Applied Neuropsychology: Child

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hapc20

Assessing language development in Arabic: The Arabic language: Evaluation of function (ALEF)

Natalia V. Rakhlin , Abdullah Aljughaiman & Elena L. Grigorenko

To cite this article: Natalia V. Rakhlin , Abdullah Aljughaiman & Elena L. Grigorenko (2021) Assessing language development in Arabic: The Arabic language: Evaluation of function (ALEF), Applied Neuropsychology: Child, 10:1, 37-52, DOI: 10.1080/21622965.2019.1596113

To link to this article: https://doi.org/10.1080/21622965.2019.1596113

View supplementary material

Published online: 10 May 2019.

Submit your article to this journal

Article views: 244

View related articles

View Crossmark data

Citing articles: 2 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hapc20

APPLIED NEUROPSYCHOLOGY: CHILD
2021, VOL. 10, NO. 1, 37–52
https://doi.org/10.1080/21622965.2019.1596113

Routledge
Taylor & Francis Group

Check for updates

Assessing language development in Arabic: The Arabic language: Evaluation of function (ALEF)

Natalia V. Rakhlin[a], Abdullah Aljughaiman[b], and Elena L. Grigorenko[c,d,e,f]

[a]Linguistics Program, Department of English, Wayne State University, Detroit, Michigan, USA; [b]Department of Special Education, King Faisal University, Hofuf, Kingdom of Saudi Arabia; [c]Department of Molecular and Human Genetics and Department of Pediatrics, Baylor College of Medicine, Houston, Texas, USA; [d]Department of Psychology, University of Houston, Houston, Texas, USA; [e]Child Study Center and Haskins Laboratories, Yale University, New Haven, Connecticut, USA; [f]Research Office, Moscow State University for Psychology and Education, Moscow, Russian Federation

ABSTRACT

Arabic is characterized by extensive dialectal variation, diglossia, and substantial morphological complexity. Arabic lacks comprehensive diagnostic tools that would allow for a systematic evaluation of its development, critical for the early identification of language difficulties in the spoken and written domains. To address this gap, we have developed an assessment battery called Arabic Language: Evaluation of Function (ALEF), aimed at children aged 3 to 11 years. ALEF consists of 17 subtests indexing different language domains, modalities, and associated skills and representational systems. We administered the ALEF battery to native Gulf Arabic-speaking children ($n = 467$; ages 2.5 to 10.92; 55% boys; 20 children in each 6-month age band) in Saudi Arabia in two data collection waves. Analyses examining the psychometric properties of the instrument indicated that after the removal of misfitting items, the ALEF subtests had reliability coefficients in the range from 0.78 to 0.98, and resulting subtest scores displayed a consistent profile of positive intercorrelations and age effects. Taken together, the results indicate that the ALEF battery has good psychometric properties, and can be used for the purpose of evaluating early language development in Gulf Arabic speaking children, pending further refinement of the test structure, examination of gender-related differential item functioning, and norming.

KEYWORDS

Arabic; developmental language disorder; language assessment; language development; psychometrics

Introduction

Arabic is the mother tongue of over 313 million people in the Middle East and North Africa, the fifth most spoken language in the world (Simons & Fennig, 2018). Even though there has been much descriptive work done on classical Arabic and its modernized version, Modern Standard Arabic, very few studies of spoken Arabic, and especially of child language acquisition and developmental language disorders (DLD), have been conducted. Although the typological differences in phonology, syntax, morphology, and orthography between Arabic and Indo–European languages, like English, necessitate such research, few studies have examined early child language development in Arabic-speaking children. One of the major reasons that research on DLD is currently so limited in Arabic is that there have been no standardized assessment devices suitable for such research. The absence of such instruments impedes not only research on child development in general and DLD in particular, but also educational and clinical practice, effectively preventing evidence-based identification and treatment of early language disorders.

In this article, we report on the development of a novel assessment battery named Arabic Language: Evaluation of Function (ALEF) developed to address the need for a theoretically-mindful and psychometrically sound assessment of spoken Arabic. Although the battery can be adapted for use with speakers of a variety of dialects, the data reported here were collected with children acquiring Gulf Arabic, a group of closely related varieties of vernacular Arabic spoken in several countries on the Arabian Peninsula along the Persian Gulf, from northern Kuwait to Oman (Hole, 1995). The core area of Gulf Arabic encompasses the eastern coast of Saudi Arabia, including Al-Hasa–Dammam, the area

CONTACT Elena L. Grigorenko elena.grigorenko@times.uh.edu TIMES, University of Houston, 4849 Calhoun Rd, Health 1, RM 372, Houston, TX 77204, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hapc.

Supplemental data for this article is available online at on the publisher's website.

© 2019 Taylor & Francis Group, LLC

where the present data were collected. Here we provide the theoretical context that motivated the development of ALEF, present the data from two waves of data collection aimed at establishing ALEF's preliminary psychometric properties, and gauge its clinical potential through the examination of subtest intercorrelations and their sensitivity to age effects.

Language acquisition in children with disorders of spoken and written language

Theories of language acquisition assume that, when unfolding typically, it progresses relatively uniformly in all languages (Guasti, 2002) and entails a sequence of genetically and experientially regulated developmental stages that result in the acquisition of language-specific representations (e.g., inventories of sounds, word roots and affixes, inflectional paradigms) and a grammar (i.e., a combinatorial system that forms the basis for the generative capacity of human language). Neurodevelopmental disorders may affect one, some, or all levels of linguistic representations and/or the capacity for manipulating them in real time, depending on the severity of the condition and its individual manifestations. These deficits are central to the cluster of primary disorders of language development and functioning, exemplified here by DLD and language-related learning disorders, highly prevalent childhood-onset neurodevelopmental disorders, still largely underidentified and understudied in most linguistic and cultural settings. These disorders are characterized by diverse patterns of deficits in phonological, morphological, lexical, syntactic, semantic, and pragmatic, as well as written language domains. In addition to considerable within-disorder heterogeneity, there also exists a large degree of cross-disorder commonality in linguistic traits (Levy & Schaeffer, 2011). For example, whereas it has been documented that phonological awareness is particularly impaired in reading disability (Frost et al., 2009; Kovelman et al., 2012; Ziegler & Goswami, 2005), its deficits have also been reported in DLD (Boudreau & Hedberg, 1999); conversely, morphosyntactic deficits are an area of weakness specifically associated with DLD (Armon-Lotem, 2012; Bedore & Leonard, 2001; Hansson & Nettelbladt, 1995; Rice, Tomblin, Hoffman, Richman, & Marquis, 2004), although such weaknesses have been reported in children with reading disability (Cantiani, Lorusso, Guasti, Sabisch, & Mannel, 2013; Rispens, Roeleven, & Koster, 2004); and even though pragmatic deficits are typically observed in Autism Spectrum Disorders (Ben-Yizhak et al., 2011; Reisinger, Cornish, & Fombonne, 2011; Young, Diehl, Morris, Hyman, & Bennetto, 2005) and social (pragmatic) communication disorder (Norbury, 2014), they have also been documented in DLD (Bishop, Chan, Adams, Hartley, & Weir, 2000).

This cross-disorder commonality of symptoms, as well as the transformation of disorder phenotypes throughout development, necessitates a dimensional and developmental approach, rather than treating each disorder as categorically distinct and invariant across individuals and across the life span. Correspondingly, to study DLD, multivariate phenotypes should be derived from assessments that tap into distinct but related aspects of language functioning, encompassing sublexical, lexical, grammatical, and pragmatic knowledge, as well as the cognitive systems that support the acquisition of this knowledge and its online application (e.g., memory systems, executive functions, and sentence processing algorithms). A language assessment that aims at being comprehensive should sample from multiple types of linguistic representations (e.g., phonological, lexical, morphological, and syntactic), modalities (e.g., comprehension and production), and relevant related cognitive processes (e.g., verbal and phonological working memory, naming fluency, phonemic awareness).

Typical and atypical acquisition of Arabic

The status of research on typical and atypical acquisition of the Arabic language, although limited in scope, has provided the foundation for the development of the ALEF. Here we review the existing relevant literature with regard to the aspects of language acquisition that are reflected in the ALEF: phonology, morphology, syntax, and literacy. We also pay special attention to the distinct features of Arabic, namely diglossia and dialectal variation in spoken language.

Phonology

In Arabic phonology, most published research has focused on typical development. For example, Amayreh and colleagues (Amayreh, 2003; Amayreh & Dyson, 1998, 2000; Dyson & Amayreh, 2000) established developmental trajectories for the acquisition of Arabic consonants in a sample of Jordanian children 2 to 6 years of age. The study focused on the standard form of Arabic ("Educated Spoken Arabic"), rather than the vernacular form children would be acquiring as their mother tongue, which likely affected the results, especially with respect to late developing sounds. Overall, these studies indicated patterns of Arabic consonant acquisition similar to the patterns typically observed cross-linguistically, but with some Arabic-specific effects.

Thus, just as has been reported for other languages, bilabial and alveolar (nonemphatic) stops and nasals were the earliest acquired consonants (acquired by the age of two to three); most fricatives, particularly sibilants, were acquired later (by the age of six), with the rhotics, interdental fricatives, affricates, and emphatic consonants being late acquired (with complete mastery not reached until the age of nine). Interestingly, these studies indicated that atypical trajectories of phonological development in Arabic might be best captured by tests that emphasize the production of consonants in word-medial positions, typically ignored in English-based tests of articulation.

