Language Assessment of Bilingual Arabic-German Heritage and Refugee Children: Comparing Performance on LITMUS Repetition Tasks

Lina Abed Ibrahim, Cornelia Hamann, and István Fekete

1. Introduction and theoretical background

In the last four decades, bilingualism and multilingualism have been in the focus of linguistic research with manifold aims, such as the exploration of a human disposition for multilingualism or of the factors determining trajectories and outcomes of (second) language learning. In Germany, research questions were heavily influenced by socio-political developments as the country has a long history of immigration and experience with heritage and refugee children in the educational system. The first studies on the untutored acquisition of German as a second language (L2) concerned adult learners (Clahsen, Meisel & Pienemann, 1983) while subsequent work focused on language development in the second generation, mostly children growing up with two languages from birth (2L1) or learning German at kindergarten age (eL2) (Meisel, 1990).

Regardless of the arrival of two major waves of refugees (from the Balkans in the 1990s and from Syria in 2015) and the heterogeneity of bilingual populations as to age and mode of acquisition, the majority of the studies interested in linguistic development investigate children and adults who are 2L1 or eL2 speakers. Most linguistic long-term research projects up to the present base their findings on such individuals or groups (see Chilla, 2008; Gagarina, 2017; Hamann & Abed Ibrahim, 2017; Rothweiler, 2006; Tuller et al., 2018 among many others). The two language assessment tools developed in Germany, standardized for monolingual and bilingual children, Russian SKRUK (Gagarina et al., 2010) and LiSe-DaZ (Schulz & Tracy, 2011), likewise are normed for these populations only. The same holds for most research on academic success of bilinguals in German schools with many studies investigating reading and writing skills based on 2L1 and eL2 children (Chilla, Krupp & Wulff, 2019; Gantefort & Roth, 2008).

* Lina Abed Ibrahim, Cornelia Hamann and István Fekete: University of Oldenburg. Corresponding author: Lina Abed Ibrahim, lina.abed.ibrahim@uol.de. This research was carried out within the BiliSAT- project financed by DFG grants to S. Chilla (CH 1112/4-1) and C. Hamann (HA 2335/7-1). We are grateful to Solveig Chilla for her support with data collection and analysis. Our special thanks go to the participating children and their parents and collaborating teachers and SLTs.

Due to the influx of almost a million refugees in September 2015, many thousands of children and adolescents with different ages of arrival, different schooling experience, different language backgrounds, different asylum status and different itineraries had to be integrated into the German school system which widely differs from state to state, see von Maurice & Roßbach (2017). Since knowledge of the German language is undoubtedly a necessary condition for academic and later professional success, these refugee children learn German in different programs and with different curricula all over the country (Ahrenholz, Fuchs & Birnbaum, 2016). Importantly, irrespective of different models of schooling in the 16 Federal States of Germany, the general design already established in the 1980s, based more or less on study of simultaneous or early successive bilingualism, remains: refugee children are expected to acquire oral and literacy L2 skills within six to eighteen months before they are to participate in regular classes. Adding to the problem is that most schools establish separate language classes irrespective of the students’ age or overall development (for details see Massumi & v. Dewitz 2015).

The short time (6-18 months) allotted to L2-acquisition is clearly problematic since more pessimistic estimates of the time needed to catch up in important areas of language are found in the recent literature (see Schönenberger, Rothweiler & Stern, 2012; Paradis & Jia, 2016). First results comparing bilingual children born in Germany (primary school age) with refugees (10-17 years of age) with respect to their vocabulary development and their ability to write an essay show a big gap in language skills between the two groups (Montanari, 2017). This study also shows that it is not easy to choose and apply language assessment tools and that the lack of bilingual norms is an additional challenge when using standardized methods. The above considerations will have made it quite evident that systematic studies on the language development of refugees who enter the school system when they are older than seven years are needed urgently not only for guiding the choice of assessment tools but ultimately for shaping expectations and models of participation.

