ELO-L: A Norm-Referenced Language Screening Test for 3 to 8-Year-Old Lebanese Children

Arab Journal of Applied Linguistics
e-ISSN 2490-4198
Vol. 4, No. 2, July 2019, 24-53

http://www.arjals.com/

Racha Zebib[1], UMR 1253, iBrain, Université de Tours, Inserm, Tours, France.
Guillemette Henry, Camille Messarra & Edith Kouba Hreich.
Institut Supérieur d'Orthophonie, Saint Joseph University of Beirut, Lebanon.
Abdelhamid Khomsi, emeritus professor of psycholinguistics at the University of Tours.

Abstract

The ELO-L (Évaluation du langage oral chez l'enfant libanais) is the first norm-referenced language-screening test in Lebanon. It is an adaptation of the ELO, a French language-screening test. The ELO-L was normed on 1,718 children aged three to eight years and divided into eight age groups with a minimum of 100 participants in each group. It is composed of five subtests targeting receptive vocabulary, expressive vocabulary, sentence comprehension, sentence production and expressive phonology. We explain how the test was adapted for Lebanese, and present the subtests, the scoring method and the normative sample. We furthermore give the first validation results, reporting on developmental sensitivity, reliability, concurrent validity and diagnostic accuracy.

Keywords*:* Lebanese Arabic, Language Assessment, Psychometric Testing

[1] racha.zebib@univ-tours.fr

Introduction

Speech and language disorders are described as communication disorders characterized by reduced vocabulary, limited sentence structure, impairments in discourse and persistent deficits in speech sound production (American Psychiatric Association, 2013). The diagnosis of these disorders depends mainly on a thorough assessment of language abilities. According to the American Psychiatric Association, language assessments should take into consideration the cultural and linguistic environment, especially in bilingual contexts. Indeed, several studies have shown that bilingualism may affect performance in a given language (Armon-Lotem, de Jong & Meir, 2015; de Jong, Cạvuṣ & Baker. 2010; Genesee, Paradis, & Crago. 2004; Paradis, 2010a; Paradis, Genesee, & Crago, 2011) and using assessment tools standardized on monolingual populations to assess bilingual children is a source of misdiagnoses (Paradis, 2010b). Therefore, taking into account bilingualism, while developing language tests, is essential in multilingual contexts (see Armon-Lotem et al., 2015).

Lebanon is a multilingual country in which the majority of the population uses Lebanese Arabic (Northern Levantine Arabic). A considerable proportion of the population is also exposed to Modern Standard Arabic (MSA) and more than half are exposed to one or two other languages, mainly French and/or English. Other minority languages, such as Kurdish and Armenian, also exist (Verdeil, Ghaleb, & Velut, 2007). It is important to emphasize that second-language learning is obligatory in Lebanese schools, regardless of the type of school: public schools, private fee-paying schools or free private schools. Private fee-paying schools are the schools with the largest student population (more than half of Lebanese students) in Lebanon (Gouteyron et al., 2000-2001, Hoyek, 2004). These schools are considered to provide higher quality education than public schools (Verdeil, Faour & Velut, 2007), which rank second in terms of number of students. Free private schools are very few in number and educate relatively few students (Gouteyron et al., 2000-2001). National law imposes bilingual education in public schools starting, at the latest, in first grade (age 6). Private schools generally introduce instruction in a second language in kindergarten. Some schools introduce a third language during the school curriculum (Shaaban, 2017). In general, the number of

instruction hours in a language other than Arabic is equivalent to or greater than the number of hours in Arabic, depending on the particular school. English and French are the two most widely used languages of instruction, besides Arabic, in all school types. Other languages of instruction are also used (e.g.,, Armenian, German, etc.). However, these languages are often not taken into account in official statistics, as they concern a small minority of schools.

According to the Lebanese Ministry of Education and the French Mission for Information on Cultural, Scientific and Technical Relations with Lebanon, Syria and Jordan, 69.5% of Lebanese students learn French and the rest learn English in addition to Arabic. It is important to note that these percentages, established in 1995-1996, have evolved since then in favor of English (Gouteyron et al., 2000-2001). According to the Center for Educational Research and Development (Ministry of Education), in 2013-2014, 54.9% of Lebanese schools had French as language of instruction against 45% for English. English appears to be continuing to gain ground over French (Shaaban, 2017). We note that private fee-paying schools are mostly attended by children from middle to upper classes, while public schools are predominant in less privileged environments (Verdeil et al., 2007). Moreover, the importance given to second language learning and the frequency of second language use seem to be related to the sociocultural environment in Lebanon. The choice of the particular second language (mainly French or English) also seems to be related to political or even religious factors (Makki, 2007).

In sum, Lebanese children are generally exposed to two or more languages at a very early age. Assessment tools used to evaluate language in children growing up in Lebanon should, therefore, take into account multilingualism. Moreover, these tools should consider the cultural and socioeconomic variability in these children, as recommended by the American Psychiatric Association, and take into consideration the fact that bilingualism in Lebanon seems to be modulated by cultural and socioeconomic factors. More generally, taking into account these factors is consistent with studies that have shown that socioeconomic factors may have an effect on some aspects of language performance (see, Letts, 2013 for a review). Furthermore, to ensure reliable diagnoses, language assessments must be based on standardized tests targeting different language

domains, in both receptive and expressive modalities (see the diagnostic criteria proposed by Tomblin, Records & Zhang, 1996). In order to be useful for diagnosis, language tests must possess important psychometric characteristics (see Chartier & Loarer, 2008 for a review):

1) Standardization: This refers to the use of strict administration and scoring methods. The aim of standardization is to guarantee that the child being assessed is evaluated in the same conditions as the normative sample.

2) Norm-reference: This corresponds to scores on the test obtained by a reference group (normative sample), which has the same characteristics of the child being assessed (age, gender, etc.). The norms will serve to evaluate the performance of the child being tested in comparison to the reference group (e.g., children of the same age). Note that, when norms are established as a function of age, a test must show good developmental sensitivity.

3) Test reliability: This refers to the stability of the test. There are different measures of reliability: Test-retest reliability, internal reliability and inter-rater reliability.

Test-retest reliability measures the consistency of the results given by a test when used with the same participant(s) over time. It aims to check if the results obtained for a given participant are reproducible and therefore stable. In other words, it measures the resistance of a test to the effects of participant internal (e.g., fatigue) or external (e.g., time of day) factors related to the moment (s)he was tested. It is generally measured via correlation analyses (e.g., Pearson correlation coefficient) applied on the scores of the same participant or the same group of participants over a period of time that does not exceed six months (time after differences in scores might be expected due to development).

Internal reliability measures the internal homogeneity of a test in order to see if the different items of the same test assess the same skill or variable. It is generally measured via Cronbach's alpha coefficient applied on the scores of a group of participants on the different items of a test.

Inter-rater reliability measures the stability of the scores as a function of raters. In other words, it aims to check if the standardization characteristics of the test (e.g.,

administration and scoring instructions) are precise enough so that different raters have similar ratings of the performance of the same participant. Inter-rater reliability is preferentially measured via interclass correlation (ICC) (see Lander, 2015).

4) Validity evidence: it corresponds to the ability of the test to measure what it aims to measure. One of the most commonly used measures of validity evidence in language assessment is concurrent validity. Concurrent validity is measured by exploring the correlations between scores obtained on the target test and scores obtained on another test measuring the same skill and administered to the same participants within the same time period.