A study of phonological disorders in Jordanian Arabic demonstrated that children with phonological deficits employ a variety of substitutions and simplifying lenitive processes, both common in normal development cross-linguistically, such as fronting, stopping, prevocalic voicing, deaffrication, assimilation, final consonant deletion, and consonant cluster reduction, as well as atypical processes also found in disordered phonological development in other languages, such as backing and glottal replacement, and Arabic-specific processes, such as de-emphasis and spirantization (Bader, 2009).

A descriptive study of phonological development in 140 Qatari children 16 to 40 months of age (Al-Buainain, Shain, Al-Timimy, & Khattab, 2012) reported examples and frequencies of occurrence for various phonological errors in children's spontaneous speech, largely consistent with patterns found in other studies of phonological acquisition of Arabic.

Finally, a recent study of speech sound acquisition in Kuwaiti Arabic found more advanced speech sound acquisition compared to previous studies, with many complex sounds in place in the speech of 4-year-olds, including pharyngeals and uvulars. The emphatic consonants, the trilled /r/, and sibilants were the groups of sounds still produced with errors in the speech of 5-year-olds (Ayyad, Bernhardt, & Stemberger, 2016).

Morphology

A unique aspect of Arabic grammar that makes it particularly interesting for the study of DLD is its pervasive morphological complexity, such that nearly all words are morphologically complex, containing at least two templatic morphemes: a tri- or quadriconsonantal root, which encodes the semantic meaning (Holes, 2004), and a vocalic pattern, which denotes grammatical information (e.g., part of speech, tense, number). The abstract templates (i.e., consonantal roots and the vocalisms) constitute separate morphemes (Habash, 2007) that have to be acquired separately by the child. As these morphemes never occur as continuous phonetic entities, they must be inferred from underlying distributional patterns (Boudelaa, Pulvermuller, Hauk, Shtyrov, & Marslen-Wilson, 2010). Another unique property of Arabic morphology is that there is no clear "division of labor" between templatic and affixal morphemes with respect to their function (inflection versus derivation), as both perform each function, thus blurring the lines between inflectional and derivational morphology.

There are few published studies of the acquisition of morphology by Arabic-speaking children. Published studies, limited by small sample sizes, focused on plural noun inflection (Abdalla, Aljenaie, & Mahfoudhi, 2013) and tense and agreement (Abdalla & Crago, 2008). The latter study of Hijazi Arabic (spoken in Jedda, Saudi Arabia) reported over 94% correct subject–verb agreement marking in both present and past tense in typically developing children between 2.0 and 5.2 years of age. In contrast, it was found that typically developing children still may not reach high proficiency with plural marking of nouns by the age of 5 (Abdalla et al., 2013), resorting to regular substitutions of irregular ("broken") plural forms as well as rule-based masculine "sound plural" with a default feminine sound plural. The findings regarding children with DLD were in accord with what has been reported cross-linguistically, namely that children with DLD have pronounced difficulty with verbal tense and agreement morphology (Abdalla & Crago, 2008). The errors included errors of omission, when children used an imperfective bare stem, thought of as an acquisitional default. Children with DLD also substantially underperformed on plural noun inflection, a grammatical category that has shown mixed results in cross-linguistic studies of DLD (Abdalla et al., 2013), but has been proposed as a potential specific language impairment (SLI) marker in Arabic.

Syntax

There is a notable scarcity of research with respect to the development of syntax in Arabic. Arabic subject–verb word order is variable and determined by a complex confluence of factors related to the properties of a given noun or a pronoun, definite or indefinite, as well as its discourse-pragmatic status, i.e., whether it expresses given or new information (Owens, Dodsworth, & Rockwood, 2009). Despite this greater variability of word order, a study of the acquisition of Egyptian Arabic (Omar, 1973), based on recorded speech samples of 37 children from 6 months to 15 years old, indicated that most aspects of syntax were acquired by the children on a timeline

comparable to what was reported for other languages (between 2 and 5 years of age), and that their word order was flexible from early on. This suggested that typically developing children successfully master the complex relationship between word order and the syntactic/pragmatic categories that determine it at an early age. With regard to the syntactic development of children with DLD, an unpublished study (Shaalan, 2010) demonstrated that child speakers of Gulf Arabic affected with DLD had pronounced difficulties with complex syntactic structures, particularly those that involved the fronting of object noun phrases and verbs.

Literacy

Studies of literacy acquisition in Arabic indicate that while the acquisition of reading skills in Arabic is heavily dependent on both visual and phonological processing skills, the development of phonemic awareness and phonological memory is an especially critical area of weakness for Arabic children with DLD and reading disability (Abu-Rabia, Share, & Mansour, 2003; Ibrahim, 2011). This supports the idea that reading acquisition follows certain universal patterns across alphabetic and consonantal (abjad) orthographies, as phonological awareness has been shown to be an important predictor of reading development across a number of languages (Ziegler et al., 2010). This research also pointed out some interesting Arabic orthography-specific effects, namely demonstrating that phonological processing might be an even more critical skill for Arabic literacy acquisition given the existence of two parallel lexicons (one of the spoken and the other of the literary language), which places great demands on the phonological abilities of the child (Bentin & Ibrahim, 1996; Ibrahim, 2011).

Language variation and diglossia in Arabic

Like many other languages, and perhaps to a much greater extent than most other languages, due to being the mother tongue in many countries across a large geographic area, Arabic is characterized by considerable variation among its spoken varieties. This presents a challenge for creating a language development assessment, as it needs to be adapted to each spoken variety or be dialect-neutral. Furthermore, because of the exclusive use of Classical or Modern Standard Arabic as the language of literacy in the Arab world, the spoken language differs from the written/formal Arabic. Thus, although the distinction between written and spoken language and formal and informal registers exists in many languages, Arabic has a more complex situation known as diglossia (Ferguson, 1959), that is, regular parallel use of two distinct languages for distinct purposes: one in everyday life, the other in formal situations and written communication. Classical Arabic, as the language of the holy Qur'an, has been carefully preserved in written form as its modernized equivalent, Modern Standard Arabic (MSA). This highly codified formal language, largely learned through formal education, exists alongside the dynamic and constantly evolving spoken vernacular. The two are highly divergent from each other. The linguistic distance between the two Arabic varieties is particularly acute in phonology, morphosyntax, word formation, and syntax, including phonemic inventories, syllabic structure, phonotactic constraints, stress patterns, and inflectional categories (Aoun, Benmamoun, & Sportiche, 1994; Aoun, Choueiri, & Benmamoun, 2010; Brustad, 2000; Holes, 2004; Watson, 2002). For example, MSA has a richer system of agreement compared to the less differentiated systems in some spoken varieties— different word order, distribution and frequency of verbal patterns, passivization, and nominal constructions (Benmamoun, 2000; Shawarbah, 2007; Shlonsky, 1997). Lexically, MSA and spoken varieties overlap only partially, with 80% of the lexicon of young children consisting of words with divergent forms in spoken and formal Arabic.

One consequence of the existence of the spoken vernaculars alongside the standard written form, each reserved for a distinct purpose, is that the latter is revered by Arabic speakers the world over as beautiful, perfect and "correct," while the former has low prestige and is not even always acknowledged as a legitimate language, but rather as "slang" (Versteegh, 1996). This attitude creates a serious challenge for developing assessments of child language development, as applying the prescriptive norms of literary Arabic to a spoken language assessment would not yield a valid measure. However, as all educated adults are trained on these prescriptive norms, it is quite challenging to overcome these language attitudes and develop norms based on the spoken language. Secondly, because of the linguistic distance between MSA and spoken Arabic, learning literary Arabic in school has been likened to learning a second language (Maamouri, 1998). For example, one study demonstrated that Arabic-speaking students faced with a lexical decision task in spoken Arabic exhibited priming effects only when primed with a spoken Arabic word and not with either Hebrew or literary Arabic words

(Ibrahim, 2009). This suggests that for Arabic speakers, who are bi-dialectal, their spoken and written lexicons are distinct, similarly to what was found with respect to lexical organization of nonbalanced bilinguals (Kroll & de Groot, 1997; Sholl, Sankaranarayanan, & Kroll, 1995). This means that for Arabic-speaking children literacy develops essentially in a second language (spoken Arabic being the primary and literary Arabic being the secondary language); hence, literacy acquisition in Arabic is a process substantially different from that in monolingual, nondiglossic environments.

Need for a standardized assessment tool of spoken Arabic

Due to the unique diglossic nature of the language learning environment in Arabic-speaking countries and the properties of the Arabic language (including its complex root-and-pattern morphology and flexible word order) and writing system (e.g., consonantal script), quantitative studies of the acquisition of Arabic with adequate sample sizes are very important, as they may contribute unique insights to the understanding of universal aspects of language development and impairment, as well as show how language-specific characteristics influence typical patterns acquisition and the presentation of DLD.

Currently, it is widely acknowledged that there is a dearth of research on DLD in Arabic (Abdalla & Crago, 2008) and an even direr need for the development of theory-rooted and psychometrically-sound assessment tools. The majority of existing language tests used in Arabic are direct translations of English tests, which have validity problems, as they show lack of sensitivity to age-related change in the age range 6–9 (Balilah & Archibald, 2018). There are few tests that have been developed for and validated on Arabic-speaking children. A recent special issue of the *Arab Journal of Applied Linguistics* (Mahfoudhi & Abdalla, 2017) included a paper that presented a set of Arabic developmental language tests for Qatari-speaking children (Shaalan, 2017), which showed limited sensitivity to age-related change. An only other assessment of language development in Arabic with a preliminary report of adequate psychometric properties is, the Egyptian Arabic Pragmatic Language Test (EAPLT) (Khodeir, Hegazi, & Saleh, 2017).

