.) (1998). *Spezifische Sprachentwicklungsstörung und Sprachlernen. Erscheinungsformen, Verlauf, Folgerungen für Diagnostik und Therapie*. Heidelberg: C. Winter.
- Schönenberger, M., Rothweiler, M., & Sterner, F. (2012). Case marking in child L1 and early child L2 German. In K. Braunmüller, & C. Gabriel (Eds.), *Multilingual individuals and multilingual societies [Hamburg Studies on Multilingualism 13]* (pp. 3–22). Amsterdam: Benjamins.
- Schroeder, C., & Dollnick, M. (2013). Mehrsprachige Gymnasiasten mit türkischem Hintergrund schreiben auf Türkisch. In C. Riemer, H. Brandl, E. Arslan, & E. Langelahn (Eds.), *Mehrsprachig in Wissenschaft und Gesellschaft. Tagungsband* (pp. 101–114). Bielefeld: University of Bielefeld.
- Schroeder, J., & Seukwa, L. H. (2017). Access to education in Germany. In A. Korntheuer, P. Pritchard, & D. Maehler (Eds.), *Structural Context of Refugee Integration in Canada and Germany. GESIS, Volume 15*. Köln: Leibniz Institute for Social Sciences.
- Schulz, P., & Tracy, R. (2011). *Linguistische Sprachstandserhebung—Deutsch als Zweitsprache (LiSe-DaZ)*. Göttingen: Hogrefe Verlag.
- Thordardottir, E. (2015). Proposed diagnostic procedures for use in bilingual and cross-linguistic contexts. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 331–358). Bristol: Multilingual Matters.

- Tuller, L.** (2015). Clinical use of parental questionnaires in multilingual contexts. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 301–330). Bristol: Multilingual Matters.
- Tuller, L., Hamann, C., Chilla, S., Ferré, S., Morin, E., Prévost, P., ... Zebib, R.** (2018). Identifying language impairment in bilingual children in France and in Germany. *International Journal of Language and Communication Disorders*, *53*, 888–904.
- Unsworth, S.** (2016). Quantity and quality of language input in bilingual language development. In E. Nicoladis & S. Montanari (Eds.), *Language and the human lifespan series. Bilingualism across the lifespan: Factors moderating language proficiency* (pp. 103–121). Berlin: Mouton de Gruyter/APA.
- van der Lely, H. K.** (1998). SLI in children: Movement, economy, and deficits in the computational-syntactic system. *Language Acquisition*, *7*, 161–192.
- von Dewitz, N., Terhart, H., & Massumi, M.** (Eds.). (2018). *Neuwanderung und Bildung. Eine interdisziplinäre Perspektive auf Übergänge in das deutsche Bildungssystem*. Weinheim: Beltz.
- von Maurice, J., & Roßbach, H.-G.** (2017). The educational system in Germany. In A. Korntheuer, P. Pritchard, & D. Maehler (Eds.), *Structural context of refugee integration in Canada and Germany* (pp. 49–53). GESIS, Volume 15. Köln: Leibniz Institute for Social Sciences.
- Wegener, H.** (1992). *Kindlicher Zweitspracherwerb. Untersuchungen zur Morphologie des Deutschen und ihrem Erwerb durch Kinder mit polnischer, russischer und türkischer Erstsprache. Eine Längsschnittuntersuchung*. Unpublished habilitation thesis, University of Augsburg, Germany.
- Willis, C. S., & Gathercole, S. E.** (2001). Phonological short-term memory contributions to sentence processing in young children. *Memory*, *9*, 349–363.
- Zebib, R., Henry, G., Khomsi, A., Messara, C., & Hreich, E.** (2017). *Batterie d'Evaluation du Langage Oral chez l'enfant libanais (ELO-L)*. Kerserwan: LTE.
- Zebib, R., Prévost, P., Tuller, L., & Henry, G.** (Eds.) (in press). *Plurilinguisme et Troubles Spécifiques du Langage au Liban*. Beyrouth: Presses universitaires de l'Université Saint Joseph.
- Zebib, R., Tuller, L., Hamann, C., Abed Ibrahim, L., & Prévost, P.** (2019). Syntactic complexity and verbal working memory in bilingual children with and without specific language impairment. *First Language*, *40*, 461–484. doi: [10.1177/014272](https://doi.org/10.1177/014272)

---

**Cite this article:** Hamann C, Chilla S, Abed Ibrahim L, and Fekete I (2020). Language assessment tools for Arabic-speaking heritage and refugee children in Germany. *Applied Psycholinguistics* *41*, 1375–1414. <https://doi.org/10.1017/S0142716420000399>

