Running Head: Arabic CDI

Adaptation and Standardization of two Arabic Communicative Development Inventories for Children Aged 8 to 30 Months and 30 to 48 Months in Egypt, Jordan and Palestine

Alshaimaa Abdelwahab[1,2*]

Caroline Floccia[1*]

Samuel Forbes[3]

Zakiyah Alsiddiqi[1]

Khalid Al-Shdifat[4]

Cristina McKean[5]

Thair Odeh[6]

Anastasia Trebacz[5,7]

and

Ghada Khattab[5]

1. School of Psychology, University of Plymouth, Plymouth, UK; (2) Department of English Language and Literature, Damietta University, Damietta, Egypt; (3) Department of Psychology, Durham University, Durham, UK; (4) Department of Rehabilitation Sciences, Jordan University of Science and Technology, Irbid, Jordan; (5) Speech and Language Sciences, University of Newcastle, Newcastle upon Tyne, UK; (6) Department of Audiology and Speech Therapy, Birzeit University, Birzeit, Palestine (7) School of Medicine, University of Sunderland, UK

*These two authors are designated as co-first authors, ordered alphabetically

Corresponding author : Caroline Floccia, School of Psychology, University of Plymouth, Drake Circus, PL4 8AA Plymouth, UK, caroline.floccia@plymouth.ac.uk, tel +44 (0)1752584822

Conflict of interest: None

Abstract

Purpose: Assessing early language skills through parental report is a cost-effective way to screen for language delays when resources are scarce. A pan-Arabic lexeme approach was tested by extending the Egyptian adaptation (Abdelwahab et al., 2021) of the MacArthur-Bates Communicative Development Inventory (CDI; Fenson et al., 1994) to Jordan and Palestine, in infants aged 8 to 30 months (Arabic CDI Toddler). We also developed a version for children aged 30 to 48 months (Arabic CDI Child) in the three countries. A fair representation of the population in terms of geographical spread, maternal education and environmental context was obtained through stratified sampling.

Methods: The Arabic CDI Toddler, containing a 100-word list with dialect variations from the three countries, was completed online or face-to-face by 427 Egyptian (including 259 from Abdelwahab et al., 2021), 455 Jordanian and 440 Palestinian mothers. The Arabic CDI Child, comprising a 100-word list, 23 simple/complex pairs of sentences and 11 language use questions, was completed by 836 Egyptian parents, 318 Jordanians and 304 Palestinians. Non-Governmental Organizations (NGOs) facilitated contact with harder-to-reach families due to illiteracy or limited internet access.

Results: Child gender and maternal education effects were found for the Arabic CDI Toddler in Jordan and Palestine, while maternal education effects were generally found across the three countries and all sections of the Arabic CDI Child. For the Arabic CDI Toddler, norms were computed for Jordan and Palestine, separately from Egypt. For the Arabic CDI Child, separate norms were provided for each country.

Conclusion: The production of a unique set of norms was contained to two countries for 8-30-month-old children, highlighting the limits of a pan-Arabic lexeme approach to the development of parental reports. The Arabic CDI Toddler and CDI Child have the potential to

be useful to parents, Early Years professionals and researchers who need to assess children's Arabic knowledge for detecting those at risk of poor language outcomes and in need of intervention.

Keywords: language development; Arabic; parental report

Adaptation and Standardization of two Arabic Communicative Development Inventories for Children Aged 8 to 30 Months and 30 to 48 Months in Egypt, Jordan and Palestine

Children's language development is the foundational skill that fosters learning, emotional and behavioral regulation (Petersen & LeBeau, 2021), social integration (Davis & Qi, 2020), and future academic achievements (Bleses et al., 2016), especially in relation to readiness for reading and writing (Bruce & Spratt, 2011). Children learn language at differing rates. However, a significant minority (~ 9%) persistently lag behind their peers, with significant consequences for their long-term outcomes especially if they do not receive additional support. Early identification of such difficulties enables the provision of early preventative interventions, which bring the greatest gains for the child (e. g. Law et al., 2017). However, the necessary tools to identify children with language difficulties in the early years are not universally available in all languages spoken around the world. This study aimed to fill a key gap through the development of easy to use, accessible language assessment tools for infants and toddlers aged 8-30 months, and for children aged 30 to 48 months learning Arabic. These tools are presented in the shape of an online app that parents and/or Early Years professionals can sign up to fill in. It is hoped that they will enable the identification of individual children at risk of poor language outcomes and in need of intervention, and also inform the development and delivery of early education and health services and policy. That is, the assessment of child language across a population could inform the fair distribution of resources to support language development and equitable outcomes (McKean & Reilly, 2023).