5) Diagnostic accuracy: it refers to the ability of the test to give accurate diagnoses. It is generally measured via the analysis of the sensitivity/specificity and the predictive accuracy of the test. In language assessment, sensitivity refers to the percentage of children with a Language Disorder (LD) correctly identified as having a LD (true positive rate) while specificity refers to the percentage of children with typical development correctly identified as not having a LD (true negative rate). Predictive accuracy measures the predictive ability of a test to correctly classify children with and without LD. It can be explored via the Receiver Operating Characteristic (ROC) curve analysis and is measured by the Area Under the Curve (*AUC*). The value of the AUC denotes the predictive accuracy of a test as follows: AUC of 1= perfect test; AUC between 0.90 and 1= excellent accuracy; AUC between 0.80 and 0.90= good accuracy; AUC between 0.70 and 0.80= fair accuracy; AUC below 0.70= poor accuracy (0.60–0.70) or worthless test (below 60) (see Swets, Dawes & Monahan, 2000).

To summarize, a language test must take into account the linguistic, cultural and socioeconomic characteristics of the participant being assessed. It should also be standardized and normed and have psychometric features such as satisfactory test reliability, validity evidence and diagnostic accuracy. In Lebanon, there is a total absence of such language tests. Speech and Language Pathologists (SLPs) often base language assessment on qualitative linguistic evaluation, spontaneous language analysis, semi-directed or directed non-normed tasks and/or translations of tests standardized and normed on foreign, monolingual populations (e.g., Clinical Evaluation of Language

Fundamentals, CELF-4, Semel, Wiig & Secord, 2003; ELO, Khomsi, 2001; a.o.). Although most of these methods should also be part of any language assessment, sole reliance on qualitative information can lead to misdiagnosis (see, Alkhamra & Al-Jazi, 2015). To address this lack of valid assessment tools, we developed the ELO-L (*Evaluation du Langage Oral chez l'enfant Libanais* 'Assessment of oral language in the Lebanese child')[2] battery. The ELO-L is the first standardized, norm-referenced speech and language screening tool in Lebanon and one of few language assessment tools in Arabic, all dialects considered (see Shaalan, 2014 for examples of Arabic tests). The ELO-L is an adaptation of the ELO (Evaluation du Langage Oral, Khomsi, 2001), [3] a French language assessment test that is commonly used by SLPs in France. It is composed of five subtests assessing the following language areas: receptive and expressive vocabulary, sentence comprehension and production and expressive phonology. There are two versions of the ELO-L: a short version for 3- to 6-year-old children and a long version for 6- to 8-year-old children. The administration time varies between approximately 30 and 45 minutes. The ELO-L is normed on 1,718 Lebanese children. The adaptation process, the subtests, the scoring method, the normative sample and the validation process are presented below.

The Adaptation Process

As mentioned above, the ELO-L is an adaptation of the ELO (Khomsi, 2001). By definition, an adaptation differs from a translation as it involves more substantial changes (e.g., item modification or substitution) in order to be linguistically or culturally relevant to the target population (Stansfield, 2003). The ELO was selected because it is a test particularly appreciated by Lebanese SLPs, who use it as a support for their qualitative assessments. The ELO-L is based on the same theoretical background and empirical approach as the ELO (see Khomsi, 2001). However, some adjustments had to be made to take into account linguistic, educational and socioeconomic specificities of Lebanon as well as the lack of scientific data on language development in Lebanon.

---

2. Zebib, Henry, Khomsi, Messarra & Kouba-Hreich (2017). *Evaluation du Langage Oral chez l'enfant Libanais (ELO-L)*. Baabda: Liban Test Editions (LTE).

3. Liban Tests Editions (LTE), detains the ELO adaptation license given by ECPA - PEARSON, France.

These adjustments mainly concerned the task items, the normative sample and the particular linguistic aspects targeted by the subtests.

Beginning with the tasks items, some items that happened to be relevant for Lebanon were translated while others were replaced with more linguistically and culturally appropriate ones. Concerning the vocabulary subtests, some lexical items in the ELO were unlikely to be part of the vocabulary stock of Lebanese children for cultural reasons (e.g., *secateurs* 'clippers') and were therefore replaced with more appropriate items in the ELO-L. The item selection process in the ELO-L differed from the procedure used in the ELO. In the latter test, lexical items in the receptive and expressive vocabulary tests were selected using a database of words containing precise information about their mean Age of Acquisition (AoA). Unfortunately, following the same procedure in the ELO-L was not possible because of the lack of studies on language development in Lebanese children and waiting for such results was in contradiction with the urgent need for norm-referenced language tests. For this reason, AoA of the words in the receptive and expressive vocabulary tests were estimated by the group of SLPs and psycholinguists who collaborated on the project, all of them professionals familiar with language in Lebanese children. It is noteworthy that this method is commonly used to establish AoA of lexical items (cf. Khomsi, & Khomsi, 2007; Morrison, Chappell & Ellis, 1997; Shaalan, 2014). The selection of the lexical item was based on the picture collection developed for tests by Khomsi and colleagues (ELO, Khomsi, 2001; BILO, Khomsi & Khomsi, 2007, etc.), as these pictures were the ones used in the subtests of ELO-L. The group of SLPs and psycholinguists reviewed the lexical items that correspond to the different pictures and estimated their AoA. Then, the items in each of the *Receptive Vocabulary* and *Expressive Vocabulary* tests were selected in a way to ensure AoA variability and to avoid floor or ceiling effects in the appropriate age slots (3 to 6 and/or 6 to 8 years).

Turning to the *Expressive Phonology* task, items were selected following the procedure used in ELO (Khomsi, 2001). Three variables were manipulated throughout the task: 1) word length (one to four syllables), 2) syllable structure, which can be with or without consonant clusters and 3) lexical familiarity. In fact, complexity of syllable

structures, item length (supported by phonological working memory) and lexical knowledge seem to affect performance in phonological production tasks (see, dos Santos & Ferré, 2018). To select the words, the SLPs and psycholinguists working on the project proposed lists of familiar (early AoA) and non-familiar words (late AoA and therefore probably processed as non-words in the target age range), with one, two, three and four syllables and with complex *vs* simple syllable structure (with or without consonant clusters). Then, they worked together to review each word in order to keep the items that were unanimously judged to be relevant.

Concerning the *Sentence Comprehension* and the *Sentence Production* tasks, selection of items was constrained by the pictures available, as in the vocabulary tasks. Most items were translated from the French version (ELO, Khomsi, 2001), taking into account what is known about morphosyntax in comprehension and in production in studies on other languages and on other Arabic dialects (Abdalla & Crago, 2008; Al-Akeel, 1998; Mustafawi & Mahfoudhi, 2005; Ravid & Farah, 1999; Shaalan, 2010; a.o.) in order to estimate the AoA of the different items and to provide variability in the morphosyntactic complexity of the items. Once the items were translated, the psycholinguists and SLPs involved in the project discussed their relevance based on their knowledge of Lebanese Arabic and on their clinical experience. Most items were approved unanimously. A few items did not reach consensus and were therefore replaced by other items assessing the same morphosyntactic feature. Finally, the tests were piloted in children in the target age range to confirm their feasibility and their relevance for use with Lebanese children.

Turning to the normative sample, unlike the ELO, the normative sample of the ELO-L was not selected randomly. Several sampling criteria were considered in order to take into account multilingualism in Lebanon. We selected several variables that have been suggested to have a quantitative and/or qualitative effect on exposure to different languages. Thus, it was important to have a representative sample of the Lebanese population, linguistically, culturally and socioeconomically. The sampling variables we controlled for were the following: type of school (public and private fee-paying schools), language of instruction (English or French), and geographic region (Beirut and its suburbs, northern Lebanon and southern Lebanon). Geographical region was taken as a

sampling variable as the geographical distribution of the Lebanese population is not homogenous in terms of socioeconomic status, predominance of type of school (Verdeil et al., 2007) and second language preference (see, CIEP, Fiche pays Liban). These sampling variables were selected because it was not possible, for practical reasons, to administer a parental questionnaire to get precise information about the multilingualism and the socioeconomic status of each child.