This means that the validation and verification of the battery presented in the current study, the ALEF battery, will provide an invaluable tool for identifying and measuring language development in Gulf Arabic-speaking children. The ALEF is a comprehensive standardized assessment of language comprised of three modules. The main battery involves a wide range of tests of spoken language, encompassing all linguistic domains in both expressive and receptive modalities: phonology (i.e., word articulation, sound discrimination), lexicon (receptive and expressive vocabulary), grammatical morphology, comprehension, production and imitation of complex syntax, and pragmatic knowledge. In addition, there are two supplementary batteries: one comprised of measures of important prerequisite cognitive processes relevant for language development (i.e., short-term memory, pseudoword repetition, serial naming speed, and phonological awareness); and another targeting basic literacy skills (word- and pseudoword decoding, spelling, paragraph reading fluency, and reading comprehension). Thus, the ALEF provides a way to systematically assess a wide variety of language skills in children from 3 to 11 years of age, constituting an invaluable tool for both clinicians and researchers.

Methods

Participants

We recruited 467 Arabic-speaking children from local kindergartens and schools in the Dammam region, where the population is relatively ethnically homogeneous and predominantly native Saudi. Only public child-care centers and schools were sampled. Based on observations in the literature (Denman et al., 2017) and preliminary power analyses, the recruitment design aimed for 20 children (10 boys, 10 girls) in each of the 13 sampled 6-month age bands. The inclusion criteria included: (a) Arabic as the native language; (b) at least two generations (parent and grandparent) of Saudi nationalities; (c) knowledge of the tribal origin of the family; and (d) Saudi citizenship. The exclusion criteria included known history of (a) any developmental or educational problems; (b) known medication for serious conditions (e.g., epilepsy); and (c) any explicit knowledge, by administrators at child care centers and schools, of ongoing source of stress in children's lives (e.g., loss of a parent, family distress).

Our goal was to develop a comprehensive, theory-driven and psychometrically-sound assessment of language development and functioning in Arabic. Correspondingly, for the vast majority of the 17 ALEF subtests, described in the ALEF Battery section, we opted to initially develop an excessive number of test items that would be calibrated in terms of their psychometric properties. Given the large size of the ALEF battery, and since the participants included young children,

Table 1. Psychometric properties of the Arabic Language: Evaluation of Function (ALEF) battery subtests.

Module	Subtest	Wave	*N*	K items	K items retained	α	ω	Var_{IRT}	r_{xx}
ALEF-SL	Expressive Vocabulary	1	236	53	47	0.89	0.89	0.24	0.89
ALEF-SL	Receptive Vocabulary	1	240	53	43	0.90	0.92	0.39	0.90
ALEF-SL	Word Articulation	1	240	46	28	0.90	0.92	0.53	0.79
ALEF-SL	Pseudoword Discrimination	1	240	42	40	0.94	0.94	0.44	0.90
ALEF-SL	Sentence Completion	2	227	33	30	0.90	0.91	0.21	0.91
ALEF-SL	Sentence Comprehension	2	227	31	23	0.86	0.88	0.43	0.78
ALEF-SL	Sentence Imitation	1	240	35	28	0.97	0.98	0.53	0.96
ALEF-SL	Pragmatic Knowledge (1)	2	203	22	18	0.84	0.86	0.45	0.81
ALEF-SL	Pragmatic Knowledge (2)	2	98	21	18	0.79	0.83	0.34	0.79
ALEF-WL	Spelling	2	104	42	36	0.95	0.96	0.72	0.89
ALEF-WL	Word Reading	2	113	66	60	0.97	0.98	0.70	0.96
ALEF-WL	Pseudoword Reading	2	113	55	50	0.95	0.96	0.69	0.92
ALEF-WL	Paragraph Comprehension (1)	2	22	10	10	0.67	0.81	–	–
ALEF-WL	Paragraph Comprehension (2)	2	63	10	10	0.73	0.78	–	–
ALEF-WL	Paragraph Comprehension (3)	2	65	10	10	0.78	0.85	–	–
ALEF-WL	Paragraph Comprehension (4)	2	36	10	10	0.73	0.82	–	–
ALEF-CP	Pseudoword Repetition	1	240	34	33	0.90	0.91	0.41	0.89
ALEF-CP	Phoneme Awareness	2	132	45	37	0.91	0.93	0.46	0.92
ALEF-CP	Digit Span	1	240	32	32	0.72	0.82	–	–
ALEF-CP	Digit Span – Forward	1	240	16	16	0.75	0.83	–	–
ALEF-CP	Digit Span – Backward	1	240	16	16	0.73	0.85	–	–
ALEF-CP	RAN (Time, Average)	1	198	4	4	0.88	0.90	–	–
ALEF-CP	RAN (Errors, Average)	1	198	4	4	0.74	0.85	–	–

Note. α = Cronbach's coefficient alpha; ω = MacDonald's coefficient omega; Var_{IRT} = proportion of variance in response patterns explained by an IRT model; r_{xx} = empirical reliability of estimated IRT ability scores; RAN = Rapid Automatized Naming.

both fatigue and logistical considerations led to a study design that involved administering the ALEF subtests in two independent data collection waves (see Table 1).

For Wave 1 data collection, we administered the following subtests—Expressive Vocabulary, Receptive Vocabulary, Word Articulation, Pseudoword Discrimination, Sentence Imitation, Pseudoword Repetition, Digit Span, and Rapid Automatized Naming—to 240 children in the age range from 2.5 to 10.92 years ($M = 6.22$, $SD = 1.71$; 119 girls, 121 boys). For the Wave 2 data collection, we administered the remaining subtests— Sentence Completion, Sentence Comprehension, Pragmatic Knowledge, Spelling, Word Reading, Pseudoword Reading, Paragraph Comprehension, and Phoneme Awareness—to $n = 227$ children in the age range from 2.75 to 9.83 years ($M = 6.40$, $SD = 1.86$; 93 girls, 134 boys). The reading subtests were only administered to children age 5 and above. Note that we administered ALEF to only two children below the age of 3 years. The removal of their data from the dataset did not result in changes in parameter estimates, and we therefore retained their data.

Informed consent was obtained from the children's parents, and all children provided an oral assent for completing the study tasks at the time of the assessment. Ten booklets were assembled to randomize the effect of specific sequences of first and last subtests; no booklet (subtest order) effect was documented. Testing was discontinued for younger children when excessive fatigue was detected.

ALEF battery

The ALEF battery is an individually administered assessment of language development in Arabic that consists of three modules: ALEF-SL (Spoken Language, the core module), ALEF-WL (Written Language, a supplemental module), and ALEF-CP (Language-Related Cognitive Processes, a supplemental module).

The choice of subtests for the ALEF was guided by two chief considerations. First, we followed the descriptive-developmental model of DLD as the one with the most clinical utility (Paul & Norbury, 2012). According to this model, the best-evidence method of language assessment for the purpose of identifying DLD includes obtaining a detailed description of a child's performance on a range of linguistic tasks involving speaking and listening (the core module ALEF-SL), and locating the child's performance in the sequence of normal development. These tasks may be augmented by, but not replaced with, an assessment of collateral skills (e.g., short-term memory; ALEF-CP). In addition, for school-age children, the evaluation of literacy skills (ALEF-WL) may be appropriate. Overall, whether used for the purposes of establishing base rate performance for selecting intervention goals, tracking intervention progress, or for grouping children in research studies, any subset of the subtests may be used in any combination, as appropriate.

Secondly, the subtests selected for each module aim at being fully comprehensive, covering a wide range

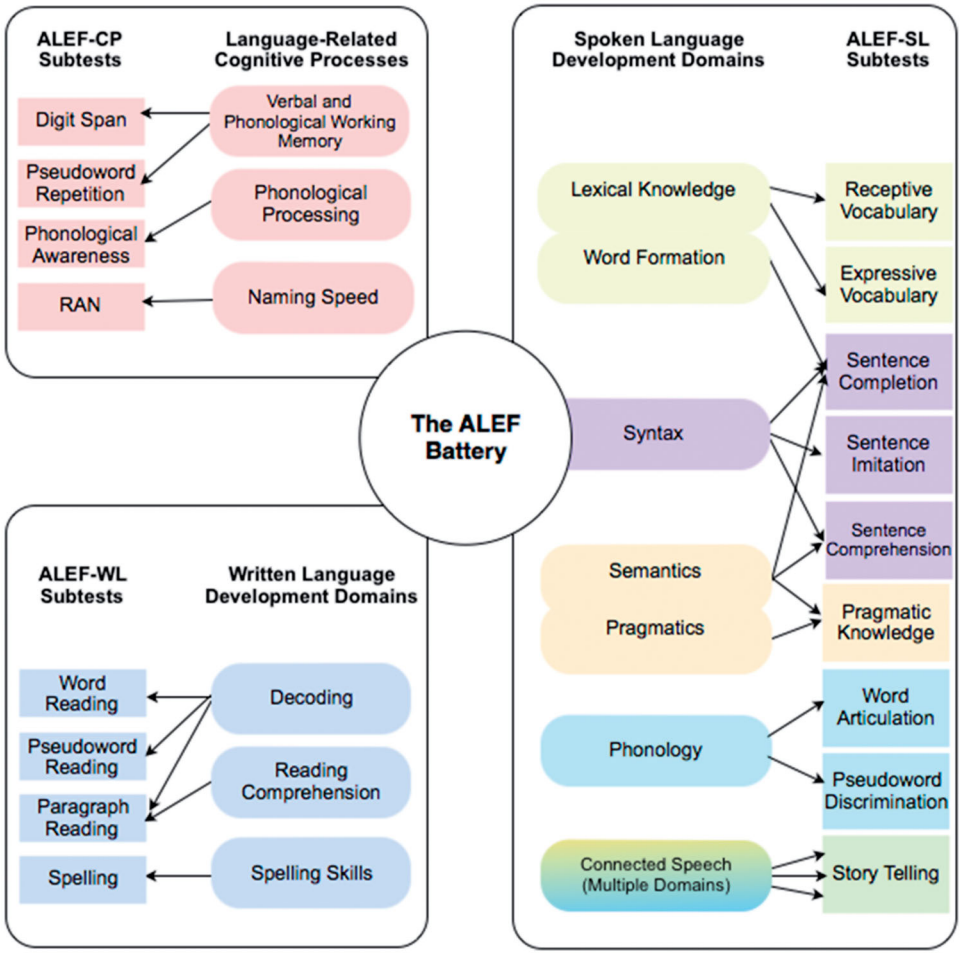

Figure 1. The modular structure of the ALEF-battery. ALEF-SL = ALEF-Spoken Language; ALEF-WL = ALEF-Written Language; ALEF-CP = ALEF-Language-Related Cognitive Processes.