Arabic is spoken by more than 260 million people in the world, making it the 5th most spoken language (Kaye, 2018). The Arabic language encompasses a wide range of dialects

spoken within countries or within areas of the same country with at least 30 to 35 different Arabic dialects recognized around the world (Versteegh, 2014). Arabic dialects significantly vary from one another, to the point that they may be mutually incomprehensible, particularly when comparing Western (or Maghrib) and Eastern (or Mashriq) dialects. The study of Arabic dialects has been historically challenging due to the lack of official orthographies and an initial focus on the study of Modern Standard Arabic, the variety that is used in educational settings and media. This has hindered work on both the adult varieties that children are exposed to as well as tools to assess children's language, which need to be based on their home dialect given that this is their primary source of linguistic exposure in the pre-school years.

Given the number of Arabic dialects and the level of resources required to develop new assessment tools, this study aimed to determine whether a pan-Arabic lexeme approach to language testing might be possible. This was done by adapting the same tools to the dialect of three Arabic-speaking countries (Egypt, Jordan and Palestine), and examining whether the data from the three countries could be collated to produce a single set of norms based around shared lexemes. This would establish a proof of concept that a similar approach may be used for other dialects of Arabic: despite local lexical variations, the trajectory of vocabulary development for a unique set of lexemes might be broadly similar across these countries. This idea was already outlined in the development of the Egyptian Short Form of the Arabic Communicative Development Inventory or CDI (Abdelwahab et al., 2021; Fenson et al., 1994, 2007), a parental questionnaire about children’s vocabulary knowledge. In the 2021 paper, the data of 436 children aged 8-30 months were collected to produce norms of vocabulary development in Egyptian Arabic. In addition, data were collected regarding lexeme knowledge from 16 other Arabic countries

(between 8 and 22 children per country, including Jordan and Palestine). It was tentatively concluded that the Egyptian norms could in principle be used for children from the other countries. However, this was based on a very small amount of data from the 16 non-Egyptian countries; in addition, the sampling of the Egyptian data was nearly entirely based on highly educated parents, from one single area of Egypt (urban Lower Egypt). Fenson et al. (2007) argue that CDIs that rely on small, non-representative samples may overestimate or underestimate the language abilities of children. They suggest that research should use larger, more diverse norming samples to improve the accuracy and reliability of CDIs, and to better understand the developmental variability in early language development. Additionally, Kidd and Garcia (2022) and Figueroa (2024) argue for the necessity to broaden the range of languages represented in the field of language development, where Indo-European languages and a handful of non-Indo-European languages (not including Arabic) are largely overrepresented. Finally, if such tools are to be used for valid clinical and policy decision-making then the removal of such sampling biases is essential. Therefore, our aim was to collect data representing the population diversity in proportions compatible with census data, in terms of levels of maternal education, geographical location, and environmental context, that is, urban versus rural, to reflect the discrepancies in resource availability to families.

Arabic dialects of Egypt, Jordan, and Palestine

The origin of the Arabic language in Egypt is linked to the arrival of Bedouin tribes from the Arabian Peninsula at the beginning of the Islamic conquest, impacting Lower Egypt first and then gradually spreading through Upper Egypt (Versteegh, 2014), but modern-day Lower Egypt varieties (including Cairo and the Delta) are mainly urban. Modern Cairene has influence across the country and beyond due to its prestige and spread through mass