Another important difference between the ELO-L and the ELO is the fact that correct lexical responses given in a language other than Lebanese Arabic are accepted as valid responses in the expressive vocabulary task in the ELO-L. Indeed, the purpose of this test is to detect a general lexical deficit (in other words, a conceptual vocabulary score, see Junker & Stockman, 2002) and not to specifically assess lexical production in Lebanese. Thus, in this test, correct answers are accepted regardless of the language used, as recommended by Zablit & Trudeau (2008) in their study of young Lebanese children. According to some authors (Holmström, Salameh, Nettelbladt & Dahlgren-Sandberg, 2015), this method reduces the over-diagnoses of language disorders (identification of language impairment in a child who in fact has no impairment), due to underestimation of lexical knowledge, in bilingual children. It is noteworthy that the responses given by the children in the normative sample supported our methodological choice, especially in younger children. For example, when children were shown a picture of the sun, 78.15% of the correct responses were given in Arabic (شمس), 16.41% in French (*soleil*) and 5.44% in English (*sun*) with more answers given in the non-Arabic language in the younger age groups (see table 1). The percentages of answers given in the different languages vary from one item to another.

Table 1 Percentages of responses given in Arabic, French and English for item 8 (Nouns) of the Expressive Vocabulary task

Age Group	3;0-3;5	3;6-3;11	4;0-4;5	4;6-4;11	5;0-5;5	5;6-5;11	6;0-6;11	7;0-7;11
% Arabic	40.7	58	74.9	77.5	89.6	90.3	96.2	98
% French	34.6	32.7	22.4	19.7	8.3	7.8	3.8	2
% English	24.7	9.3	2.7	2.9	2.1	1.8	0	0

The Subtests

The ELO-L is composed of five subtests, each of which has separate norms. All of the pictures used in the test are black and white drawings taken from the picture collection developed for tests by Khomsi and colleagues (ELO, Khomsi, 2001; BILO, Khomsi & Khomsi, 2007, etc.). The pictures for each subtest are presented in a separate booklet. The expressive phonology subtest has no pictures. The subtests are presented below following the order of administration when the entire test battery is used.

The *Receptive Vocabulary* subtest is composed of one trial item and 34 test items, in both the short and the long versions. In this task, the child is asked to point to one picture, out of four pictures presented on the same page, which corresponds to a word given by the test administrator. Among the four pictures, one corresponds to the word given and is therefore the target picture, one is a phonological distractor (forming a minimal pair with the word given), one is a semantic distractor and the last one has no link with the target word. Figure 1 provides an example: the target picture is a tent (**خيمة**; /χajme/) the phonological distractor is a cloud (/ɣajme/), the semantic distractor is a house, and the fourth picture, a clock, is not linked to the word 'tent'. This procedure makes it possible to assess the precision of lexical representations: phonological distractors give insight into the quality of the child's phonological representations and into his/her ability to discriminate words that are phonologically similar, while semantic distractors assess the precision of semantic representations (see Khomsi, 2001). In other words, this task assesses receptive vocabulary as well as the robustness of semantic and phonological representations. As mentioned earlier, the items in the test vary according to the estimated Age of Acquisition (AoA) of the words.

The Receptive Vocabulary test provides three scores: The first one corresponds to the total number of correct responses (maximum score of 34), the second one corresponds to the total number of designations of phonological distractors and the third one corresponds to the total number of designations of semantic distractors.

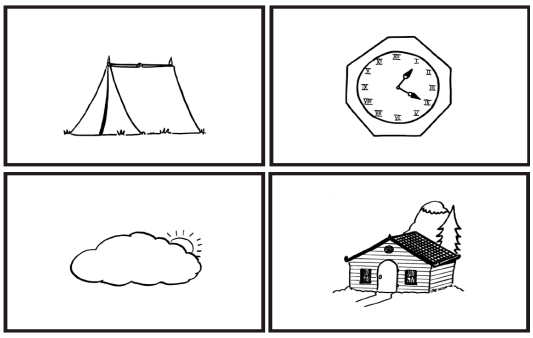


Figure 1. Example of an item from the Receptive Vocabulary test

The *Sentence Comprehension* test assesses listening comprehension of sentences. It is composed of one trial item and 15 test items in the short version and 27 items in the long version. The additional items in the long version add complexity to the task to avoid ceiling effects in older children (those aged 6 to 8). In this task, the child is asked to point to a picture, choosing from among four pictures the one that corresponds to the sentence given by the test administrator. The items vary in terms of morphosyntactic complexity and inferential load. Concerning morphosyntactic complexity, the sentences vary from simple sentences (V-O) to more complex sentences that require processing of morphological features (number/gender of nouns and pronouns, verb inflection, etc.) and syntactic computational complexity, such as syntactic movement or clausal subordination (e.g., subject and object relatives). Turning to inferential load, some sentences require an inference to be made in order to be understood, while others are directly related to the pictures. Inference skills have, in fact, been shown to predict listening comprehension performance in young children (Lepola et al., 2012) and performance in reading comprehension in older children (Cain, Oakhill & Bryant, 2004). Figure 2 provides an example of a non-inferential item (left), where the picture is directly related to the sentence **عم تجلي** (Prog. fem-wash (the dishes) 'She is washing the dishes'), and an example of an inferential item (right), where selection of the correct picture for the sentence **البنت وقعت؟** (The-girl fell-fem? 'Did the girl fall?') requires inference. The distractors are designed to control the specific morphosyntactic and semantic variables involved in the processing of each sentence. In sum, this test assesses comprehension of sentences by tapping into different component skills of comprehension (see Bishop, 2014): some sentences require basic cognitive, semantic and morphosyntactic processing

skills involved in language comprehension while others also involve inference.

This test provides one score that corresponds to the total number of correct answers (maximum score of 15 for the short version and of 27 for the long version).

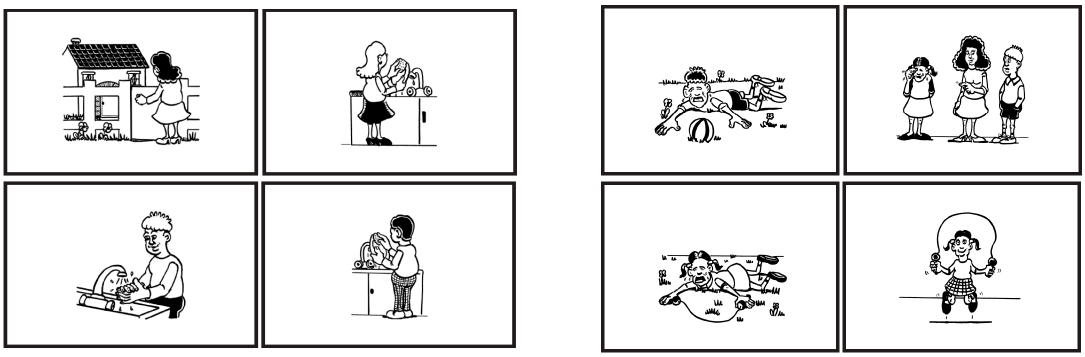
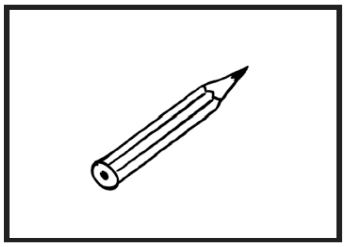


Figure 2. Examples of a non-inferential item (left) and of an inferential item (right) of the Sentence Comprehension test

The *Expressive Vocabulary* test is composed of 2 trial items (one noun and one verb) and of 54 items (27 nouns and 27 verbs) in the short version and 70 items (35 nouns and 35 verbs) in the long version. It assesses lexical knowledge and retrieval of nouns and verbs, as knowledge of verbs *vs.* nouns was shown to be different at a developmental level and because children with language disorders show differences in the processing of nouns *vs.* verbs (see Haman, Luniewska & Pomiechowska, 2015, for a review). In this task, the child is asked to name pictures by answering the question *What is this?*, for the nouns, and *What is he/she doing?*, for the verbs (see Figure 3 for an example). As in the *Receptive Vocabulary* test**,** the items in this task vary according to the estimated AoA of the words.