It is at this point that our investigation makes a contribution. We compare a group of Syrian refugees with different educational backgrounds to younger Arabic heritage speakers, i.e. 2L1 or eL2 speakers. In the wake of COST Action IS0804, cross-linguistically valid tools were developed. Since non-word and sentence repetition have high accuracy for diagnosing Developmental Language Disorder (DLD) in monolinguals (Conti-Ramsden et al., 2001), these LITMUS tools (Language Impairment Testing in Multilingual Settings, Armon-Lotem et al., 2015) included Nonword Repetition Tasks (NWRT; Grimm et al., 2014) and Sentence Repetition Tasks (SRT, Marinis & Armon-Lotem, 2015 Hamann et al., 2013). Both these LITMUS-tools have recently been shown to identify DLD in bilingual populations and proved to have reasonable to good – in combination even excellent - diagnostic accuracy for the identification of DLD in 2L1 and eL2 children (Abed Ibrahim & Fekete, 2019; Armon-Lotem & Meir, 2016; Chiat & Polišenská, 2016; Hamann & Abed Ibrahim, 2017; Tuller et al., 2018). This is an
encouraging result since the tasks are easy and fast to administer and could thus serve as a first assessment in many contexts, including schools.

Repetition tasks have been associated with verbal working memory (VWM) and verbal short-term memory (VSTM) in so far as NWRTs are usually taken to measure VSTM (Archibald & Gathercole, 2006; Bishop, Adams & Norbury, 2006). 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, Polišenská, Chiat & Roy, 2014; or Klem et al. 2014 and Meir, 2017 for SRT). Recently, Abed Ibrahim & Hamann (2017) confirmed that linguistic complexity has a decisive influence on performance on both German SRT and NWRT. Moreover, a detailed investigation of the factors influencing performance of bilingual children on SRT in French has shown that differences in performance of the typically developing (TD) group and the group with DLD cannot only be explained by differences in VSTM or VWM but crucially depend on syntactic complexity (Zebib et al., 2019). In line with these results from previous studies, it can be stated that the German LITMUS-NWRT and SRT as well as the Arabic SRT, which was developed in parallel to the LITMUS French SRT (see Henry et al., to appear), were constructed as language measures and include structures that are known to be difficult for children with DLD.

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). Such background information allows exploring factors such as Early Development, socio-economic status (SES), age of onset of L2 (AoO) and length of exposure to L2 (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 DLD including age of first words and first sentences, and not language exposure and use, generally explained more of the variance in the performance in the German NWRT and SRT.

Clearly, in a refugee context, other factors such as interrupted education, living and study conditions and trauma incurred during flight could influence performance in language tasks. Thus, questions about schooling, access to language courses, itinerary, means of transport and past and present living conditions were added to the PaBiQ for the purposes of this study.

In this paper, the investigation includes a group of refugees. On the one hand, we want to ensure that a child shows typical language acquisition and can therefore be expected to encounter no extraordinary difficulties in her second language development. On the other hand, we want to make sure that, if there is an impairment, the child will be diagnosed and will have access to language therapy. This is even more important 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. This aim can be achieved by assessing the home language of a child, which, in the case of refugees, should be the dominant language and should show a possible impairment. Comparing L1
development in refugee and heritage children in more depth will enable us to judge whether certain L1 tests are fair for both groups, and in particular for heritage children, who might not (or no longer) be dominant in their L1 after extensive L2-exposure in kindergarten or school. Heritage children might thus be at a disadvantage in L1 tests given the problems with L1-maintenance in heritage speakers (Montrul 2008). We will therefore analyze performance in a standardized Arabic language assessment tool, as well as in a LITMUS SRT in Arabic (Zebib et al., to appear).

As teachers or speech and language therapists capable of assessing the L1 in bilinguals are rare, application of German (L2) assessment tools is inevitable. Since the LITMUS-SRTs and NWRTs have shown promising results for heritage children, this study asks whether the same tests render reliable and fair assessments of language abilities also for refugee children and explore how much language exposure is required to achieve fair results.

In addition, we want to know which factors influence variance of performance. Since it has been discussed that performance in L2 repetition tasks can be influenced by AoO, LoE, and SES, but crucially by morphological and lexical knowledge (see Chiat & Polišenská, 2016 on properties of differently constructed NWRTs and Tuller et al., 2018 for an investigation of SRTs), we want to investigate in how far such factors, but also quality of input, current use and WM factors, might influence performance. In particular, we want to explore whether the tests allow a fair and quick classification after an exposure of more than a year. Final points of investigation are in how far early and present language exposure/use in L1 or L2 determine later language learning and lead to an advantage in L2 development.