media. Phonological features that distinguish Egyptian Arabic dialects from other dialects groups include the realization of (dʒ) as [g] (e.g. /d͡ʒabal/ *mountain* as [gabal]) in much of Cairene and the Central Delta, the shortening of long vowels in unstressed positions (e.g. [joːm] *day* vs [joˈmen] *two days*), the -sh suffix for negation ([maʕraf-ʃ] *I don't know*), and breaking the second consonant of CCC clusters from the third (e.g. /jiktbu/ → [jiktibu] *they write*). Other Egyptian dialect features include lexical items (e.g. [jiddi] *he gives*; [gallaːbijja] *garment*; [dilwaʔt] *right now*; and [kida] *this way*), and the placement of demonstratives and interrogatives in post-position, e.g. [il-walad da] *boy this* (this boy); [katab ʔeː] *wrote*-3rd pers. sing. *what* (What did he write?) (Bishai, 1964; Versteegh, 2014).

The origin of the Arabic Language in the Syro-Lebanese region, which includes Jordan and Palestine, pre-dates the Islamic conquest due to the presence of Arabic-speaking tribes in the region. The new Arabic varieties brought by the Arab conquerors (Versteegh, 2014) were more urban by nature, but modern-day Jordanian and Palestinian are still more influenced by Bedouin than urban varieties compared with Lebanese and South Syrian. This is reflected in all aspects of the grammar including the phonology, where most Jordanian and Palestinian varieties have the Bedouin [g] for (q), though with urban [ʔ] is gaining ground. In contrast with Egyptian, Jordanian and Palestinian break the first consonant from the second (/jiktbu/ → [jikitbu] *they write*) and use the prefix ba- for the imperfect bi/bji as in [ba-ktub] *I write* (Cotter & de Jong, 2019; Owens, 2013).

Existing language tests in Egypt, Jordan and Palestine

The availability of culturally adapted norms for early language assessment varies across countries, with Egyptian practitioners having access to a broader range of assessment tools than Jordan and Palestine (see Table 1). In Jordan, assessment tools tend to be adapted directly from Western tests and target not only language, but also other domains

such as cognition and socio-emotional development. In Palestine, there is a general lack of standardized tests for assessing language development in young children. In this context, a standardized parental report like the Arabic CDI will possibly be welcomed by a range of Early Years professionals.

Table 1 here

Furthermore, in the three countries of interest, language skills are not routinely assessed in early childhood through public healthcare. Language assessments only take place if children attend specialist assessment such as with a speech pathologist or pediatrician. The provision of an Arabic adaptation of the free to use and easily accessible CDI should enable universal access to language assessment. In this way, we aim to support families without access to private healthcare to evaluate their child's language development, as well as provide policy-makers and services with means to understand the level of need in their populations. Ultimately, we anticipate that such knowledge can lead to increased access to support and intervention for children, enhancing their long-term life chances and wellbeing (Law et al., 2017).

The multiple forms and uses of the CDI

CDIs are the most popular parental report instruments for assessing children's early language acquisition, as they have been adapted in more than 100 languages to reflect the culture and structure of these languages (https://mb-cdi.stanford.edu/adaptations.html). Parental reports have proven to be trustworthy sources in relation to the information they provide about children's communicative abilities given that parents are close observers of their children. They are also an affordable method for evaluating a child's language abilities,

particularly in this critical period before the child becomes able to engage in face-to-face activities with strangers (Kristoffersen et al., 2013). The exact content of published CDIs varies depending on the age range and the version, but their core feature is a list of familiar words that parents indicate to be understood and/or produced by their child based on their own knowledge about their child's vocabulary. These instruments can be administered in several ways such as an interview with the parents, on paper, or online (Kristoffersen et al., 2013). There are three forms of the original American English MacArthur Bates CDI: 1) the CDI-Words and Gestures (Fenson et al., 1994) for children up to 18 months which measures early gestures, comprehension, and production through a checklist of 63 gestures and 396 words; 2) the CDI-Words and Sentences for children between 16 to 30 months which assesses expressive vocabulary through a list of 680 words, and includes sections assessing morphology and syntax (Fenson et al., 1994); 3) the CDI-III (Fenson et al., 2007) designed for children aged 36 to 48 months which assesses expressive vocabulary through a checklist of 100 words and includes sections about advanced grammar, semantic knowledge and comprehension. Multiple studies have developed short forms of the CDIs (primarily forms 1 and 2) to save time and resources, and because they have proven to have comparable statistical properties (e.g., in Galician: Pérez-Pereira & Resches, 2007; American English: Fenson et al., 2000; British English: Floccia et al., 2018; Finnish: Vehkavuori & Stolt, 2018; German: Mayor & Mani, 2019; Egyptian Arabic: Abdelwahab et al., 2021).