of skills in all major linguistic domains in different modalities. For example, the subtests in ALEF-SL cover skills in all oral language domains in both expressive and receptive modalities within the form-content-use framework (Bloom & Lahey, 1978; Paul & Norbury, 2012): phonology, morphology and syntax (form), lexical semantics (content) and pragmatics (use). The subtests in ALEF-WR cover phonological decoding, reading fluency and reading comprehension, as well as spelling. Finally, ALEF-CP combines subtests that can be grouped under the umbrella of "phonological processing," broadly defined, and include subtests similar to those included in the Comprehensive Test of Phonological Processing, CTOPP-2 (Wagner et al., 1999), widely used to asses language and reading-related skills in children.

With respect to specific skills, the core, spoken language (ALEF-SL) module (see Figure 1) consists of the following subtests: Receptive Vocabulary, Expressive Vocabulary, Word Articulation, Pseudoword Discrimination, Sentence Comprehension, Sentence Completion, Sentence Imitation, Story Telling[1], and Pragmatic Knowledge. ALEF-WL contains the following subtests: Word Reading, Pseudoword Reading, Paragraph Reading, and Spelling. In addition, ALEF-CP contains the following subtests: Pseudoword Repetition, Digit Span, Rapid Automatized Naming, and Phonological Awareness. Together, these 17 subtests represent key linguistic skills tested both receptively and expressively, which differ in their modes of administration and response (e.g., picture selection, picture naming, cloze format).

The development of ALEF was carefully aligned with the considerations presented in the Introduction, and was conducted by an international team comprised of specialists in language acquisition, psychometrics, neuro-cognitive development and the Arabic language. The development of the test structure and test items was completed in several iterations with

[1]Results of the Story Telling Task will be presented separately, as they require transcription and detailed linguistic analyses.

three rounds of reviews of the test materials (including stimuli, instructions, and training and scoring materials for the test administrators and data collectors) conducted by a panel of specialists from King Faisal University. Illustrations for test items were commissioned from a professional graphical artist, and were carefully reviewed for cultural appropriateness as well as construct validity with respect to the targeted lexical and grammatical categories. Test administration was performed by a set of trained school teachers who received a two-day training course on the basics of the assessment procedures in general and the details of the ALEF administration in particular.

Spoken language (ALEF-SL) module

Receptive Vocabulary and Expressive Vocabulary. The aim of these two subtests is to assess the receptive and expressive knowledge of concrete vocabulary and basic concepts using developmentally appropriate lexical items sampled from main grammatical categories (nouns, verbs, adjectives, adverbs, and prepositions), as well as various semantic classes (animals, transportation, home/household items, body parts, nature, food, clothing, materials, emotions, actions, colors, and shapes). In addition, they assess the knowledge of verbs sampled from all Arabic verb classes (I-X). Words were selected and arranged according to a three-tier taxonomy: (a) basic, high-frequency words learned in everyday social interactions; (b) high- and moderate-frequency words mainly learned through educational experiences; and (c) more advanced words, learned from exposure to scholarly texts. The words were selected from a frequency dictionary of Arabic (Buckwalter & Parkinson, 2014) and their appropriateness for inclusion was confirmed independently by native Arabic speaking experts with advanced degrees in Psychology and Linguistics. The procedure for the Expressive Vocabulary subtest is picture naming: the child is shown pictures and asked to name each one. In the Receptive Vocabulary subtest, the child is shown three pictures for each item (one depicting the target object, concept or action, and two distractors) and asked to point to the picture depicting the target given by the test administrator. The subtests took around 10 minutes to complete (each).

Sentence Comprehension and Sentence Completion. These two subtests measure the child's receptive and expressive grammar skills. They are designed to target the following specific aspects of Arabic grammar: formation of plural nouns, verb tenses, subject–verb agreement, and comparative forms of adjectives. During the receptive subtest, Sentence Comprehension, the child is shown three pictures (one corresponding to the target sentence and two foils) and asked to point to the picture that corresponds to the target sentence. In the corresponding Sentence Completion test, the child is shown one picture and given the sentence describing it, then shown a second, related picture and asked to complete the sentence describing this picture using a cloze format. For example, for an item targeting the child's knowledge of plural noun formation, he or she is given sentences with their corresponding pictures analogous to the following: "In this picture, there is a cat, and here we have two _____" (expected response, "cats"). The administration of these two subtests took around 7 minutes per subtest.

Sentence Imitation. Research in (a)typical language acquisition has established the important role of sentence imitation as a probe into the child's internalized grammatical system. Elicited Sentence Imitation allows us to assess children's knowledge of precise grammatical factors, and, based on any changes (omissions or substitutions) that the child makes in his or her response, to evaluate his or her stage of grammatical development. This assumption is based on the observation confirmed by a large amount of empirical evidence that in order for the child to imitate a structure, the structure must already be part of the child's grammatical competence (Lust, Chien, & Fynn, 1987). In addition, processing and imitating complex sentence structures taps working memory resources (Just & Carpenter, 1992) and is one of the best tasks for differentiating children with DLD from their typically developing peers . Our subtest contains sentences of varying and increasing complexity that target specific syntactic structures (e.g., wh-questions of varying complexity, sentences with negation, sentences containing complex predicates and conjoined clauses, sentences with subordinate clauses of various types and complexity). The child is asked to repeat them exactly as they are spoken by the adult, and the score for each item is assigned based on the number and type of errors (e.g., omission, substitution, permutation). The subtest took around 7 minutes to complete.

Story Telling. Research has shown that eliciting narratives is a valid measure of language development both in typically and atypically developing children (Norbury & Bishop, 2003). For the purpose of the assessment, we developed a supplementary subtest in the form of a culturally appropriate wordless picture-story (analogous to the one widely used in the West, "*Frog, Where are You?*" by M. Meyer) with original illustrations. The child is asked to look through all of the pictures and then tell a story about them.

The narrative is recorded and may be scored using two different options, as dictated by the specific needs of the researcher/clinician. First, there is a brief option that does not require transcription and therefore can be used in clinical practice by implementing a "rough-and-ready" scoring procedure in the following scales: (a) intelligibility; (b) amount of output; (c) grammatical well-formedness; and (d) global narrative quality. The second option is more detailed and involves transcribing the narrative and scoring it on a number of lexical, syntactic and pragmatic characteristics, which provide a rich source of information with regard to the child's level of productive language, measured in an ecologically valid way (as part of connected discourse). Given the open-ended nature of this subtest and the resource demands of appropriate transcription and scoring procedures, we do not report the data on Story Telling in this manuscript, while explicitly stipulating that such an analysis will be performed separately through the lens of clinical linguistics and language acquisition. The subtest took around 8 minutes to complete.

Pragmatic Knowledge. This subtest consists of two parts. Part 1 (Social Use of Language) is aimed at measuring the child's familiarity with the conventions guiding the use of language in social situations. This part is designed for use with children 3–11 years of age and consists of orally presented situations (accompanied by pictures as memory aids) followed by a question. Each item describes an aspect of daily life that requires using language in a certain way to accomplish a particular goal (e.g., greeting, offering help, asking for help, requesting food, comforting a friend in distress, congratulating). Each response is judged on whether it is (a) appropriate for the situation; (b) in a socially appropriate form; and (c) formulated correctly. Part 2 of the subtest (Inferences) is aimed at measuring more advanced pragmatic skills, namely being able to infer implied meaning from stated information. In some items the inference depends on world knowledge, in others on the knowledge of semantically complex words, and in still others on familiarity with conversational conventions. Thus, it assesses children's knowledge of those aspects of meaning construction that go beyond lexical semantics and the meaning directly asserted in the sentence. This part is aimed at children 6–12 years of age and includes mini-scenarios and a question posed to the child at the end of each situation. The child is instructed to listen to the stories and answer the questions using clues from the story. Items targeting different types of implied meaning and requiring imputation of the speaker's intention are included. The scoring rule for each item is explicitly provided. This subtest is specifically targeting the language difficulties typical in children with Autism Spectrum Disorders and Social Communication/Pragmatic Language Disorder. The subtest took around 12 minutes to complete.