2. Method
2.1. Participants

This study compares 11 2L1 and eL2 children (5;6-9;0) with Arabic as L1 (Lebanese) to 11 refugee children from Syria (7;7-11;6) who all 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 or schools for at least 18 months. All participants scored above percentile rank 9 (IQ score ≥ 80 according to Wechsler’s IQ scale) in the German version of Raven’s Colored Progressive Matrices (Bulheller & Häcker, 2002), and were not at risk of DLD. The participants of the heritage group had previously been classified as typically developing children using a comprehensive assessment procedure outlined in Hamann and Abed Ibrahim (2017), including standardized tests in the L1 and L2, respecting dominance effects on test performance (Thordardottir, 2015). For the refugee participants, we do not refer classification to L2-tests, but rely on a standardized L1-test given their relatively short exposure to the L2.

Participants were recruited in kindergartens, schools, community associations or places of worship from different federal states, representing a spectrum of schooling in the L2. The refugee children in our study are typical for the diversity
of programs since they are resident in different parts of Germany. Bremen, for example offers a preparatory language support class before children switch to the regular class. North Rhine-Westphalia offers language support classes and, simultaneously, courses in regular education. Many primary school children between 6 and 8 years of age attend regular classes from the beginning and receive additional L2 support in separate courses. Table 1 gives a participant overview.

<table>
<thead>
<tr>
<th>(Mean, SD and range)</th>
<th>Heritage speakers (n=11)</th>
<th>Syrian Refugees (n=11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at testing (months)</td>
<td>88,4 (12,8), 70-108</td>
<td>114 (19,5), 91-138</td>
</tr>
<tr>
<td>Age of Onset (months)</td>
<td>39 (13,6), 24-75</td>
<td>91,5 (21,7), 66-120</td>
</tr>
<tr>
<td>Length of Exposure, (months)</td>
<td>53,2 (19,7), 32-97</td>
<td>22,8 (3,7), 18-27</td>
</tr>
<tr>
<td>Socio-economic status, SES (mother’s education in years)</td>
<td>13,7 (2,2), 9-18</td>
<td>15,3 (5,3), 8-22</td>
</tr>
<tr>
<td>Length of L1- schooling (months)</td>
<td>0,63 (0,5), 0-2</td>
<td>11,63 (11,5), 0-32</td>
</tr>
<tr>
<td>Gender</td>
<td>3 M, 8 F</td>
<td>7 M, 4 F</td>
</tr>
</tbody>
</table>

2.2. Procedures and assessment tools

In addition to standardized L1 and L2 tests, all children performed German versions of the LITMUS NWRT and SRT, as well as the Arabic LITMUS-SRT. Relevant background information was gathered with the Questionnaire for Parents of Bilingual Children (PaBiQ; Tuller, 2015).

We also chose two working memory measures, Forward Digit Span (FDS, WISC-IV, Petermann & Petermann, 2011), associated with VSTM, and Backward Digit Span (BDS) involving storage and processing and assessing VWM.

2.2.1. Standardized L1 and L2 tests

The lack of age adequate tests and standardized tools posed a serious problem for background assessment for this investigation, which includes children up to the age of 12 years. Nevertheless, the following standardized language tests were used. For assessing Arabic, we used the ELO-L (Zebib et al., 2017), standardized and normed in Lebanon on large samples of children growing up in the kind of institutional bilingualism typical for Lebanon. It was adapted to other varieties of Arabic by linguistically trained native speakers of these varieties including the Syrian Arabic version used here. Since the age-range of the norming population is 3;0-7:11, norm-referencing was not applicable to a subset of our sample.

We encountered similar difficulties for some standardized L2 tests. For assessment of vocabulary, we used WWT (Glück, 2011), which provides norms
for monolingual children (5;6 -10;11). For morphosyntax, we used the LiSe-DaZ (Schulz & Tracy, 2011). The test is normed for eL2 bilinguals as well as monolinguals. Norms for bilinguals are available for the ages 3;0 till 7;11. 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. For the older group of refugees we also used the TROG-D (Test for Reception of Grammar – German, Fox, 2009). The TROG-D can be reliably used as a test for language development in the age range of 3;0 – 9;11. For L1, we expect ceiling effects in 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 dominance adjusted norm in the oldest age range with an available norm.