In terms of research, CDI data are useful for example for exploring the variation in early word production (Fenson et al., 1994), the varied relationship between lexical and grammatical development (Bates & Goodman, 2001), and the variation in the early composition of vocabulary (Bates et al., 1994). CDIs have also been used for screening and identifying children for diagnostic evaluations and interventions (e.g. Kim et al., 2016),

although a recent review by Eriksson (2022) has argued that there is insufficient evidence to date that the CDIs are valid screening tools for language difficulties. From their review of 9 studies having used various forms or adaptations of the CDI to identify children with language difficulties, sample sizes were low, biases frequent, cut-off values often decided based on the sample, and resulting diagnostic accuracy low. This does not necessarily invalidate the use of the CDIs in the identification of children at risk but rather urges caution in their use as "screeners" suggesting such tools can be usefully embedded within more comprehensive systems of developmental surveillance and support (McKean & Reily 2023; Reily & McKean 2023). Embedding vocabulary checklists within a staged or triaged approach to identification for example can enable the benefits of such tools to be exploited whilst managing their limitations (Wilson et al., 2022). Furthermore, when the choice in a population is between an imperfect first assessment tool and no tool at all for the majority of the public, then the CDI becomes a pragmatic solution that cannot be disregarded in the absence of any reasonable alternative. It is clearly vital that assessment tools are culturally and linguistically valid and there are risks in adapting a list of vocabulary items rather than starting with a "blank slate" of bringing cultural and linguistic biases to the vocabulary chosen. However, there is a pressing need for rapid development of tools across the many under-represented languages in the world with very limited normative language acquisition data from which to begin the process of assessment development. This is coupled with limited resource, especially within Low and Middle-Income Countries where other priorities for children's health and development often take precedence. Adaptation, if conducted with meaningful and detailed engagement with key stakeholders can be a pragmatic method for accelerating progress toward the development of valid tools and interventions (Moore et al., 2021; Ohana & Armon-Lotem, 2023). In this study extensive work was conducted to

maximize the cultural-linguistic validity of the adaptation and we are confident that valid conclusions can be drawn abut children's language abilities. Also, we posit that such tools can be the first steps in gathering the necessary large-scale data to begin the development of valid characterization of individual differences in specific cultural-linguistic contexts.

In this paper we describe the development of two tools: (1) the Arabic CDI Toddler (8-30 months-of-age), a short form of the MacArthur CDI Words and Gestures (Fenson et al., 1994), addressing words only, and (2) the Arabic CDI Child (30 – 48 months-of-age) which is the equivalent of the CDI-III (Fenson et al., 2007). The Arabic CDI Toddler should technically be called Short Form Arabic CDI Word Only (Abdelwahab et al., 2021) but for concision we have chosen to call it Arabic CDI Toddler in this manuscript. Both are parental reports of the child's language knowledge.

Aims of the study

The first aim of this study was to develop and standardize two accessible language assessment tools for infants and toddlers aged 8 to 30 months and children aged 30 to 48 months who are learning Arabic, while considering the diversity of Arabic dialects. Specifically, the Arabic CDI Toddler and the Arabic CDI Child were adapted to the dialects of Egypt, Jordan and Palestine. We examined whether data collected from each country could be combined to produce a unique set of norms, which would serve as a proof of concept for a similar pan-Arabic lexeme approach for other dialects of Arabic. The Arabic CDI Toddler data in Egypt augmented the data already collected in Abdelwahab et al. (2021) with a more representative sample, given that the 2021 norms for the Egyptian Arabic CDI Toddler were based on data from families from advantaged socio-economic circumstances (as measured by maternal education) located in urban areas of Lower Egypt.