This test provides three scores: The first one corresponds to the total number of correct responses on the entire test (maximum score of 54 for the short version and of 70 for the long version), the second one corresponds to the total number of correctly named nouns and the third one corresponds to the total number of correctly named verbs. As mentioned before, this test was designed to assess conceptual vocabulary and, therefore, appropriate answers given in languages other than Lebanese are counted as correct. We note that a list of accepted answers is given for each item.

What is this? (Nouns)



What is he doing? (Verbs)

Figure 3. Example of a noun item (left) and a verb item (right) from the Expressive Vocabulary test

The *Sentence Production* test assesses morphosyntactic skills in production. It is composed of 23 items in the short version and 37 items in the long version. As is the case for the Sentence comprehension task, the additional items in the long version add complexity to the task to avoid ceiling effects in older children. This test is a sentence completion task: the child sees a picture and hears a related sentence, and then (s)he sees another picture and hears the beginning of a sentence. The child then has to complete this second sentence while taking into account the morphosyntactic transformations induced by the picture. So, for the item given in Figure 4, the child sees Picture 1 (right) and hears "هون الصّبي نايم" (here the-boy asleep-masc 'Here, the boy is asleep') and then he/she sees Picture 2 (left) and hears "...و هون البنت" ('And here, the girl…'). In order to complete the sentence, the child has to transform the masculine form of the first sentence into a feminine form. Test items vary in morphosyntactic complexity.

This test provides one score corresponding to the sum of the sub-scores obtained on the different items. In the short version, one item has a maximum score of 2 points (corresponding to 2 morphosyntactic features targeted by the sentence). In the long version, 4 items have a maximum score of 2 points and 2 have a maximum score of 3 points (maximum score of 23 for the short version and of 37 for the long version). Precise scoring instructions are given for each item.

Picture 2 Picture 1

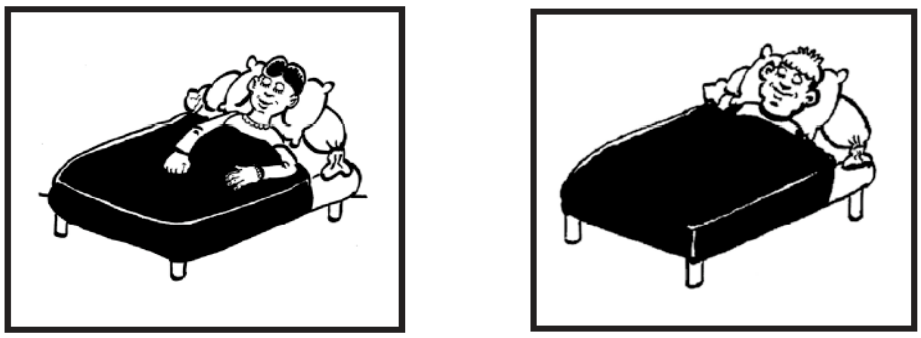


Figure 4. Example of an item of the Sentence Production test

The *Expressive Phonology* task is composed of 28 items in both the short and the long versions. In this test, the child is simply asked to repeat the words produced by the test administrator as accurately as possible. As mentioned above, the items vary according to syllable length from 1 to 4 (e.g., /kaff/ - كفّ – 'palm' is monosyllabic, /sˤɑbe/ - صبي 'boy' has two syllables, /ʃamsijje/ - شمسيّة 'umbrella' has three syllables, /banadu:ra/ - بندورة 'tomato' has four syllables), phonological complexity, mainly syllable structure, which can be with or without consonant clusters (ex: /sfenʒe/ -سفنجة 'sponge' *vs* /tˤɑ:be/ طابة – 'ball'), and familiarity, with some (unfamiliar) words most likely being processed by young children as nonwords, while others were chosen for their early AoA (ex: /mestˤawsˤɑf/ - مستوصف 'dispensary' *vs* /tˤɑ:be/ - طابة 'ball'). In sum, this test assesses expressive phonological skills through a quick screening of phonological performance in contexts that vary in the degree to which phonological working memory (length of the items) and lexical knowledge (AoA of the items) are involved. Articulatory deficits and regional accents are not penalized (since the target domain is phonology).

This test provides one score that corresponds to the total number of correctly repeated words (maximum of 28).

The Scoring Method

Raw scores on the five subtests are obtained as mentioned in the previous section. Then, they are transposed into percentiles, which are juxtaposed to provide an individual profile for each child. This allows the examiner to observe, in addition to the specific performance of the child in each task, the homogeneity/heterogeneity of his/her

linguistic profile. Between the ages of 3;0 and 5;11, norms are provided for each six-month interval to account for the rapid evolution of language during this developmental period. Between the ages of 6;0 and 7;11, norms are provided for one-year age spans (6;0-6;11 and 7;0-7;11). Profiles are given in percentiles because some distributions did not follow the normal distribution. However, age means and standard deviations are provided to allow clinicians to calculate z-scores, when appropriate. In addition, since, after controlling for age, other variables had a significant effect on performance (on at least one task and for at least one age group), norms are also provided by grade level (from kindergarten until second grade), geographical region (Beirut, southern Lebanon and northern Lebanon), school type (private/public), school language (French or English) and gender (girls/boys).

The Normative Sample and Procedure

The normative sample is composed of 1,718 children recruited from public and private schools located in three different geographical regions in Lebanon: Beirut and its suburbs (35.4%), the south (35.3%) and the north (29.3%) of Lebanon. Children with language deficits were not excluded, as recommended by some authors (e.g., McCauley & Swisher, 1984). Indeed, the scores of these children are part of a continuum and it is difficult to set a cut-off as an exclusionary criterion. Thus, the normative sample covers the whole continuum. Of the 1718 children, 1329 are aged between 3;0, and 5;11 years and passed the short version of the battery and 389 are aged between 6;0 and 7;11 years and passed the long version. As mentioned earlier, the children aged 3;0 to 5;11 were divided into six age groups of six months each. The older population is divided into two age groups, 6-year-olds and 7-year-olds. The minimum number of participants in each age group was 100, as recommended (Salvia & Ysseldyke, 1995). The percentage of boys (53.3%) in the whole population is slightly higher than the percentage of girls (46.7%). The number of children, the mean age and the percentage of girls in each age group are presented in Table 2. Children were recruited in public (49.7%) and fee-paying private schools (50.3%). 60.2% of the children were recruited from schools which have French and Arabic as languages of instruction and 39.8% from schools using English and Arabic. The distribution of children in French versus English schools is therefore not balanced.

However, as mentioned earlier, the percentage of children in French schools is higher in Lebanon, which is in accordance with our data. It should be noted that all of the children were assessed individually in a quiet room within their school. Third- and fourth-year SLP students trained in the administration of the battery conducted the assessments.