Since there is no norming sample matching our oldest children, raw values are additionally used for meaningful group comparisons for the L1 and L2 standardized tests in line with other researchers (Montanari, 2017).

2.2.2. The German LITMUS NWRT

The NWRT chosen here (Grimm et al., 2014) relies on increasing phonological complexity, not increasing numbers of syllables (dos Santos & Ferré, 2018; Grimm & Schulz, 2020). Using the most common vowels and consonants of the world’s languages, it presents one-, two- and three-syllable non-words of different phonological complexity with complex onsets or codas and combinations thereof (see Abed Ibrahim & Fekete, 2019; Grimm & Schulz, 2020 for more detailed descriptions and analysis of the task’s properties). The items are presented to the child in an appealing PPT in pseudo-randomized order through headphones. The task takes about 5-10 minutes to administer and is scored according to whole item accuracy. Minimally different vowels and, crucially, voicing of consonants are nonetheless not counted as errors in order not to disadvantage bilingual children.

2.2.3. The German and the Arabic LITMUS SRTs

The German SRT (Hamann et al., 2013) and the Lebanese Arabic SRT (Henry et al., in press) target children between 5;6 and 9;0. 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 DLD in the cross-linguistic literature. In addition to simple sentences varying in agreement and/or tense, Marinis & Armon-Lotem (2015) recommend to include object (which) questions, passives, finite complement clauses, (object) relative
clauses and topicalizations, which are all included in the German version (for more details and examples see Hamann & Abed Ibrahim, 2017). 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, non-finite and finite complement clauses as well 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 additionally included the sentence bracket, which is considered a milestone in the acquisition of German as early L2. An overview of the common structures in Arabic and German LITMUS-SRTs is given in Table 2.

The current versions of the German and the Arabic SRT include 45 and 36 sentences respectively. The stimuli are presented via a child friendly PPT in randomized order. Administration of the tasks takes 5-10 minutes. The tasks can be rated by “identical repetition, SRT_Id” counting only exact repetition as correct, disregarding only phonological errors, or it can be rated by “target structure, SRT_Tar”, 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, incorrect case if it is not crucial for the target structure.

### Table 2. Common structures in German and Arabic LITMUS-SRTs

<table>
<thead>
<tr>
<th>target structure</th>
<th>less complex</th>
<th>more complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>monoclausal (SVO)</td>
<td>present (imperfective)</td>
<td>past (perfective)</td>
</tr>
<tr>
<td>Object wh-question</td>
<td>Who-questions (German)</td>
<td>Which NP-questions</td>
</tr>
<tr>
<td>Clausal embedding</td>
<td>- finite embed.</td>
<td>+ finite embed.</td>
</tr>
<tr>
<td>Relative clause</td>
<td>Subj. relatives</td>
<td>Obj. relatives</td>
</tr>
</tbody>
</table>

### 2.3. Data Analysis

All standardized tests, as well as the German LITMUS-NWRT and the German and Arabic LITMUS-SRTs were administered as per description (cf. Hamann & Abed Ibrahim, 2017; Tuller et al., 2018, Zebib et al. to appear). As already indicated, norm-referencing was possible only for certain age groups in the different standardized tests. Thus basic comparisons will be performed on raw values (percent-correct scores) as well.

IBM SPSS 26 (IBM Corp. Released, 2019) and R-Studio (R-Studio Team, 2015) were used to conduct the statistical analyses. To determine factors influencing performance, we developed a series of linear mixed-effects (LME) regression models (Pinheiro & Bates, 2000), employing the lme4 R-package (Bates et al., 2015). An advantage of LME models is that they allow for the comparison of significant predictors in terms of their contribution or weight in explaining the dependent measure, just as in the case of a simple multiple regression model. To that end, we report on the Beta-estimates per significant predictor.
3. Results