The second aim was to assess the role of key factors such as gender and socio-economic circumstances as indexed by maternal education and environmental context (urban versus rural) on early language development. This was done not only to establish whether the current data align with what is typically found in other CDI standardizations, as will be discussed below, but also to examine whether these factors have a comparable effect in Middle-Eastern and Western populations.

Effect of child gender on language development

Several studies using CDIs have indicated an advantage for girls as compared to boys for various aspects of language development, especially in the early stages of lexical development (e.g., Bornstein et al., 2004; Fenson et al., 2007; Floccia et al., 2018). Typically, on average girls produce their first gestures and combine words earlier than boys (Eriksson et al., 2012; Özçalışkan & Goldin-Meadow, 2010), with data from more than 3000 CDIs on Wordbank (Frank et al., 2017) showing that American English-learning 2-year-old girls produce up to 100 words more than boys at the same age.

On the other hand, some studies found only a slight difference between boys and girls in their vocabulary acquisition (Fenson et al., 1994), particularly between 8 and 30 months where gender accounted for only 1-2% of variance in language ability. Andersson et al. (2011) reported a small vocabulary size difference between 4-24 month-old Swedish boys and girls, which only appeared by the end of the second year of their life. Similarly, in a meta-analysis of Slovenian data, using various tools including the CDIs, Marjanovič-Umek and Fekonja-Peklaj (2017) found that the effect of gender on production vocabulary size increased with age but showed a slight to null effect between 8 to 30 months.

However, all the studies mentioned so far have examined Western populations. In the Middle East, Abdelwahab et al. (2021) found no effect of gender on Egyptian Arabic-

speaking children's vocabulary size with the Arabic CDI Toddler for children aged 8-30 months, which they attributed to the low age range of the sample and, perhaps, the privileged socio-economic circumstances (SEC) background of the families. Similarly, Musa et al. (2021) investigated the effect of gender on the language skills of Arabic Speaking Sudanese Children aged between 2 and 8 years, using the Modified Preschool Language Scale-4 (Abu Haseeba et al., 2011) and the Sudanese Arabic Speaking Semantic Skills Test (SASST). No significant differences in performance were found between girls and boys. If confirmed, this suggests that gender differences in language skills between the East and the West could be due to the different ways that boys and girls are encouraged to use language in the home environment and in the community. The current study will re-evaluate the absence of gender difference in Abdelwahab et al. (2021) with a larger, more inclusive sample.

Effect of socio-economic circumstances (SECs)

In representative samples internationally it has been demonstrated that early language skills follow the social gradient. Children with more disadvantaged SECs have on average poorer language skills than their counterparts living in more advantaged SECs and this effect is a gradient one, with each decrement in SEC associated with lower language levels (McKean et al., 2018; Reilly et al., 2014; Wareham et al., 2021). This gradient has been shown to exist as early as 18 months of age (Brushe et al., 2021). This is generally attributed to families from more disadvantaged background having more difficulties in providing language enriching environments due to limitations of time, physical resources and the effects of stress and other pressures (e.g. Eadie et al., 2022; Hoff, 2003; Huttenlocher et al., 2010; Wareham et al., 2021). In the literature, studies have commonly used a combination of parents' educational attainment, occupational status, and income as measures of socio-

economic circumstances (Bradley & Corwyn, 2002). Hoff (2003) found that more highly educated mothers, specifically those with college education and higher status jobs, engaged in more child-directed speech and used more complex language with their 24-month-old children than mothers with more disadvantaged SECs. However, when CDIs are used to measure language skills, effects of SEC are not systematically observed, sometimes because the sampling does not always encompass sufficient variability in SEC or because there can be some over-reporting in families from disadvantaged SECs (Feldman et al., 2000; Fenson et al. 1993; Reese & Read, 2000). For example, using the Arabic CDI, Abdelwahab et al. (2021) found no impact of SECs as indexed by maternal education, which could have been expected given that within the self-selected sample of mothers 85% were university educated. Similarly, Hamilton et al. (2000) found no impact of SECs on comprehension or production of British English children when collecting the Oxford CDI data in the UK with parents from mostly middle-class backgrounds. This was also documented by Fenson et al. (1994) for the first standardization of the MacArthur CDI in the US, as 77% of the parents had reached a level of university education. No impact of SECs on vocabulary skills was reported before 3 years of age, with a small effect after that age. These studies demonstrate that without due attention to sample characteristics, and making specific efforts to reach families less likely to self-select, collecting CDI data tends to lead to samples of highly educated parents. In the current study, we aimed to recruit a proportion of mothers whose education fairly represents the distribution of education levels in each country, and investigate the effect of socio-economic circumstances (with maternal education as a proxy) on early language skills.