Table 2 Number of children, mean age and percentage of girls in each age group

Age group (range)	Nb. of Participants	Mean Age (*SD*)	Gender: % of girls
3;0-3;5	100	3.30 (0.14)	54.0
3;6-3;11	184	3.76 (0.14)	45.1
4;0-4;5	262	4.26 (0.15)	42.0
4;6-4;11	258	4.72 (0.14)	48.4
5;0-5;5	308	5.24 (0.15)	45.1
5;6-5;11	217	5.72 (0.14)	44.7
6;0-6;11	186	6.54 (0.25)	48.9
7;0-7;11	203	7.43 (0.27)	51.2

The Validation Process

In this section, psychometric characteristics of the ELO-L are presented. We note that we prefer to use the word *validation* (see Dickes, Tournois, Flieller & Kop, 1994) instead of *validity* because the data concerning the validity of a test are generally limited when it is first published and the search for its validation is an ongoing process that continues after its publication (see Chartier & Loarer, 2008). The developmental sensitivity, the reliability and the validity evidence of the ELO-L are presented below.

Developmental sensitivity of the 5 subtests.

Results on the *Receptive Vocabulary* test showed a linear increase in performance between the youngest and the oldest age groups (see Table 3). A one-way ANOVA with Bonferroni correction revealed a significant effect of age group ($F(7, 1710)$=266.64; p=.000). Pairwise comparisons showed significant differences between all adjacent age groups (p <.01), except between the children aged 3;6-3;11 and the ones aged 4;0-4;5 (p

=.284). However, the difference in performance between the children aged 3; 6-3; 11 and 4; 6-4; 11 was significant (p =.000).[4]

Table 3 Means and Standard Deviations on the Receptive Vocabulary Task as a Function of Age Group

	3;0-3;5	3;6-3;11	4;0-4;5	4;6-4;11	5;0-5;5	5;6-5;11	6;0-6;11	7;0-7;11
M	13.65	15.35	16.29	18.54	19.75	21.41	25.12	26.68
(SD)	*(3.73)*	*(3.55)*	*(3.71)*	*(3.73)*	*(3.81)*	*(3.78)*	*(4.02)*	*(3.80)*

Concerning the *Sentence Comprehension* task, results also revealed a linear increase in performance in both the short and the long versions (see Table 4). A one-way ANOVA with Bonferroni correction applied on the short version showed a significant effect of age Post-hoc pairwise comparison did not show significant differences in performance between the first (3;0-3;5) and the second (3;6-3;11) (p =.1) or the third (4;0-4;5) (p =.09) age groups, between the second (3;6-3;11) and the third (4;0-4;5) age groups (p =.1) or between the third and the fourth age groups (p =.056). However, the difference between the second and the fourth age groups was statistically significant (p =.000), as well as the difference between the third and the fifth (5;0-5;5) age groups (p =.000). The differences in performance on the short version between the three older adjacent age groups were also significant (p <.01). Concerning the long version used with the children aged 6 to 8 years, a significant effect of age group was also found ($F(1, 387)$= 14.90 ; p =.000).

Table 4 Means and Standard Deviations on the Sentence Comprehension Task as a Function of Age Group
group ($F(5, 1323)$=37.25; p =.000).

	Short version						Long version	
	3;0-3;5	3;6-3;11	4;0-4;5	4;6-4;11	5;0-5;5	5;6-5;11	6;0-6;11	7;0-7;11
M	4.77	5,13	5.44	5.97	6.58	7.23	17.73	19.23
(SD)	*(2.0)*	*(2.03)*	*(1.96)*	*(2.05)*	*(2.19)*	*(2.18)*	*(4.01)*	*(3.68)*

[4]. The results also showed a significant effect of age on the scores that correspond to the total number of designations of phonological ($F(7,1710)$=78.03; p=.000) and semantic ($F(7, 1710)$= 95.59; p=.000) distractors, with a linear decrease by age group.

Results on the *Expressive Vocabulary* task showed a linear increase in performance as a function of age group in both the short and the long versions (see Table 5). A one-way ANOVA with Bonferroni correction revealed a significant effect of age group on children's performance on the short version (*F(5, 1323)*=147.20; *p*=.000). Post-hoc pairwise comparisons showed significant differences between all adjacent age groups (p <.001).[5] A significant effect of age group was also found on performance in the long version (*F(1, 387)*=20.00; *p*=.000)

Table 5 Means and Standard Deviations on the Expressive Vocabulary Task as a Function of Age Group

	Short version						Long version	
	3;0-3;5	3;6-3;11	4;0-4;5	4;6-4;11	5;0-5;5	5;6-5;11	6;0-6;11	7;0-7;11
M	21.10	24.65	28.24	32.84	35.39	38.02	52.20	55.40
(SD)	*(6.42)*	*(7.34)*	*(7.84)*	*(6.93)*	*(7.07)*	*(6.10)*	*(7.42)*	*(6.71)*

Similarly, results on the *Sentence Production* task revealed a linear increase in performance as a function of age group in both the short and the long versions (see table 6). A one-way ANOVA with Bonferroni correction revealed a significant effect of age group on performance on the short version (*F(5, 1323)*=123.01; *p*=.000). Post-hoc pairwise comparisons revealed significant differences between all adjacent age groups (*p*<.05). A significant effect of age group was also found for the long version (*F(1, 387)*=12.63; *p*=.000)

Table 6 Means and Standard Deviations on the Sentence Production Task as a Function of Age Group

	Short version						Long version	
	3;0-3;5	3;6-3;11	4;0-4;5	4;6-4;11	5;0-5;5	5;6-5;11	6;0-6;11	7;0-7;11
M	5.64	6.83	7.84	9.71	10.56	12.21	21.45	22.97
(SD)	*(2.55)*	*(2.95)*	*(2.86)*	*(3.04)*	*(3.14)*	*(2.90)*	*(4.52)*	*(3.91)*

[5]. Results on the subscores of the expressive vocabulary task also showed a significant effect of age group (Nouns: Short version/27: *F(5, 1323)*=138.99; *p*=.000; Long version/35: *F(1, 387)*= 14.22; *p*=.000); Verbs: Short version/27: *F(5, 1323)*=104.50; *p*=.000; Long version/35: *F(1, 387)*=17.88; *p*=.000).

Finally, results on the *Expressive phonology* task revealed a linear increase in performance as a function of age group (see Table 7). A one-way ANOVA with Bonferroni correction showed a significant effect of age group on performance (*F(7, 1710)*=106.29; *p*=.000). Post-hoc pairwise comparisons showed significant differences between the second (3;6-3;11) and the third (4;0-4;5) age groups (*p*=.000), between the fifth (5;0-5;5) and the sixth (5;6-5;11) age groups (*p*=.001) and between the sixth and the seventh age groups (*p*=.000) but not between the first (3;0-3;5) and the second (*p* =.124), the third and the fourth (4;6-4;11) (*p* =.058), the fourth and the fifth (*p* =1.0) and the seventh (6;0-6;11) and the eighth (7;0-7;11) (*p*= 1.0) age groups. Although many differences between adjacent age groups were not significant, the differences between all non-adjacent groups were significant (*p* <.05). We note that the two oldest groups have mean scores close to the maximal score and relatively low standard deviations, which can be interpreted as a ceiling effect of the task in older populations.

Table 7 Means and Standard Deviations on the Phonology Task as a Function of Age Group

	3;0-3;5	3;6-3;11	4;0-4;5	4;6-4;11	5;0-5;5	5;6-5;11	6;0-6;11	7;0-7;11
M	16.34	17.9	21.02	22.28	22.42	24.19	26.42	*27.24*
(SD)	*(5.64)*	*(6.17)*	*(4.60)*	*(4.30)*	*(5.86)*	*(4.80)*	*(2.14)*	*(1.42)*

In sum, results on the five subtests of the ELO-L showed a linear increase in performance as a function of age groups. Inferential statistics revealed a good developmental sensitivity for *Receptive Vocabulary, Expressive Vocabulary* and *Sentence in Production* with significant differences between (almost) all adjacent age groups. The difference in performance between adjacent age groups on the *Sentence Comprehension* and the *Phonology* tasks was not always significant. However, differences between non-adjacent groups were generally significant.