Word Articulation and Pseudoword Discrimination. These are two phonetic tests aimed at evaluating the sound inventory development of spoken Arabic. In the Word Articulation subtest, the child is given an opportunity to pronounce a list of words, each containing either a certain developmentally significant consonant or a difficult to articulate consonant cluster in the initial, medial or final position. To avoid having to provide an auditory prompt containing the target sound (to avoid the child imitating the adult), the child is prompted by a picture depicting each item and asked to complete a sentence spoken by the test administrator using the word depicted in the visual prompt. In the Pseudoword Discrimination subtest, the child is given pairs of contrasting pseudowords (i.e., identical except for one consonant or consonant cluster) interspersed with identical pairs and asked to judge whether the two in each pair are the same or different. This procedure is aimed at testing the child's ability for sound discrimination, a skill crucial for both spoken and subsequent written language development. Both subtests took around 8 minutes to complete each.

Written language (ALEF-WL) module

The written language module aims at evaluating a range of literacy skills and includes the following subtests: *Spelling, Word Reading, Pseudoword Reading, and Paragraph Reading*. This part of the assessment includes the measures typically used to assess a child's reading and spelling skills: namely, (a) a dictated Spelling task (a list of words of varying frequency and complexity are presented first in isolation and then in a syntactic frame, and the child is asked to spell each word; 5 minutes); (b) Word Reading and Pseudoword Reading (a list of words and pseudowords of varying complexity spelled in the vowelized script that the child is asked to read out loud; the number of items correctly read in 1 minute is recorded); and (c) a Paragraph Reading subtest that indexes reading fluency and reading comprehension (two age-appropriate passages that the child is asked to read, while the examiner records the number of words read correctly in one minute; the child is then asked to answer multiple questions based on the content of the passages to assess reading comprehension; 15 minutes). This module is designed to be used with children of school ages and is aimed at identifying difficulties with literacy

acquisition, characteristic of children with various types of DLD.

Collateral cognitive processes (ALEF-CP) module

This part of the assessment evaluates cognitive components identified as relevant prerequisites for language and literacy development, namely, verbal working memory (Digit Span), phonological short-term memory (Pseudoword Repetition), naming speed (Rapid Automatized Naming), and Phonological Awareness (Elision task). These are commonly used short assessments that involve asking the child to repeat series of digits, forward and backward, repeat pseudowords of increasing length and complexity, name as fast as they can a serially arranged recurring set of letters, numbers, colors and objects (depicted on 4 cards, one for each type of stimuli), and to repeat a word and then modify it by saying it without one sound or syllable elided either from the beginning, the middle or the end of the word. Pseudoword repetition tasks have been previously identified as a robust clinical marker for DLD (Conti-Ramsden, Botting, & Faragher, 2001). Phonological awareness and rapid automatized naming have been shown to be impaired in children with reading disability (Vellutino, Fletcher, Snowling, & Scanlon, 2004; Wolf, Bowers, & Biddle, 2000). Finally, verbal working memory has been previously identified as an important factor in both reading comprehension (Carretti, Borella, Cornoldi, & De Beni, 2009) and sentence processing (Montgomery & Evans, 2009). The ALEF-CP module took around 20 minutes to complete.

Statistical analyses

To establish the ALEF's psychometric properties, we used a set of complimentary approaches. First, all of the ALEF subtests were examined for their internal consistency (as indexed by Cronbach's coefficient alpha α) to evaluate reliability. Following recent recommendations in the literature (Dunn, Baguley, & Brunsden, 2014), we also computed McDonald's coefficient omega, ω, a recent alternative to coefficient alpha based on the congeneric (compared to the tau-equivalent) model, which estimates the general factor saturation of a subtest, and follows similar interpretational rules.

In addition, we fit a set of item response theory (IRT) models to item-level response pattern data. Unidimensional IRT models were fit using the *mirt* package for R (Chalmers, 2012): 2PL (estimating item difficulty and discrimination) models were fit for Receptive Vocabulary, Word Articulation, Pseudoword Discrimination, Pseudoword Repetition, Spelling, Phoneme Awareness, Sentence Comprehension, Pragmatic Knowledge, Word and Nonword Reading. A generalized partial credit model (GPCM) was fit to the data for subtests that allowed partial credit and required rubric-based scoring: Sentence Imitation, Sentence Completion, and Expressive Vocabulary. Subtest scores were computed using the expected *a posteriori* (EAP) method after the removal of items that showed <1% or >95% average response accuracy (for dichotomous items), showed a negative item-total correlation, or displayed significant IRT misfit according to the plausible value statistic (PV $p < .05$).

Age and gender effects were established by fitting a set of linear models to the data with centered age and age^2 (to account for possible nonlinearity) terms and dummy-coded gender (0 = girl, 1 = boy) as fixed effects predicting subtest scores. P-values were obtained using permutations for the linear model as implemented in the *lmPerm* R package.

If the child failed to comply with the instructions or did not provide a single correct response for the entire subtest, their data were coded as missing.

Results

Preliminary psychometric analyses of the ALEF battery found that all of the subtests (but one paragraph in Reading Comprehension) had high to very high internal consistency estimates (generally convergent for Cronbach's and McDonald's estimates). After the removal of poorly performing items (from 2.9% removed items in Pseudoword Repetition to 18.9% removed items in Receptive Vocabulary), internal consistency coefficients ranged from $\alpha = .67$ to .98 (Median = .88) and $\omega = .78$ to .98 (Median = .89), providing strong support for ALEF's reliability as an assessment of early language development in multiple language domains and modalities (see also Table 1 for a summary of key psychometric properties).

IRT models explained from 21% (Sentence Completion) to 72% (Spelling) of variance in item response patterns (Median estimate = 44.5%), with estimated empirical reliabilities for EAP scores ranging from .78 (Sentence Comprehension) to .96 (Word Reading and Sentence Imitation; Median estimate = .90). Obtained item characteristic curves (ICCs) are presented in the Supplementary Data.

Subtest scores were generally normally distributed, but with significant (Shapiro-Wilk's $p < .05$ corrected for multiple testing using the Bonferroni correction) deviations from normality established for Word Articulation

Table 2. Estimated regression model parameters for Arabic Language: Evaluation of Function (ALEF) subtests (with permutation-based *p*-values).

	Intercept	Age		Gender		Age*Gender		Age^2		Age^2 x Gender		Overall model			
		B	*P*	*B*	*P*	*B*	*P*	*B*	*P*	*B*	*P*	R^2_{ADJ}	*F*	df	*p*
Expressive Vocabulary	0	1.62	2×10^{-16}	0.05	0.7059	0.58	0.0144	−0.11	2×10^{-16}	−0.04	0.0170	0.2738	18.72	5,230	1.3×10^{-8}
Receptive Vocabulary	0	0.85	2×10^{-16}	0.04	0.5100	0.03	1	−0.05	2×10^{-16}	0.00	1	0.1192	7.47	5,234	1.6×10^{-6}
Word Articulation	0	0.67	0.0266	0.26	2×10^{-16}	0.85	2×10^{-16}	−0.05	0.0406	−0.07	2×10^{-16}	0.1447	9.09	5,234	6.5×10^{-8}
Pseudoword Discrimination	0	0.37	0.0435	−0.02	1	−0.11	0.4155	−0.02	0.1463	0.01	0.4082	0.0536	3.71	5,234	0.0030
Sentence Imitation	0	1.74	2×10^{-16}	0.01	0.9800	−0.03	0.9800	−0.13	2×10^{-16}	0.00	1	0.2259	14.95	5,234	9.8×10^{-13}
Pseudoword Repetition	0	1.02	2×10^{-16}	−0.02	0.5275	0.43	0.0261	−0.07	2×10^{-16}	−0.03	0.0267	0.1207	7.4	5,228	1.9×10^{-6}
Digit Span	6.92	4.07	2×10^{-16}	0.35	0.0450	1.21	0.8040	−0.27	2×10^{-16}	−0.09	0.6550	0.2055	12.9	5,225	4.9×10^{-11}
Digit Span - Forward	5.01	1.51	2×10^{-16}	0.28	0.0186	0.63	0.1952	−0.10	0.0002	−0.04	0.2213	0.0882	5.45	5,225	9.3×10^{-5}
Digit Span - Backward	1.91	2.56	2×10^{-16}	0.08	0.2413	0.58	0.0402	−0.17	2×10^{-16}	−0.05	0.0393	0.2489	16.24	5,225	1.1×10^{-13}
RAN - Errors	6.30	−15.41	2×10^{-16}	−0.70	0.1006	−3.06	0.0355	1.00	2×10^{-16}	0.26	0.0341	0.5443	48.07	5,192	2.2×10^{-16}
RAN - Time	69.44	7.71	0.7451	−6.39	0.0020	−8.76	0.3961	−1.31	0.0853	0.51	0.4110	0.1951	10.55	5,192	5.8×10^{-9}
Sentence Completion	0	0.26	2×10^{-16}	−0.06	0.8200	−0.02	0.4150	–	–	–	–	0.2535	26.58	3,223	9.8×10^{-15}
Sentence Comprehension	0	−0.01	0.7450	−0.03	0.3730	0.03	0.2250	–	–	–	–	0	0.46	3,223	0.7096
Pragmatic Knowledge (1)	0	1.01	2×10^{-16}	−0.10	0.6667	−0.11	1	−0.08	0.0002	0.00	1	0.0619	3.67	5,197	0.0034
Pragmatic Knowledge (2)	0	4.16	0.0600	−0.43	2×10^{-16}	4.69	0.0234	−0.27	0.0670	−0.32	0.0329	0.1238	3.74	5,92	0.0040
Phoneme Awareness	0	0.36	2×10^{-16}	−0.10	0.0708	0.08	0.1990	–	–	–	–	0.2017	9.59	3,99	1.3×10^{-5}
Spelling	0	1.52	0.1680	−0.01	0.8630	−0.22	1	−0.07	0.2410	0.03	1	0.1632	4.978	5,97	0.0004
Word Reading	0	0.25	0.0064	0.39	0.1000	−0.03	0.6154	–	–	–	–	0.1784	9.11	3,109	2.0×10^{-5}
Nonword Reading	0	3.28	0.0132	−0.04	0.8235	2.13	0.0600	−0.19	0.0219	−0.13	0.0593	0.0300	1.71	5,101	0.1384
Reading Comprehension	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Note. Gender was dummy coded as 0-girl, 1-boy. RAN = Rapid Automatized Naming.

and Pseudoword Discrimination (due to ceiling effects), as well as Receptive Vocabulary, Sentence Imitation, Spelling, Word and Nonword Reading subtests, and the RAN measures. Visual analyses of the subtest score distributions suggested that the departures from normality in Receptive Vocabulary, Sentence Imitation, Spelling, and reading subtests was likely due to the presence of a subpopulation with low scores that formed the second peak in the distributions, potentially pointing to possible DLD and RD-related deficits.