Another factor presumably related to SEC is the broad distinction in environmental context between urban and rural environments. Increased opportunities for urban children

to interact with others and greater access to diverse social and recreational activities in cities would appear to lead to the hypothesis that urban children learn words faster than rural children. This is indeed what was reported by Vogt et al. (2013) when comparing CDI scores for Mozambican children raised in urban versus rural communities. However, the small sample size (N=14) and the possible confound with maternal education makes it difficult to generalize. What is clearer is that the amount of speech directed to children in urban environments is generally higher than in rural environments for countries outside of North America, as shown by Cristia's (2023) systematic review of infant-directed speech research across the world. If this observation also applies to Middle Eastern countries, it is unknown whether this would translate into lower scores for children living in rural rather than urban contexts in Egypt, Jordan, and Palestine, above and beyond any effects explained by maternal education.

In summary, starting from the Arabic CDI Toddler developed and normed for Egypt by Abdelwahab et al. (2021), we will adopt a pan-Arabic lexeme approach by comparing vocabulary data from 8-30-month-olds in Jordan, Palestine, and Egypt, and examine whether a single set of norms for lexemes can be produced (Study 1). In Study 2, we will develop and standardize the Arabic CDI Child in these three countries, for children aged 30 to 48 months. In the two studies and in each country, we will attempt to fully represent the population in terms of geographic spread, maternal education, and environmental context. We will also examine how factors known to affect early language skills in Western societies such as SECs and child gender modulate language development in this region of the world.

Research questions

1/ In developing and standardizing language assessment tools for 8-30-month-old Arabic-learning toddlers (Study 1) and 30-48-month-old Arabic-speaking children (Study 2),

can a pan-Arabic approach produce a unique set of norms for the three countries (Egypt, Jordan and Palestine)?

2/ How do child gender and socio-economic circumstances contribute to emerging language skills in these countries?

Study 1 – Arabic CDI Toddler

Methods

In this first study, building on the Arabic CDI standardized for Egypt by Abdelwahab et al. (2021), we refined its adaptation to Jordan and Palestine and collected data in the three countries from parents of children aged 8 to 30 months. Data were collected online through a dedicated app or face to face to reach out to families from disadvantaged SECs, with the help of NGOs when necessary.

CDI adaptation

Abdelwahab et al. (2021) describes the procedure of selecting the 100 lexical items comprising the Arabic CDI Toddler and adapting it to 17 different Arabic dialects. In the current study, the existing lists of the Jordanian and Palestinian CDIs were sent to three native speakers in each country to check whether any other lexical varieties could be used for the same words. In the Jordanian form, lexical varieties for the same lexemes were added to 12 words (bib, wall, medicine, money, slide, grandpa, nap, shush, please, walk, hot, none). For the Palestinian list, lexical varieties were added to 15 words (monkey, mouse, spider, balloon, ice cream, nappy, glasses, pyjamas, nose, broom, slide, grandpa, nap, old, how). The resulting lists were implemented on the testing platform, and this was piloted by sending the link to 20 families in Palestine and Jordan to test for its user-friendliness and clarity, including the accuracy of the list of words. This was not done for Egypt due to the team's previous experience with using a similar platform in this country

(Abdelwahab et al., 2021). No further changes were suggested during this phase. Validity and reliability of the Arabic CDI were established in Abdelwahab et al. (2021) as follows. In addition to face validity, concurrent validity was assessed by administering the CDI together with the Arabic Language Test (Rifaie et al., 2021; Rifaie & Hassan, 2004; Rifaie et al., 1994) to 23 participants. Reliability was examined through test-retest with 21 participants. Because there was no change to the content of the 2021 test, apart from adding lexical variants and moving to a slightly different platform, we did not re-assess validity or reliability in the current study (the Arabic CDI Toddler can be found in the Supplemental material). Parents could access the platform by using this link: https://www.psy.plymouth.ac.uk/ArCDI/Toddler/