Reliability of the five subtests.

Reliability of the subtests of the ELO-L was measured via test-retest reliability, internal reliability and inter-rater reliability. Test-retest reliability of the subtests was measured in a group of 20 subjects who were tested twice with a time interval of one month. These children are spread across the different age groups of the normative sample. Pearson

correlation coefficients revealed moderate test-retest reliability for *Receptive Vocabulary* (r = .77; p <.001) and good reliability for *Sentence Comprehension* (r = .89; p <.001), *Expressive Vocabulary* (r = .81, p <.0001), *Sentence Production* (r = .94, p <.001) and *Expressive Phonology* (r = .92; p <.001).

Internal reliability was measured via Cronbach's alpha coefficient applied on the scores of the normative sample. The results revealed satisfactory internal reliability for the *Receptive Vocabulary* task (α =.79); low reliability for the *Sentence Comprehension* task (α= .36 for the short version and α =.60 for the long version); satisfactory reliability for the *Expressive Vocabulary* task (α=.89 for the short version and α=.88 for the long version); acceptable reliability for the short version of *Sentence Production* (α=.75) and low to borderline reliability for the long version (α=.67); and satisfactory reliability for the *Expressive Phonology* task (α=.88).

Inter-rater reliability was measured only for *Expressive Phonology*, as it is the task that is potentially the most rater-dependent: In the receptive tasks, the test administrator has only to report if the child pointed to the target picture and, in *Expressive Vocabulary* and *Sentence Production*, the list of the accepted answers and the scoring instructions are specified for each item, which gives the rater little leeway. Indeed, inter-rater reliability has been shown to be almost perfect in these conditions (see Chartier & Loarer, 2008). Inter-rater reliability of *Expressive Phonology* was measured via a double rating, by two different SLPs, of the productions of 14 children with Developmental Language Disorders (M_{age}=8;1; SD= 15;4). Results based on a two-way random interclass correlation revealed excellent inter-rater agreement with an average ICC of .996 with a 95% confidence interval from .989 to .999 (F(1, 13)= 277.824, p<.001).

Validity evidence.

As the ELO-L is the first language assessment tool that has been standardized and normed in Lebanon, it was not possible to measure its concurrent validity by referring to other standardized tests. However, correlations between four ELO-L subtests (Receptive Vocabulary, Expressive Vocabulary, Sentence Production and Expressive Phonology)

and four experimental tasks – Lebanese LITMUS tests[6] - were explored (see, Zebib, Prévost, Tuller, & Henry (Eds.), *in press*). The Lebanese LITMUS tasks (LITMUS-LB) consist of a receptive vocabulary task, an expressive vocabulary task, a sentence repetition task assessing morphosyntax in production and a nonword repetition task assessing expressive phonology. These tests were administered to 42 children, aged between 5;7 and 7;10, including 32 typically developing children and 10 children diagnosed by SLPs as having a Developmental Language Disorder (with predominant phonological and morphosyntactic deficits) [7] . The results revealed satisfactory correlations (see Chartier and Loarer, 2008: 97) between each of the ELO-L subtests and the LITMUS task constructed to measure the same language ability. Thus, Spearman correlation coefficients were significant between *Sentence Production* of the ELO-L and the LITMUS-LB Sentence Repetition test (r_s = .56; p <.0001); between *Expressive Phonology* of the ELO-L and the LITMUS-LB Non-Word Repetition test (r_s = .64; p<.0001); between *Expressive Vocabulary* of the ELO-L and the LITMUS-LB Expressive Vocabulary test (r_s = .50; p <.01); and between *Receptive Vocabulary* of the ELO-L and the Lebanese LITMUS-LB Receptive Vocabulary test (r_s = .64; p <.0001).

Diagnostic accuracy.

Diagnostic accuracy of the ELO-L was measured using the data of the same population of 42 children described above (32 children with typical development and 10 diagnosed with DLD). To calculate Sensitivity and Specificity, the diagnostic criterion proposed by

6. LITMUS tests (Language Impairment Testing in Multilingual Settings) were developed as part of COST action IS0804 Language Impairment in a Multilingual Society: Linguistic Patterns and the Road to Assessment (see, Armon-Lotem, de Jong & Meir, 2015).

7. These data were collected as part of a project funded by the French Ministry of Foreign and European Affairs and the French Ministry of Higher Education and Research (Projet Cèdre) on the identification of Specific Language Impairment in multilingual contexts (*Dépistage du trouble spécifique du langage dans des contextes plurilingues*) (see, Zebib et al. (Eds.), *in press*).

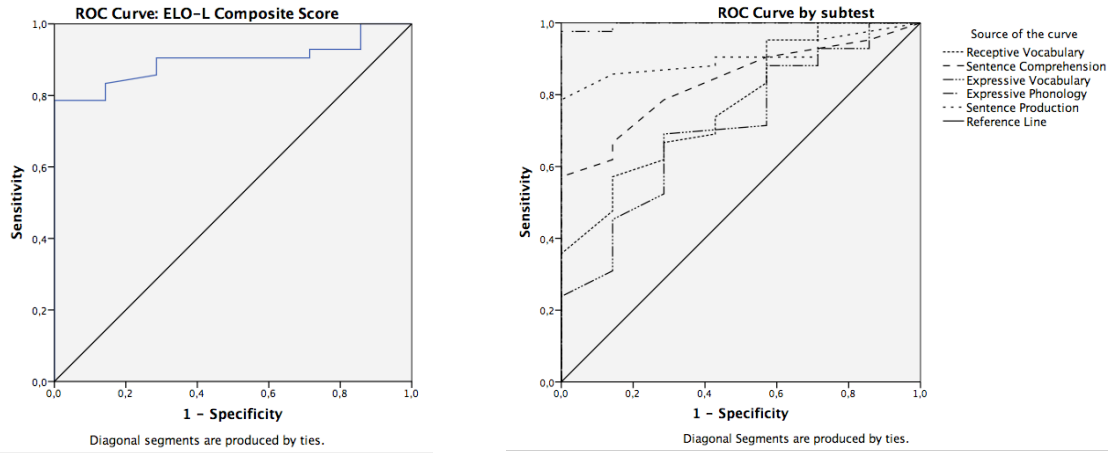


Figure 5. ROC curves applied on the ELO-L composite score and on each of the subtests.

Tomblin et al. (1996) was applied. A child was considered as having a Language Disorder if (s)he scored below norms on at least two language areas, as measured by two ELO-L subtests. Results revealed a perfect sensitivity of 100% and an excellent specificity of 93.75%. The predictive accuracy applied on a composite score that includes the five subtests revealed excellent accuracy (*AUC*=.90; *p*=.001). Moreover, predictive accuracy was good for the *Sentence Comprehension* test (*AUC*=.84; *p*=.005), excellent for the *Sentence Production* test (*AUC*=.90; *p*=.001), almost perfect for the *Expressive Phonology* test (*AUC* = .997; *p* =.001), fair for the *Receptive Vocabulary* test (*AUC*=.78; *p* = .021) and non-significant for the *Expressive Vocabulary* test (*AUC* = .71; *p* =.076) (see Figure 5). We note that children with DLD in this sample have mainly phonological and morphosyntactic deficits, which is in line with the better predictive accuracy obtained in the tests targeting phonology and morphosyntax (*Expressive Phonology, Sentence Comprehension* and *Sentence Production*).