Gender and age effects were evaluated jointly in a set of linear models summarized in Table 2. Demographic variables jointly explained from 0 (Sentence Comprehension) to 54% (RAN Errors) variance in subtest scores, with the median estimate of $R^2 = 0.16$. With the exception of Spelling, Sentence Comprehension, and RAN Time, all ALEF subtests were sensitive to age effects (in many cases the quadratic effect for age was also significant) in the predicted direction (i.e., ability estimates increased positively with age), pointing to the validity of ALEF as a test of language development.

Corrected for multiple comparisons, only two subtests displayed significant gender effects – boys scored slightly higher than girls on Word Articulation (with the standard effect estimate of $B = 0.26$), and lower than girls on the Pragmatic Knowledge – Part 2 subtest ($B = -0.43$), indicating that this subtest is sensitive to potential gender differences in the developmental trajectories for pragmatic language skills.

Analyses of subtest intercorrelations were performed separately for wave 1 and wave 2 of the data collection, and revealed a predictable positive manifold of intercorrelations (see Tables 3 and 4). The strongest correlations for wave 1 were obtained between Digit Span and Expressive Vocabulary ($r = .60$, $p < .001$) and Digit Span and Pseudoword Repetition ($r = .51$, $p < .001$). Notably, naming error counts from the RAN subtest were also negatively related with most wave 1 subtests. For wave 2, the strongest correlations were established between Nonword Reading and Reading Comprehension ($r = .61$, $p < .001$), and Nonword Reading and Word Reading ($r = .52$, $p < .001$). Interestingly, although Phoneme Awareness scores did not correlate with reading subtests, they were positively related to Spelling ($r = .41$, $p < .001$), suggesting a close relationship between phonological skills and spelling ability in Arabic.

Overall, the results indicated that our newly developed assessment of Arabic language development and functioning showed high internal consistency, substantial

Table 3. Intercorrelations among the Arabic Language: Evaluation of Function (ALEF) battery subtests (Wave 1).

	1. EV	2. RV	3. WA	4. PD	5. PR	6. DS	7. DS-F	8. DS-B	9. RAN-E
1. EV									
2. RV	.35***								
3. WA	.26***	.02							
4. PD	.23***	.22***	.16*						
5. PR	.57***	.21**	.25***	.09					
6. DS	.60***	.34***	.15*	.26***	.51***				
7. DS-F	.44***	.23***	.11	.19**	.43***	.86***			
8. DS-B	.59***	.35***	.14*	.25***	.42***	.82***	.41***		
9. RAN-E	−.59***	−.29***	−.22**	−.17*	−.44***	−.43***	−.22**	−.50***	
10. RAN-T	−.08	−.08	.13	−.09	−.11	−.15*	−.15*	−.10	.20**

Note. Pearson's r coefficients are presented. EV = Expressive Vocabulary; RC = Receptive Vocabulary; WA = Word Articulation; PD = Pseudoword Discrimination; PR = Pseudoword Repetition; DS = Digit Span (F = forward; B = backward); RAN = Rapid Automatized Naming (E = Errors; T = Time).
*$p < .05$.
**$p < .01$.
***$p < .001$.

Table 4. Intercorrelations among the Arabic Language: Evaluation of Function (ALEF) battery subtests (Wave 2).

	1. SC	2.SCL	3. PK (1)	4. PK (2)	5. NWR	6. WR	7. RC	8. SP
1. SC								
2. SCL	.40***							
3. PK (1)	.33***	.49***						
4. PK (2)	.15	.18	.70**					
5. NWR	.29**	.23*	.24*	.26*				
6. WR	.23*	.31***	.21*	.18	.52***			
7. RC	.43***	.24	.33*	.17	.61***	.47***		
8. SP	.26**	.31**	.27*	.11	.24*	.40***	.03	
9. PA	.25*	.32**	.17	.06	.11	.19	.22	.41***

Note. Pearson's r coefficients are presented. SC = Sentence Comprehension; SCL = Sentence Completion; PK = Pragmatic Knowledge (Parts 1 and 2); NWR = Nonword Reading; WR = Word Reading; RC = Reading Comprehension; SP = Spelling; PA = Phoneme Awareness.
*$p < .05$.
**$p < .01$.
***$p < .001$.

age sensitivity, minimal gender bias, and predictable patterns of intercorrelations among subtests.

Discussion

The main goal of the study (representing the pilot stage of a long-term effort) was to establish the preliminary psychometric properties of the individual subtests of the ALEF. To date, there is no comparable comprehensive language test of Arabic with adequate psychometric properties. A recent study investigated the validity of the existing language development tests available for Arabic children, the majority of which represent direct translations of English language instruments (Balilah & Archibald, 2018), including the Arabic Expressive-Receptive Vocabulary Test (El-Halees & Wiig, 1999), the Arabic Language Screening Test (El-Halees & Wiig, 1999), as well as the Arabic Picture Vocabulary Test and the Arabic Language Test (Shaalan, 2010). The study found that Arabic language tests were not sensitive to age-related differences across the 6–9 year age range, and that there was a need for tests measuring more complex language skills for older children. Another study pointed to a similar conclusion (Shaalan, 2017) that tests suitable for a broader age range of children are needed.

The ALEF, a comprehensive standardized assessment tool for Arabic-speaking children covering all language domains in expressive and receptive modalities, was constructed with a careful consideration of the unique linguistic features of Gulf Arabic, with test stimuli controlled for the relevant key linguistic and cultural characteristics. As a result, we developed an instrument that is theory-driven, psychometrically-sound, and culturally- sensitive. We used a set of complementary analyses to identify a number of misfitting items, which were then removed to shorten the subtests while preserving each subtest's psychometric fidelity. After the removal of misfitting items, all ALEF subtests demonstrated high reliability estimates, generally convergent between classical test theory and IRT, suggesting that they may be used as reliable tools for indexing spoken and written language abilities in Arabic-speaking children. All of the subtests of the ALEF – Spoken Language (with the exception of Sentence Comprehension) showed expected (both linear and nonlinear) age effects, indicating that the ALEF is a valid test of language development. Furthermore, the positive manifold of correlations among the subtests suggests that the ALEF can be effectively (in a flexible way, as permitted by the assessment's structure) used to evaluate Arabic language development in children as young as three years of age. The resulting shorter length makes the ALEF more practical for clinical use.

The test provides researchers and clinicians with a new tool for the evaluation of language development in Arabic-speaking children whose home language is Gulf Arabic, with a possibility of adapting it to a wide variety of spoken Arabic dialects. Because of its breadth, the ALEF can be used to measure selectively

and in depth a child's performance in all subdomains of language, either expressively or receptively, to evaluate a child's level of literacy acquisition, and/or to get additional indicators measuring cognitive skills known to be pre- (or co-requisite) of language and literacy acquisition (phonological and verbal working memory, rapid automatized naming and phonemic awareness). The subtests can be used in any combination, as age- and developmentally appropriate for each individual child.

The results reported here are too preliminary to recommend the ALEF, in its current state, as a freestanding diagnostic tool. Even after a sufficiently large and representative norming sample for the ALEF is obtained in the future, as with any clinical evaluation, a combination of methods would be required to determine a differential diagnose (including language sampling, a behavioral observation and a caregiver report). However, even in its current form the ALEF represents an important clinical and research assessment tool for the study of language development and impairment in Arabic. For example, an important clinical application of the instrument would include obtaining a comprehensive description of a child's baseline linguistic profile, including language comprehension (in the domains of Syntax, Semantics, and Pragmatics) and production (Phonology, Morphology, Syntax, Semantics and Pragmatics), which can be used to determine areas of strength and weakness, to establish intervention goals, and to track a child's progress towards achieving therapy goals. Another useful feature of the battery is the inclusion of a number of processing-based tasks (such as pseudoword repetition), known to be sensitive clinical markers of language impairment (Conti-Ramsden, 2003). Furthermore, this task was shown to minimize cultural bias in language assessment (Oetting & Cleveland, 2006), an important consideration for Arabic speakers, given a substantial cultural and linguistic diversity in this population. A clinician may adapt a combination of tests (e.g., Expressive vocabulary and Sentence Imitation), in conjunction with the Pseudoword Repetition test as an informal screener (until test development process is completed and normative tables are published) to identify children with language difficulties. Finally, the test can be used with school-age children, whose main weakness main be manifested as difficulty acquiring literacy skills (including decoding, reading comprehension, and spelling).