Participants

The data from a total of 3084 children were collected between 2020 and 2022, of which 1778 were discarded for the reasons described in the participant flow chart in Figure 1, leaving 1306 usable data. The target age range was 7.5 to 30.4 months. Parents were asked on the app which dialect they spoke themselves, and after data collection, data were excluded if the child was exposed to a language other than their country's Arabic dialect for more than 10% of the time; note that in Jordan, both Jordanian and Palestinian dialects were considered as the country's dialect. This was done as a practical approach to represent the majority of children in each country; a full representation of all dialects spoken in each country would be beyond the scope of the current study. This criterion also excluded children exposed to another language than Arabic more than 10% of the time (although it did not necessarily exclude children exposed to other languages in preschool settings, which is a limitation to this study). This decision was based on the recurrent findings that vocabulary size in multilingual children in each language is generally smaller than that of

monolingual children, and varies as a function of exposure (e.g. Cattani et al., 2014; Floccia et al., 2018; Thordardottir, 2011). To obtain reliable norms for multilingual children was beyond the scope of the current study. The majority of these data were collected online using a variety of social media specific to each country, dynamically adjusting the focus onto different geographical areas. In order to achieve a representative sample of the population, about a third were collected by local research assistants in face-to-face interviews (152 in Jordan, 165 in Egypt and 91 in Palestine), most of whom were living in rural areas, and/or with parents who were either illiterate, had primary education, and/or were unable to access the online website themselves, or without internet connection enabling them to sign up. In this way those families not reached through social media or unable to fill in the CDI unsupported were represented in the sample.

Insert Figure 1 here

Stratification

Census data (For Egypt: Central Agency for Public Mobilization and Statistics 'CAPMAS', 2020, retrieved [12 November 2022], https://www.capmas.gov.eg/ ; For Jordan: National Centre for Statistics and Information for Jordan, 2021, retrieved [15 November 2022], https://dosweb.dos.gov.jo/; For Palestine: Palestinian Central Bureau of Statistics, 2021, retrieved [20 November 2022], https://www.pcbs.gov.ps/site/lang__en/1/default.aspx, 2021) were used to create a representative sampling frame with regards to region, environment (urban/rural) and maternal education. It must be noted that each country may have a different definition of these categories and they were not always defined in each census. Details of the census

calculations are found in Supplemental material. Table 2 shows the final distribution of the proportion of women needed in urban versus rural areas in each country to obtain a representative sample of the population in terms of education level and population density. In Jordan, we aimed at recruiting at least 10% of families from rural areas, predominantly found in Karak, Ma'an and Mafraq. In Palestine, the census data reported no rural area in Gaza for historical reasons, so we could not set up a target in terms of rural/urban context in that area. We aimed at recruiting mainly families in urban areas in Gaza, and 75% in the West Bank (mainly in Jenin, Nablus, and Ramallah & Albireh). In Egypt, our target was twice as many parents from Lower Egypt than Upper Egypt, with an equal proportion of families from rural than urban areas.

The initial target sample sizes were 400 children in Jordan, 400 in Palestine and 200 in Egypt to complement the existing 436 from Abdelwahab et al. (2021). A sample size of 200 Egyptian children was needed for modelling an effect of ~0.04 (the smallest effect found in the 2021 Egyptian data) with a 0.80 power (Berndtson, 1991). As for the Jordan and Palestine targets, typical CDI standardizations collect between 20 and 50 datapoints per one-month age bin (e.g. 22 in Slovenianet al., 2013; 33 in Estonian, Urm & Tulviste, 2021; 46 in Mandarin Chinese, Hao et al., 2008). Pooling the data from Egypt, Jordan and Palestine together (N=1436) would result in 60 children per bin, which is in the top range of what is usually found in the literature in terms of sample size (see Fenson et al., 2000, for disputing the necessity of very large sample sizes).