Discussion and Conclusion

The ELO-L is the first norm-referenced speech and language screening assessment test in Lebanon. It is composed of five subtests assessing Receptive Vocabulary, Sentence Comprehension, Expressive Vocabulary, Sentence Production and Expressive Phonology. It is normed on 1,718 Lebanese children aged between 3 and 8 years and divided into 8 age groups with a minimum of 100 participants in each group. Two versions of the ELO-L exist: a short version for children aged 3;0 to 5;11 years (divided

into 6 age groups of 6 months each) and a long version for children aged 6;0 to 7;11 years (divided into 2 age groups of 1 year each). In order to have a representative sample of Lebanese children, and thus include variety related to SES and related cultural variables, participants were recruited from private fee-paying and public schools, having English or French as language of instruction besides Arabic and located in Beirut and its suburbs, southern Lebanon and northern Lebanon.

Besides striving to take into account these sociocultural variables inherent in the contemporary Lebanese context, the adaptation process also included a systematic review of the content of the subtest items. These items were selected by a group of psycholinguists and SLPs who, because of the lack of scientific studies on language development and language pathology in Lebanon, had to rely on their clinical experience and on the results obtained in studies on other Arabic dialects or other languages. Although this procedure is acknowledged, especially for the selection of lexical items (Khomsi, & Khomsi, 2007; Morrison, Chappell & Ellis, 1997; Shaalan, 2014), it is probably not the ideal method for ensuring the best possible content validity. Indeed, it would have been more appropriate to start by exploring language development in Lebanese children in order to have more scientific data for item selection in the ELO-L. However, the urgent need for norm-referenced tests in Lebanon pushed us to forego the latter methodological choice. Similarly, the methodology used to take into account variability in bilingualism and socio-economic status in the Lebanese population (the sampling criteria of the normative sample) could have been improved through use of a parental questionnaire to control these variables more precisely. Interviewing the parents of 1718 children spread all over the Lebanese territory was simply not feasible. However, multilingualism was directly integrated into the task where possible. This was the case for the Expressive Vocabulary task, where appropriate answers given in a language other than Lebanese Arabic are counted as correct. Multilingualism was not directly taken into account in the other tasks for feasibility reasons. In our view, standardizing French and English tests on the Lebanese population, in addition to the ELO-L, seems to be the only way to fully address this question.

Regarding the psychometric characteristics of the ELO-L, the first results of the validation process were quite promising. We discuss the results of this process for each of the Receptive Vocabulary, Expressive Vocabulary, Sentence Comprehension, Sentence Production and Expressive Phonology tasks, in turn. Results on the *Receptive Vocabulary* task, revealed good developmental sensitivity with a linear and statistically significant increase in mean scores between almost all adjacent age groups, moderate test-retest reliability, satisfactory internal reliability and significant concurrent validity. Results on the *Expressive Vocabulary task* revealed good developmental sensitivity with a linear and statistically significant increase in mean scores between all adjacent age groups in both the short and the long versions. Moreover, it has good test-retest reliability, satisfactory internal reliability and significant concurrent validity. Results on the *Sentence Comprehension* task showed good general developmental sensitivity with a significant effect of age on the children's performance and significant pairwise differences between the oldest age groups. However, pairwise comparisons did not show significant differences between the youngest adjacent age groups (four groups aged between 3;0 and 4;6), although the mean scores increased linearly, which explains the significant difference generally found between non-adjacent age groups. Thus, the scores of the youngest children on this task should be interpreted with caution as it may lack sensitivity in these age slots. Note that we preferred not to group adjacent age groups when no significant difference in performance was found in a particular subtest as the scoring method used in ELO-L allows the examiner to look at the general profile of a child in comparison to his age group and therefore to identify eventual peaks and valleys in performance. The results on the *Sentence Comprehension* task also showed good predictive accuracy and test-retest reliability, but low internal reliability. This latter result can be explained by the nature of the task, as it was designed to assess different component skills of language comprehension (linguistic and inferential skills) and because the number of inferential items was inferior to the number of non-inferential items. This item distribution was applied because children as young as the ones in our normative sample rely mostly on non-inferential processing to interpret sentences (Khomsi, 1987; 1999; 2001). Nevertheless, these results suggest that this task could be

improved. Results on the *Sentence Production* task revealed good developmental sensitivity with significant differences between all age groups, excellent predictive accuracy, good test-retest reliability, significant concurrent validity and acceptable internal reliability for the short version but low to borderline reliability for the long version. This latter result may be due to the particularly long sentences that were added in the version for 7- o 8-year-olds. These sentences may involve different processing abilities in comparison to the sentences in the first part of the task (sentences of the short version), with greater involvement of working memory. This hypothesis should, however, be verified in a subsequent study. Finally, the *Expressive phonology* task has good general developmental sensitivity with a significant effect of age on children's performance and linear increase in mean scores between the different age groups. Although pairwise comparison did not give significant differences between all age groups, the differences between all non-adjacent age groups were significant. Note that a ceiling effect was observed in the two oldest age groups, which explains the absence of significant difference between them. This result is in accordance with the scientific literature on speech sound development (Daviault, 2011). Moreover, the *Expressive Phonology task* has good test-retest reliability, excellent inter-rater reliability, satisfactory internal reliability, significant concurrent validity and almost perfect predictive accuracy.

In sum, the ELO-L has promising psychometric characteristics in general. Moreover, the entire test showed excellent predictive accuracy with perfect sensitivity and excellent specificity, which are key measures of the relevance of psychometric tests. These diagnostic accuracy results were obtained based on a study of 42 children, 32 children whose language development was typical (as verified by educators and parental survey) and 10 children who had been diagnosed by SLPs for a DLD (also verified with the results of a parental questionnaire) subjected to thorough testing with experimental language tasks. It should be noted, however, that this result should be approached with caution, as it is possible that children currently diagnosed as having a DLD by SLPs in Lebanon appear to have severe language deficits, which naturally gives rise to the perfect sensitivity of the battery. Indeed, it is likely that less severely affected children are under-diagnosed in Lebanon, precisely because of the lack of psychometric

tests up until now. The wide availability of the ELO-L and its ensuing systematic use by SLPs may alleviate this problem.

In conclusion, the ELO-L is the first norm-referenced language test for Lebanon. It is composed of five subtests that target different language domains and has promising psychometric characteristics with a large normative sample. It was designed to take into account, as far as possible, the linguistic, socioeconomic and cultural specificities of the Lebanese population. As is the case for other language screening tests, the ELO-L should be used in addition to detailed anamnesis, clinical observation and qualitative linguistic analysis to ensure more accurate diagnosis and more appropriate therapeutic projects. It can also be used as a language screening measure in research on typical and atypical language acquisition (see de Almeida et al., 2017; dos Santos & Ferré, 2018; Hamann & Abed Ibrahim, 2017; Khoury Aouad Saliby, Dos Santos, Kouba Hreich & Messarra, 2017; Tuller et al., 2018; Zebib et al., *in press*; a.o.).

References

Abdalla, F., & Crago, M. (2008). Verb morphology deficits in Arabic-speaking children with specific language impairment. *Applied Psycholinguistics*, *29*(2), 315-340. doi:10.1017/S0142716408080156

Al-Akeel, A. (1998). The Acquisition of Arabic Language Comprehension by Saudi Children. Unpublished Ph.D. Thesis, University of Newcastle upon Tyne, UK.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders.* (5th edn) (DSM-5™). Arlington, VA: American Psychiatric Publishing.

Armon-Lotem, S., de Jong, J. & Meir, N. (2015). *Assessing multilingual children: Disentangling bilingualism from language impairment*. Multilingual Matters.

Bishop, D. (2014). *Uncommon understanding* (Classic edn). London: Psychology Press.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, *96*, 31–42.