3.1. Standardized L1-Test (ELO-L)

For the L1, the refugee group showed higher lexical skills, both for receptive (raw scores: \( U = 12.0, p = .001, r = .685 \), Z-scores: \( U = 10.0, p < .001, r = .709 \)) and expressive vocabulary (raw scores: \( U = 17.5, p < .01, r = .604 \), Z-scores: \( U = 16.0, p < .01, r = .624 \)). Furthermore, the refugee children outperformed their heritage peers on morphosyntax production (raw scores: \( U = 6.00, p < .001, r = .765 \), Z-scores: \( U = 17.0, p < .01, r = .610 \)). Interestingly, both groups performed on par in morphosyntax comprehension (raw scores: \( U = 55.0, p = .748 \), Z-scores: \( U = 49.5, p = .478 \)) and in phonology if raw scores (but not Z-scores) are considered. Mean raw scores (% correct) and Z-scores for the ELO-L subtests are given in Figure 1a and Figure 1b.

![Fig. 1a. ELO-L (raw scores)](image)

![Fig. 1b. ELO-L (Z-scores)](image)

3.2. Standardized L2-Tests

With regard to L2-lexical abilities, both groups performed poorly in both expressive and receptive subparts of the WWT. In fact, no significant group-differences emerged and both performed within the language impaired range if standardized Z-scores are considered, even in receptive vocabulary (Heritage: \( M = -2.83, SD = 1.42 \) vs. Refugee: \( M = -2.57, SD = 2.14 \)).
Considering morphosyntax, while heritage children were able to reach age appropriate norms in the LiSe-DaZ, more than half of the refugee children (6/11) would be classified as at risk for DLD by the test with only children with more than 24-27 months of L2-exposure reaching appropriate norms. Similar results were observed for performance of the refugee group on TROG-D applying dominance-adjusted cut-offs: all children with an LoE<24 months (5/11) were identified as at risk of DLD.

3.3. LITMUS Repetition Tasks
3.3.1. Arabic LITMUS-SRT

As can be seen in Figure 2, heritage children were not disadvantaged by Arabic LITMUS-SRT unlike in most of 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 the task, especially on SRT_Ar_Tar (Heritage: \( M = 85.01, SD = 15.9 \); Refugee \( M = 94.4, SD = 10.9 \)), no significant between-group differences were found for either SRT_Ar_Id or SRT_Ar_Tar.

![Fig. 2. LITMUS-Ar_SRT: %correct identical rep. & target structure](image)

3.3.2. German LITMUS NWRT and LITMUS-SRT

No significant group differences emerged between heritage and refugee children on LITMUS NWRT. As for the German-SRT (Figure 3), the heritage group performed generally better and showed less variance on both SRT_Id (Heritage: \( M = 60.2, SD = 24.9 \); Refugee \( M = 49.5, SD = 31.3 \)) and SRT_Tar (Heritage: \( M = 78.9, SD = 17.8 \); Refugee \( M = 65.6, SD = 23.9 \)). However, these differences did not reach statistical significance. When children’s performance on both SRT measures is plotted against LoE to the L2 (see Figure 4), it becomes apparent that most refugee children with less than 24 months of exposure perform below the cut-off scores separating typically developing from language impaired children in Hamann and Abed Ibrahim (2017). Note that these cut-offs were determined for a population of younger bilinguals (5;6-9;0) including a subset of the present heritage sample.
3.4. 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 individual variation in the current paper is limited to performance in the LITMUS-SRTs.

Background, WM (FDS & BDS) and linguistic variables previously discussed in the literature were probed as potential predictors using nonparametric Spearman-correlations. Only variables yielding significant bivariate correlations with the dependent performance variables SRT_Ar_Id, SRT_Ar_Tar, SRT_Id and
SRT_Tar were selected as fixed factors for a series of linear mixed-effects (LME) regression models. Although mean age differed significantly between the two groups, it did not show significant correlations with the dependent measures. Thus, it was not assessed as a random factor. The variable SES was correlated with several dependent measures. Therefore, it was entered as a random factor into all models to partial out its effect. Further factors such as L1/L2 schooling were also included as random factors. To avoid model-overfitting, the number of fixed factors per model was limited to three. For the sake of simplicity, we mainly report on the models with significant fixed factors.