While this represents an important advance, it is only a stepping-stone in the creation of a fully functional diagnostic tool. Future research is required to conduct further tests of validity, construct age norms (norm tables separate for boys and girls), adapt the test for use in Arabic-speaking populations from other regions, and determine the test's diagnostic accuracy (i.e., sensitivity, specificity, and predictive value). Further research would allow us to create, in addition to individual subtest scores, composite scores addressing separate modalities and/or language subdomains. For example, subtests may be combined to derive composite Index scores covering such areas as Core Language (e.g., combining Receptive Vocabulary, Sentence Completion, and Sentence Comprehension), Expressive Language (e.g., combining Expressive Vocabulary, Sentence Completion, and Sentence Imitation), and Receptive Language (e.g., combining Receptive Vocabulary, Sentence Comprehension, and Inferences from Pragmatics), as performed, for instance, in standardized instruments in English. This will require a new, larger scale data collection than reported here, now made possible by the present pilot stage. Which subtests should be combined in each index will be determined using a data-driven approach. Having index scores, in addition to the individual subtest scores and composite module scores, would allow the administration of the ALEF to be individualized, time-efficient, and flexible.

Further research will be aimed at overcoming several other limitations of the present study. For example, given the large initial pool of items, it was necessary to administer the assessment in two data waves, with no overlap between individuals in waves 1 and 2, thereby prohibiting us from conducting a covariance-based set of procedures aimed at the construct validation of ALEF (i.e., the confirmatory factor analysis of the structure depicted in Figure 1). Because of the size of the battery at the initial stage of test development, it was also not feasible to collect additional data using other assessment measures for evaluating ALEF's convergent and divergent validity. For example, given the aims of the pilot and the length of the battery, we opted not to add to an already large battery an assessment of general cognition. In the future, such an assessment should be added to rule out participants with intellectual disability, and to provide a measure of divergent validity. Our task was also complicated by the absence of an existing "gold standard" for language assessment in Arabic, which, we hope, the ALEF has the potential to become.

We also did not examine differential item functioning (DIF) using ALEF subtest data, as DIF analyses generally require the availability of at least 5 to 10 individuals in the dataset per response category for all items. Although this was not the case in our dataset,

future studies using the revised ALEF battery will directly examine and remove items that show gender (or other, for example, age, family SES, and parental education) bias. Of note is that the gender effects observed in this study were small in magnitude, and the only significant differences after corrections for multiple testing were obtained for Word Articulation and Pragmatic Knowledge, with boys outperforming girls on the former and underperforming on the latter. The presence of gender differences in Pragmatic Knowledge is notable given that it indexes social use of language skills that form diagnostic features for disorders with traditionally higher prevalence indicators for boys such as autism spectrum disorders and social pragmatic communication disorder.

In sum, this article presents the first step in the development and the pilot usage of the battery, which is now ready to be administered to a larger representative sample of Arabic-speaking children in the Kingdom of Saudi Arabia in order to complete the development of this comprehensive and precise clinical diagnostic tool.

As the development of the ALEF continues, we are confident that future standardization and norming efforts (as well as adaptations to closely related spoken Arabic dialects) will result in a polished product that will advance our understanding of (a)typical language development and related neurodevelopmental disorders.

Due to the limitations of this research, the ALEF in its current stage is not intended to serve as a standalone norm-based diagnostic tool. Furthermore, because of the considerable linguistic and cultural diversity of the Arabic-speaking world, much work is needed to understand how or even whether a single instrument can be used across various Arabic speaking populations. However, the existence of this instrument and its first encouraging psychometric properties in the scientific literature should trigger researchers working with Arabic-speaking populations in different settings to utilize it in their studies. Collectively, the field should be able to arrive to a situation where one of the largest language groups in the world can count on the availability of a standardized instrument for the assessment of language competencies in children.

Acknowledgements

This work was supported by the King Faisal University, Saudi Arabia. We are grateful to Dr. Sergey Kornilov for a major contribution to the development of the ALEF Battery and his analyses of the collected data; Dr. Abdullah Alsaadat for his invaluable assistance and support with all administrative issues; Mrs. Nabilah Halal for her contribution to the linguistic stimuli; and Mrs. Natalie Banker for the artwork she created for the assessment. We also thank numerous colleagues from Saudi Arabia who assisted us with data collection and entry, and the children who contributed their time and energy to this work.

Funding

This work was supported by the King Faisal University, Saudi Arabia.

References

Abdalla, F., Aljenaie, K., & Mahfoudhi, A. (2013). Plural noun inflection in Kuwaiti Arabic-speaking children with and without Specific Language Impairment. *Journal of Child Language*, *40*(01), 139–168. doi:10.1017/S0305000912000499

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

Abu-Rabia, S., Share, D., & Mansour, M. S. (2003). Word recognition and basic cognitive processes among reading-disabled and normal readers in Arabic. *Reading and Writing*, *16*(5), 423–442. doi:10.1023/A:1024237415143

Al-Buainain, H., Shain, K., Al-Timimy, F., & Khattab, G. (2012). Baseline data for Arabic acquisition with clinical applications: some phonological processes in Qatari children's speech. *International Journal of Business and Social Research*, *2*(6), 18–33.

Amayreh, M. M. (2003). Completion of the consonant inventory of Arabic. *Journal of Speech, Language, and Hearing Research*, *46*(3), 517–529. Doi 10.1044/1092-4388(2003/042) doi:10.1044/1092-4388(2003/042)

Amayreh, M. M., & Dyson, A. T. (1998). The acquisition of Arabic consonants. *Journal of Speech, Language, and Hearing Research*, *41*(3), 642–653. doi:10.1044/jslhr.4103.642

Amayreh, M. M., & Dyson, A. T. (2000). Phonetic inventories of young Arabic-speaking children. *Clinical Linguistics & Phonetics*, *14*, 193–215.

Aoun, J., Benmamoun, E., & Sportiche, D. (1994). Agreement, word-order, and conjunction in some varieties of Arabic. *Linguistic Inquiry*, *25*(2), 195–220.

Aoun, J., Choueiri, L., & Benmamoun, E. (2010). *The syntax of Arabic*. Cambridge, UK: Cambridge University Press.

Armon-Lotem, S. (2012). Between L2 and SLI: Inflections and prepositions in the Hebrew of bilingual children with TLD and monolingual children with SLI. *Journal of Child Language*, *41*(1), 3–33. doi:10.1017/S0305000912000487

Ayyad, H. S., Bernhardt, B. M., & Stemberger, J. P. (2016). Kuwaiti Arabic: Acquisition of singleton consonants. *International Journal of Language & Communication Disorders*, *51*(5), 531–545. doi:10.1111/1460-6984.12229

Bader, S. (2009). Speech and language impairments of Arabic-speaking Jordanian children within natural phonology and phonology as human behaviour. *Poznań Studies in Contemporary Linguistics*, *45*, 191–210.

Balilah, A., & Archibald, L. (2018). The measurement of language ability and impairment in Arabic-speaking children. In S. Hidri (Ed.), *Revisiting the assessment of second language abilities: From theory to practice* (pp. 65–84). Cham, Switzerland: Springer International Publishing.

Bedore, L. M., & Leonard, L. B. (2001). Grammatical morphology deficits in Spanish-speaking children with Specific Language Impairment. *Journal of Speech, Language, and Hearing Research*, *44*(4), 905–924. doi:10.1044/1092-4388(2001/072)

Ben-Yizhak, N., Yirmiya, N., Seidman, I., Alon, R., Lord, C., & Sigman, M. (2011). Pragmatic language and school related linguistic abilities in siblings of children with autism. *Journal of Autism and Developmental Disorders*, *41*(6), 750–760. doi:10.1007/s10803-010-1096-6

Benmamoun, E. (2000). *The feature structure of functional categories: A comparative study of Arabic dialects*. Oxford, UK: Oxford University Press.

Bentin, S., & Ibrahim, R. (1996). New evidence for phonological processing during visual word recognition: The case of Arabic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(2), 309–323. doi:10.1037//0278-7393.22.2.309

Bishop, D. V. M., Chan, J., Adams, C., Hartley, J., & Weir, F. (2000). Conversational responsiveness in specific language impairment: Evidence of disproportionate pragmatic difficulties in a subset of children. *Development and Psychopathology*, *12*(2), 177–199. doi:10.1017/S0954579400002042

Bloom, L., & Lahey, M. (1978). *Language development and language disorders*. New York: Wiley.

Boudelaa, S., Pulvermuller, F., Hauk, O., Shtyrov, Y., & Marslen-Wilson, W. (2010). Arabic morphology in the neural language system. *Journal of Cognitive Neuroscience*, *22*(5), 998–1010. doi:10.1162/jocn.2009.21273

Boudreau, D. M., & Hedberg, N. L. (1999). A comparison of early literacy skills in children with specific language impairment and their typically developing peers. *American Journal of Speech-Language Pathology*, *8*(3), 249–260. doi:10.1044/1058-0360.0803.249

Brustad, K. (2000). *The syntax of spoken Arabic: A comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. Washington, DC: Georgetown University Press.

Buckwalter, T., & Parkinson, D. (2014). *A frequency dictionary of Arabic*: Core vocabulary for learners. Abington, UK: Routledge.

Cantiani, C., Lorusso, M. L., Guasti, M. T., Sabisch, B., & Mannel, C. (2013). Characterizing the morphosyntactic processing deficit and its relationship to phonology in developmental dyslexia. *Neuropsychologia*, *51*(8), 1595–1607. doi:10.1016/j.neuropsychologia.2013.04.009

Carretti, B., Borella, E., Cornoldi, C., & De Beni, R. (2009). Role of working memory in explainingthe performance of individuals with specific reading comprehension difficulties: A meta-analysis. *Learning and Individual Differences*, *19*(2), 246–251. doi:10.1016/j.lindif.2008.10.002

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*(6), 1–29.