Chartier, P. et Loarer, E. (2008) *Evaluer l'intelligence logique.* Paris : Dunod.

CIEP, Fiche pays Liban (http://www.lefildubilingue.org/book/export/html/267)

Daviault, D. (2011). *L'émergence et le développement du langage chez l'enfant.* Montréal: Chenelière Education.

de Almeida, L., Ferré, S., Morin, É., Prévost, P., dos Santos, C., Tuller, L. & Zebib, R., (2017). "Identification of Bilingual Children with Specific Language Impairment in France", Linguistic Approaches to Bilingualism, *7*(3/4), pp. 331-358.

de Jong, J., Çavus, N. & Baker, A. (2010). Language impairment in Turkish-Dutch bilingual children. In S Topbas, & M Yavas, *Communication disorders in Turkish* pp. 290- 302. Bristol: Multilingual Matters.

Dickes, P., Tournois, J., Flieller, A., Kop, J-L. (1994). *La psychométrie.* Vendôme : Editions puf.

dos Santos, C. & Ferré, S. (2018) A nonword repetition task to assess bilingual children's phonology, *Language Acquisition*, *25*(1), 58-71, doi: 10.1080/10489223.2016.1243692

Genesee, F., Paradis, J. & Crago, M. (2004). *Dual language development and disorders: A handbook on bilingualism and second language learning*. Baltimore: Brookes

Publishing Company.

Gouteyron, A., Bernadaux, J., Renar, I., Jeambrun, P., Pourtaud, D., Dupont, J.-L. & Miraux, J.-L. (2000-2001). Rapport d'information n°52 de la mission d'information sur les relations culturelles, scientifiques et techniques de la France avec le Liban, la Syrie et la Jordanie. [En ligne]. http://www.senat.fr/rap/r00-052/r00-052_mono.html#toc4.

Hamann, Cornelia & Abed Ibrahim, Lina (2017). Methods for identifying specific language impairment in bilingual populations in Germany. *Frontiers in Communication*, *2*, (16).

Haman, E., Łuniewska, M., & and Pomiechowska, B. (2015). "Designing cross-tasks (CLTs) for bilingual preschool children," S. Armon-Lotem, J. de Jong, & and N. Meir (Eds.), *Methods for assessing multilingual children: Disentangling bilingualism from Language impairment*, 194-238. Bristol: Multilingual Matters.

Holmström, K., Salameh, E.-K., Nettelbladt, U., & Dahlgren-Sandberg, A. (2015). Conceptual scoring of lexical organization in bilingual children with language impairment. *Communication Disorders Quarterly,* 1-11.

Hoyek, S. (2004). Le français dans l'enseignement scolaire et universitaire au Liban. *Cahiers de l'Association Internationale des Etudes Françaises*, *56*, 49-56.

Junker, D.A., & Stockman, I.J. (2002). Expressive vocabulary of German-English bilingual toddlers. *American Journal of Speech-Language Pathology*, *11*(4), 381-394.

Khomsi, A. (1987). Epreuve d'évaluation des strategies de comprehension en situation orale: O-52. Paris: Editions du Centre de Psychologie Appliquée (ECPA).

Khomsi, A. (1999). *Lecture de mots et compréhension – révisée* (LMC-R). Paris: Editions du Centre de Psychologie Appliquée (ECPA).

Khomsi, A. (2001). *Evaluation du langage oral* (ELO). ECPA : Paris.

Khomsi, A. & Khomsi, J. (2009). *Bilan informatisé de langage oral pour le cycle 1*. ECPA: Paris.

Khoury Aouad Saliby, C., dos Santos, C., Kouba Hreich, E. & Messarra, C. (2017). Assessing Lebanese bilingual children: The use of cross-linguistic lexical tasks in Lebanese Arabic, *Clinical Linguistics & Phonetics, 31*, 11-12, 874-892, doi:

10.1080/02699206.2017.1308554

Landers, R. (2015). Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS, *The Winnower 5*:e143518.81744. doi: 10.15200/winn.143518.81744

Lepola, J., Lynch, J., Laakkonen, E., Silvén, M. and Niemi, P. (2012), The Role of Inference Making and Other Language Skills in the Development of Narrative Listening Comprehension in 4–6-Year-Old Children. *Read Res Q, 47*, 259-282. doi:10.1002/rrq.020

Letts C., Edwards S., Sinka I., Schaefer B., Gibbons W. (2013). Socio-economic status and language acquisition: Children's performance on the new Reynell developmental language scales. *International Journal of Language Communication Disorders, 48*, 131–143. doi : 10.1111/1460-6984.12004

Makki, M. (2007). La langue française au Liban: Langue de division. Langue de consensus? *Hérodote, 126*(3), 161-167.

McCauley, R., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49*, 34-42.

Mustafawi, E. & Mahfoudhi, A. (2005). The development of binding principles in Qatari Arabic. *Al-'Arabiyya*, *38-39*, 19-44.

Paradis, J. (2010a). The interface between bilingual development and specific language impairment. *Applied Psycholinguistics, 31*, 227–252.

Paradis, J. (2010b). Bilingual children's acquisition of English verb morphology: Effects of language exposure, structure complexity, and task type. *Language Learning 60*, 651–680.

Paradis, J., Genesee, F. & Crago, M. (2011). *Dual language development and disorders : A handbook on bilingualism and second language learning* (2nd edn). London: Brookes.

Ravid, D. & Farah, R. (1999). Learning about noun plurals in early Palestinian Arabic. *First Language*, *19*, 187-206.

Salvia, J. & Ysseldyke, J.E. (1995). *Assessment (6th edn).* Boston: Houghton Mifflin.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical evaluation of language fundamentals, fourth edition (CELF-4)*. Toronto, Canada: The Psychological Corporation/A

Harcourt Assessment Company.

Shaaban, K. (2017). The ongoing rivalry between English and French in Lebanon. In Atta Gebril (Ed.), *Applied linguistics in the Middle East and North Africa: Current practices and future directions* (pp.162-182). Amsterdam: John Benjamins.

Shaalan, S. (2010). *Investigating Grammatical Complexity in Gulf Arabic-Speaking Children with Specific Language Impairment (SLI)*. Doctoral Dissertation. UCL (University College London), London, United Kingdom.

Shaalan, S. (2014). Reliability and validity of four Arabic language tests: A comparison of performance of Qatari school-aged children with and without language impairment. *Arab Journal of Applied Linguistics, 2*(1), 20-48.

Stansfield, C.W. (2003). Test translation and adaptation in public education in the USA. *Language Testing*, *20*(2), 189-207.

Swets J. A., Dawes R. & Monahan J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1, 1–26.

Tomblin J.B., Records N.L. & Zhang X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research, 39*, 1284–1294.

Tuller, L., Hamann, C. Chilla, S. , Ferré, S. , Morin, E. , Prevost, P. , dos Santos, C. , Abed Ibrahim, L. & Zebib, R. (2018), Identifying language impairment in bilingual children in France and in Germany. *International Journal of Language & Communication Disorders, 53*, 888-904. doi:10.1111/1460-6984.12397

Verdeil, E., Ghaleb, F. & Velut S. (2007). *Atlas du Liban: Territoires et société*. Liban : Editions IFPO/CNRS.

Zablit, C. & Trudeau, N. (2008). Le vocabulaire chez les jeunes enfants libanais arabophones, francophones et bilingues. *Glossa, 103*, 36-52.

Zebib, R., Prévost, P., Tuller, L. & Henry, G. (ed.) (*in press*). *Plurilinguisme et troubles spécifiques du langage au Liban*. Presses universitaires de l'Université Saint Joseph : Beyrouth.