3.4.1. Predictors of performance on the Arabic LITMUS-SRT

As for non-linguistic background variables, length of L1-schooling, and current L1 use as well as L1-linguistic richness significantly predicted performance on both measures SRT_Ar_Id and SRT_Ar_Tar (LME models 3.1, 3.2, 3.3). The only linguistic variable that emerged as a significant predictor of performance on both measures of Arabic SRT was L1-receptive morphosyntax (LME model 3.4). Notably, no significant effects were observed for AoO_L2, LoE_L2, L1-vocabulary, morphosyntax in production or phonology.

Table 3. LMEMs analysis: Predictors of performance on Arabic-SRT

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>SRT_Ar_Id</th>
<th>SRT_Ar_Tar</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Background variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 3.1: Fixed factors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(LoE_L2, current L1 use &amp; length_L1_schooling)</td>
<td>Length_L1_schooling: $\beta = .777, SE = .360$, $T(13.5) = 2.15, p = .049^*$</td>
<td>Length_L1_schooling: $\beta = .736, SE = .301$, $T(13.1) = 2.44, p = .029^*$</td>
</tr>
<tr>
<td>Random factors: SES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 3.2: Fixed factors</td>
<td>a. Linguistic richness L1: $\beta = 5.04, SE = 1.89, T(15.9) = 2.66$, $p = .017^*$</td>
<td>a. Linguistic richness L1: $\beta = 3.58, SE = 1.62$, $T(16.0) = 2.20$, $p = .042^*$</td>
</tr>
<tr>
<td>(AoO_L2, current L1 use &amp; linguistic richness L1)</td>
<td>b. Current L1 use: $\beta = -4.35, SE = 1.41$, $T(14.6) = .309$, $p = .007^{**}$</td>
<td>b. Current L1 use: $\beta = -3.16, SE = 1.21$, $T(15.1) = -2.60$, $p = 0.019^*$</td>
</tr>
<tr>
<td>Random factors: SES+L1-schooling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 3.3: Fixed factors</td>
<td>a. Linguistic richness L1: $\beta = 5.30, SE = 1.89, T(15.9) = 2.79$, $p = .012^*$</td>
<td>Current L1 use: $\beta = -3.42, SE = 1.22$, $T(15.4) = -2.80$, $p = .013^*$</td>
</tr>
<tr>
<td>(LoE_L2+current L1 use+linguistic richness L1)</td>
<td>b. Current L1 use: $\beta = -4.10, SE = 1.43$, $T(15.2) = -2.85$, $p = .011^*$</td>
<td></td>
</tr>
<tr>
<td>Random factors: SES+L1-schooling</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>WM and linguistic variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 3.4: Fixed factors</td>
<td>MorphR_ELO-L: $\beta = 2.88, SE = 0.67, T(15.5) = 4.28$, $p &lt; .001^{***}$</td>
<td>MorphR_ELO-L: $\beta = 2.65, SE = 0.45$, $T(15.2) = 5.58$, $p &lt; .001^{***}$</td>
</tr>
<tr>
<td>(FDS_span+BDS_span +MorphR_ELO-L)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random factors:SES</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
3.4.2. Predictors of performance in the German SRT

As can be seen in LME models 4.1-4.3, the only background variable that emerged as a significant predictor for performance in both SRT_Id and SRT_Tar was current L2 use.

As for WM and L2 language measures, expressive vocabulary skills emerged as a significant predictor for performance on SRT_Id, while a combination of FDS and L2 expressive vocabulary accounted for performance on SRT_Tar (LME model 4.4). With respect to L2-morphosyntax, the performance in the subtest assessing German sentence complexity level (ES “Entwicklungsstufe”, i.e. developmental stage) combined with FDS was predictive of performance on SRT_Tar but not on SRT_Id (LME model 4.5). Plotting performance in SRT_Tar against the developmental level showed that only children who reached developmental level 4, i.e. produced embedded clauses, were capable of performing above the pathology cut-offs established for younger bilinguals in Hamann and Abed Ibrahim (2017).