Conti-Ramsden, G. (2003). Processing and linguistic markers in young children with specific language impairment (SLI). *Journal of Speech, Language, and Hearing Research*, *46*(5), 1029–1037. doi:10.1044/1092-4388(2003/082)

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry*, *42*(6), 741–748. doi:10.1111/1469-7610.00770

Denman, D., Speyer, R., Munro, N., Pearce, W. M., Chen, Y.-W., & Cordier, R. (2017). Psychometric properties of language assessments for children aged 4-12 years: A systematic review. *Frontiers in Psychology*, *8*, 1515. doi:10.3389/fpsyg.2017.01515

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. doi:10.1111/bjop.12046

Dyson, A. T., & Amayreh, M. M. (2000). Phonological errors and sound changes in Arabic-speaking children. *Clinical Linguistics & Phonetics*, *14*(2), 79–109.

El-Halees, Y., & Wiig, E. H. (1999). *Arabic receptive-expressive vocabulary test*. Arlington, Texas: Schema Press.

Ferguson, C. A. (1959). Diglossia. *Word*, *15*(2), 325–340. doi:10.1080/00437956.1959.11659702

Frost, S. J., Landi, N., Mencl, W. E., Sandak, R., Fulbright, R. K., Tejada, E. T., … Pugh, K. R. (2009). Phonological awareness predicts activation patterns for print and speech. *Annals of Dyslexia*, *59*(1), 78–97. doi:10.1007/s11881-009-0024-y

Guasti, M. T. (2002). *Language acquisition: The growth of grammar*. Cambridgde, MA: MIT Press.

Habash, N. (2007). Arabic morphological representations for machine translation. In A. Soudi, A. van den Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and empirical methods* (Vol. 38, pp. 263–285). Netherlands: Springer.

Hansson, K., & Nettelbladt, U. (1995). Grammatical characteristics of Swedish children with SLI. *Journal of Speech, Language, and Hearing Research*, *38*(3), 589–598. doi:10.1044/jshr.3803.589

Hole, C. (1995). Modern Arabic: Structures, functional and varieties. New York, USA: Longman Linguistic Library.

Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties*. Washington, DC: Georgetown University Press.

Ibrahim, R. (2009). The cognitive basis of diglossia in arabic: Evidence from a repetition priming study within and between languages. *Psychology Research and Behavior Management*, *2*, 93–105.

Ibrahim, R. (2011). Literacy problems in Arabic: Sensitivity to diglossia in tasks involving working memory. *Journal of Neurolinguistics*, *24*(5), 571–582. doi:10.1016/j.jneuroling.2010.10.003

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122. doi:10.1037//0033-295X.99.1.122

Khodeir, M. S., Hegazi, M. A., & Saleh, M. M. (2017). Development and standardization of a test for pragmatic language skills in Egyptian Arabic: The Egyptian Arabic Pragmatic Language Test (EAPLT). *Folia Phoniatrica et Logopaedica*, *69*(5–6), 209–218. doi:10.1159/000485656

Kovelman, I., Norton, E. S., Christodoulou, J. A., Gaab, N., Lieberman, D. A., Triantafyllou, C., … Gabrieli, J. D. E. (2012). Brain basis of phonological awareness for spoken

language in children and its disruption in dyslexia. *Cerebral Cortex*, *22*(4), 754–764. doi:10.1093/cercor/bhr094

Kroll, J. F., & de Groot, A. M. B. (1997). Lexical and conceptual memory in the bilingual: Mapping form to meaning in two languages. In J. F. Kroll & A. M. B. d. Groot (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives* (pp. 169–199). Mahwah, NJ: Erlbaum.

Levy, Y., & Schaeffer, J. (Eds.). (2011). *Language development across populations: Towards a definition of specific language impairment*. Mahwah, NJ: Lawrence Erlbaum.

Lust, B., Chien, Y.-C., & Fynn, S. (1987). What children know: Comparison of experimental methods for the study of first language acquisition. In B. Lust (Ed.), *Studies in the acquisition of anaphora, vol. 2: Applying the constraints* (pp. 271–356). Dordrecht, Netherlands: Reidel.

Maamouri, M. (1998). Language education and human development: Arabic diglossia and its impact on the quality of education in the Arab region. Paper presented at the The World Bank Mediterranean Development Forum, Marrakesh, Morocco.

Mahfoudhi, A., & Abdalla, F. (2017). Language development and impairment in Arabic-speaking children: Introduction to the special issue. *Arab Journal of Applied Linguistics*, *2*(1), 1–11.

Montgomery, J. W., & Evans, J. L. (2009). Complex sentence comprehension and working memory in children with specific Llanguage impairment. *Journal of Speech, Language, and Hearing Research*, *52*(2), 269–288. doi:10.1044/1092-4388(2008/07-0116)

Norbury, C. F. (2014). Practitioner Review: Social (pragmatic) communication disorder conceptualization, evidence and clinical implications. *Journal of Child Psychology and Psychiatry*, *55*(3), 204–216. doi:10.1111/jcpp.12154

Norbury, C. F., & Bishop, D. V. (2003). Narrative skills of children with communication impairments. *International Journal of Language & Communication Disorders*, *38*(3), 287–313. doi:10.1080/136820310000108133

Oetting, J. B., & Cleveland, L. H. (2006). The clinical utility of nonword repetition for children living in the rural south of the US. *Clinical Linguistics & Phonetics*, *20*(7–8), 553–561. doi:10.1080/02699200500266455

Omar, M. (1973). *The acquisition of Egyptian Arabic as a native language*. Washington, DC: Georgetown University Press.

Owens, J., Dodsworth, R., & Rockwood, T. (2009). Subject-verb order in spoken Arabic: Morpholexical and event-based factors. *Language Variation and Change*, *21*(01), 39–67. doi:10.1017/S0954394509000027

Paul, R., & Norbury, C. (2012). *Language disorders from infancy through adolescence-E-Book: Listening, speaking, reading, writing, and communicating*. Amsterdam, Netherlands: Elsevier Health Sciences.

Reisinger, L. M., Cornish, K. M., & Fombonne, E. (2011). Diagnostic differentiation of autism spectrum disorders and pragmatic language impairment. *Journal of Autism and Developmental Disorders*, *41*(12), 1694–1704. doi:10.1007/s10803-011-1196-y

Rice, M. L., Tomblin, J. B., Hoffman, L., Richman, W. A., & Marquis, J. (2004). Grammatical tense deficits in children with SLI and nonspecific language impairment: Relationships with nonverbal IQ over time. *Journal of Speech, Language, and Hearing Research*, *47*(4), 816–834. doi:10.1044/1092-4388(2004/061)

Rispens, J., Roeleven, S., & Koster, C. (2004). Sensitivity to subject–verb agreement in spoken language in children with developmental dyslexia. *Journal of Neurolinguistics*, *17*(5), 333–347. doi:10.1016/j.jneuroling.2003.09.001

Shaalan, S. (2010). *Investigating grammatical complexity in Gulf Arabic speaking children with specific language impairment (SLI)*. London, UK: UCL (University College London).

Shaalan, S. (2017). 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.

Shawarbah, M. (2007). *The Bedouin Dialect of the Tiyaha in the Negev: Phonology, morphology and some syntactic issues. (Ph. D.)*. Beersheba, Israel: Ben-Gurion University.

Shlonsky, U. (1997). *Clause structure and word order in Hebrew and Arabic: An essay in comparative Semitic syntax* (Vol. 11). Oxford, UK: Oxford University Press.

Sholl, A., Sankaranarayanan, A., & Kroll, J. F. (1995). Transfer between picture naming and translation: A test of asymmetries in bilingual memory. *Psychological Science*, *6*(1), 45–49. doi:10.1111/j.1467-9280.1995.tb00303.x

Simons, G. F., & Fennig, C. D. (Eds.). (2018). *Ethnologue: Languages of the World* (21st ed.). Dallas, Texas: SIL International.

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry*, *45*(1), 2–40. doi:10.1046/j.0021-9630.2003.00305.x

Versteegh, K. (1996). Linguisitic attitudes and the origin of speech in the Arab world. In A. Elgibali (Ed.), *Understanding Arabic*. Cairo, Egypt: American University in Cairo.

Wagner, R., Torgesen, J., Rashotte, C., Pearson, N. A., Wiig, E., Secord, W., … Denckla, M. B. (1999). *CTOPP-2: Comprehensive test of phonological processing–Second edition*. Austin, TX: Pro-ed, 1999.

Watson, J. (2002). *The phonology and morphology of Arabic*. Oxford, UK: Oxford University Press.

Wolf, M., Bowers, P. G., & Biddle, K. (2000). Naming-speed processes, timing, and reading: A conceptual review. *Journal of Learning Disabilities*, *33*(4), 387–407. doi:10.1177/002221940003300409

Young, E. C., Diehl, J. J., Morris, D., Hyman, S. L., & Bennetto, L. (2005). The use of two language tests to identify pragmatic language problems in children with autism spectrum disorders. *Language, Speech, and Hearing Services in Schools*, *36*(1), 62–72. doi:10.1044/0161-1461(2005/006)

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., … Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science*, *21*(4), 551–559. doi:10.1177/0956797610363406

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131*(1), 3–29. doi:10.1037/0033-2909.131.1.3