## Appendix A: Overview of background variables in the bilingual refugee group

| Child (gender) | Family size | Age          | AoO | LoE | SES | Schooling (months) |       |         | Schooling interruption (months) | L1 literacy | Itinerary | Stay in refugee shelter | Current permit |
|----------------|-------------|--------------|-----|-----|-----|--------------------|-------|---------|---------------------------------|-------------|-----------|-------------------------|----------------|
|                |             |              |     |     |     | Syria              | Other | Germany |                                 |             |           |                         |                |
| N1<br>M        | 6           | 100<br>8;4   | 73  | 27  | 18  | 6                  | 7     | 24      | 6                               | no          | plane     | no                      | A              |
| N2<br>M        | 6           | 124<br>10;4  | 97  | 27  | 18  | 18                 | 7     | 24      | 6                               | yes         | plane     | no                      | A              |
| N3<br>F        | 10          | 131<br>10;11 | 111 | 20  | 14  | 12                 | 24    | 20      | 24                              | yes         | plane     | no                      | C              |
| N4<br>M        | 10          | 115<br>9;7   | 95  | 20  | 14  | 0                  | 12    | 20      | 24                              | yes         | plane     | no                      | C              |
| N5<br>F        | 8           | 93<br>7;9    | 72  | 21  | 18  | 0                  | 12    | 21      | 10                              | no          | plane     | no                      | B              |
| N6<br>M        | 8           | 135<br>11;3  | 111 | 24  | 18  | 24                 | 20    | 21      | No                              | yes         | plane     | no                      | B              |
| N7<br>M        | 8           | 97<br>8;1    | 75  | 22  | 8   | 4                  | NA    | 15      | 12                              | no          | sea       | 7 mo.                   | C              |
| N14<br>M       | 8           | 78<br>6;6    | 46  | 32  | 8   | 0                  | NA    | 3       | No                              | no          | sea       | 7 mo.                   | C              |
| N8<br>F        | 4           | 93<br>7;9    | 66  | 27  | 22  | 6                  | NA    | 27      | No                              | no          | plane     | no                      | C              |
| N9<br>M<br>M   | 4           | 93<br>7;9    | 66  | 27  | 22  | 6                  | NA    | 27      | No                              | no          | plane     | no                      | C              |

(Continued)

(Continued)

| Child (gender) | Family size | Age         | AoO | LoE | SES | Schooling (months) |       |         | Schooling interruption (months) | L1 literacy | Itinerary | Stay in refugee shelter | Current permit |
|----------------|-------------|-------------|-----|-----|-----|--------------------|-------|---------|---------------------------------|-------------|-----------|-------------------------|----------------|
|                |             |             |     |     |     | Syria              | Other | Germany |                                 |             |           |                         |                |
| N10<br>M       | 5           | 138<br>11;6 | 120 | 18  | 8   | 38                 | NA    | 18      | 18                              | yes         | sea       | 6 mo.                   | C              |
| N11<br>F       |             | 138<br>11;6 | 120 | 18  | 8   | 38                 | NA    | 18      | 18                              | yes         | sea       | 6 mo.                   | C              |
| N12<br>F       | 3           | 152<br>12;8 | 111 | 41  | 14  | 20                 | NA    | 36      | 6                               | yes         | sea       | 3 mo.                   | C              |
| N13<br>F       | 6           | 145<br>12;1 | 106 | 39  | 18  | 24                 | NA    | 36      | 6                               | yes         | sea       | 3 mo.                   | C              |
| N15<br>F       | 6           | 88<br>7;4   | 56  | 32  | 7   | NA                 | NA    | 7       | 6                               | yes         | sea       | 6 mo.                   | C              |

## Appendix B: Standardized tests used to assess cognitive abilities and language skills in Arabic and German

| Language       | Test                                      | Language skill tested |                      |                       |   |   | Scoring system                             | Norm group                                     |
|----------------|---|-----------------------|----------------------|-----------------------|---|---|--|--|
|                |   | Phonology             | Vocabulary reception | Vocabulary expression | Morphosyntax comprehension                      | Morphosyntax production                       |  |  |
| IQ             | Colored Progressive Matrices <sup>a</sup> |                       |                      |                       |   |   |  | 3;9–11;8                                       |
| Working memory | BDS/FDS from WISC-IV <sup>b</sup>         |                       |                      |                       |   |   |  | 6;0–16;11                                      |
| Arabic         | ELO-L <sup>c</sup>                        | 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 <sup>d</sup>                     | —                     | Picture selection    | Picture naming        | —   | —   | Individual subtest scores                  | 5;6–10;11                                      |
|                | TROG-D <sup>e</sup>                       |                       |                      |                       | Picture-sentence matching                       |   | Individual scores,                         | 3;0–10;11 and adults                           |
|                | LiSe-DaZ <sup>f</sup>                     | —                     | —                    | —                     | Picture-sentence matching, Truth value judgment | Story, sentence completion, lead-in questions | Individual subtest scores                  | 3;0–6;11 (monolinguals), 3;0–7;11 (bilinguals) |

Note: <sup>a</sup>Bulheller & Häcker, 2002; <sup>b</sup>Fox 2009; <sup>c</sup>Glück, 2011; <sup>d</sup>Petermann & Petermann, 2011; <sup>e</sup>Schulz & Tracy, 2011; <sup>f</sup>Zebib, Henry, Khomsi, Messara & Hreich, 2017.