Table 4. LMEMs analysis: Predictors of performance on German-SRT

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>SRT_Id</th>
<th>SRT_Tar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random factor: SES</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Model 4.1: Fixed factors</strong> (AoO_L2, current L2 use &amp; linguistic richness L2)</td>
<td>Current L2 use: β = 6.96, SE = 2.36, T(17.9)= 2.94, p = .008*</td>
<td>Current L2 use: β = 4.39, SE = 1.93, T(17.7) = 2.27, p = .036*</td>
</tr>
<tr>
<td><strong>Model 4.2: Fixed factors</strong> (AoO_L2, current L2 use &amp; L2_schooling_months)</td>
<td>Current L2 use: β = 5.86, SE = 2.11, T(17.9) = 2.77, p = .012*</td>
<td>Current L2 use: β = 4.35, SE = 1.68, T(17.9) = 2.58, p = .018*</td>
</tr>
<tr>
<td><strong>Model 4.3: Fixed factors</strong> (LoE_L2, current L2 use &amp; linguistic richness L2)</td>
<td>Current L2 use: β = 5.34, SE = 2.29, T(17.97) = 2.32, p = .031*</td>
<td>n.s.</td>
</tr>
<tr>
<td><strong>WM and linguistic variables</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| **Model 4.4: Fixed factors** (FDS_span, BDS_span & LexP_L2) | LexP_L2: β = 2.48, SE = 1.07, T(12.6) = 2.32, p = .03* | a. FDS_span : β = 6.88, SE = 2.59, T(10.9) = 2.65, p = .022*  
  b. LexP_L2: β = 2.40, SE = .72, T(13.0) = 3.32, p = .005** |
| **Model 4.5: Fixed factors** (FDS_span, BDS_span & ES) | n.s. | a. LIS_ES: β = 21.603, SE = 9.84, T(14.8) = 2.19, p = .044*  
  b. FDS_span : β = 10.3, SE = 3.39, T(14.1) = 3.04, p = .008** |

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
As for TROG-D, all refugee children (5/11), who were identified as at risk of DLD applying dominance-adjusted cut-offs, performed below the respective pathology cut-offs in SRT_Id and SRT_Tar.

4. Discussion and conclusion

This study confirms previous research showing that the German LITMUS-NWRT is suitable for all groups of bilingual children (Grimm and Schulz, 2020; Hamann & Abed Ibrahim, 2017) whereas the German LITMUS-SRT leads to accurate results only after longer exposures. The German LITMUS-SRT thus very reliably assesses heritage children (see also Abed Ibrahim & Fekete, 2019) but can at best be indicative during the first 24 months of exposure in case of refugee children.

When other morphosyntactic measures were investigated, it turned out that only children who were able to produce subordinate clauses in the subtest on sentential complexity of the LiSe-DaZ, were capable of performing reasonably in the L2 SRT. This is in line with previous studies showing that eL2 learners acquire the German sentence structure in main clauses fast, but take longer to acquire verb placement in subordinate clauses. These findings clearly demonstrate that refugee children, during the first 24 months of exposure will have difficulties with tests that include such milestones of L2 acquisition (Clahsen et al., 1983). This is unfortunate since 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. It could be established that the same children diagnosed as DLD on TROG-D 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. This aligns with previous findings, also on spontaneous speech data, in that refugee children perform in the range of children with DLD (Chilla, 2008; Håkansson & Nettelbladt, 1993; Schöler, Fromm & Kany, 1998).

These findings are complemented by the results on L1 assessments. Here, the refugee children have an advantage over the heritage children in most standardized measures. They perform on par with the heritage children on the Arabic LITMUS-SRT, however. Interestingly, the only linguistic measure predicting performance on both measures, identical repetition and target structure, is the performance in L1-receptive morphosyntax. Since performance was on par for this measure in both groups, we can conclude that the Arabic SRT is fair, not only for refugee children, but also for heritage children at school age. Our analyses also showed that neither working memory, nor L2-exposure variables (AoO_L2, LoE_L2) predicted performance in the Arabic SRT. The latter results indicate that heritage children are not disadvantaged and that the Arabic LITMUS-SRT can be applied with both groups. 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 will have to include L1-measures, particularly in refugee populations.

Using L2 and L1 measures for bilinguals has long been 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 also reliably measures morphosyntactic competence, which, unfortunately, does not develop as fast as many schooling and integration models presuppose. This, in particular, is a result which needs to be discussed widely and which should be used to correct the often too optimistic expectations of teachers and educators.

References