Insert Table 2 here

Procedure

Ethical approval for this project (both studies) was granted by the Faculty of Health Ethics Committee at the University of Plymouth (reference 2020-2200-1182), as well as by the Humanities, Arts and Social Sciences Ethical Review Committee at the University of Newcastle (21/05/2020), the Institutional Review Board at Jordan University of Science and Technology (reference 27/133/2020), and the Scientific Research Committee at Birzeit University, Palestine (reference BZUPNH2001).

A bespoke platform created ty the Technical Team of the School of Psychology at the University of Plymouth was used for data collection, very similar to that used in Abdelwahab et al. (2021). After reading and signing a consent form for the use of their data, parents were asked to provide demographics including their country of residence, their governorate and environment (rural or urban), their spoken dialect, and their highest qualification and occupation (for each parent when applicable). They are also asked to indicate their child's date of birth and gender, country of birth and residence, the dialect the child was most familiar with, any other language than Arabic spoken at home or outside and to what extent through indicating the percentage of exposure to any additional dialect(s) or language(s) according to a scale. Finally, parents were asked to indicate whether their child had any diagnosed developmental delay, hearing problems, and/or prematurity. This was followed by the CDI questionnaire where the 100 words were presented one by one; parents had to indicate for each one whether their child produced it, or understood it without producing it yet, or none of the above.

Research assistants (RAs) in each country recruited participants via online and face-to-face methods. Paid social media posts and paper leaflets containing a QR code link to the online platform were used to target potential participants. Nurseries that agreed to support recruitment shared the details of the study with parents. NGOs facilitated contact

with families that were harder to reach due to illiteracy or limited access to digital technology. These families were paid a small amount for participation to encourage more parents to complete the questionnaire. Parents provided fully informed consent and were able to withdraw from the study at any time. Completion of the Arabic CDI Toddler took approximately 10-30 minutes. For face-to-face data collection, RAs carried printed forms of the CDI, and parents decided whether they needed help to fill it out or not, usually depending on their educational background. RAs started by reading the instructions aloud to all parents, who were then offered the chance to go through the list themselves, or have the RA read the list word by word to them.

Results

Plan of analyses

The first step was to examine how each sample represents the country's population in terms of geographical spread, environmental context and maternal education. The second step was to compare the 2021 Egyptian sample (Abdelwahab et al., 2021) and the current Egyptian one, and examine how vocabulary scores are predicted by age, SEC factors (indexed by maternal education and environmental context) and child gender in each country dataset. The third step was to compare the three datasets to one another (Egypt, Jordan and Palestine), and examine the effects of age, gender and SEC factors to determine if a single set of norms could be produced. Linear regressions were used to analyze the effects of the different factors (SEC, gender, age) and to compare the datasets with one another. In the last step we examined the psychometric properties of the Arabic CDI Toddler through Principal Component Analyses and McDonald's omegas (Hayes & Coutts, 2020), and binomial generalized linear models were used to generate final norms.

Jordan descriptive data

Of the 566 Jordanian children, 424 (75%) came from urban areas and 142 (25%) from rural. Our target was 90% and 10% respectively. The over-representation of the rural population was due to a large number of children in rural areas of the Irbid governorate (*N* = 76), following an opportunist recruitment in this area motivated by the ambition to reach families living in disadvantaged circumstances. In terms of education, 5% of mothers were illiterate (target 7%), 10% had a primary level of education (target 11%), 19% had a secondary level of education (target 54%) and 66% had a degree or above (target 28%). The final distribution was therefore skewed towards highly educated mothers, although our sample is close to target in terms of the proportion of the less educated mothers (illiterate and primary), and therefore more representative of diversity within the population than many previous CDI standardizations. In terms of geographical areas, Table A1 in the Appendix displays the proportion of children in each governorate, as well as the estimated distribution of the population across governorates. As planned, most of the children live in Amman, Irbid and Zarqa (64%), with all other governorates contributing 36% (range 0.4 to 6%).

Most parents who filled in the CDI were mothers (80.4%), as is generally observed with parental questionnaires (Cepanec, et al., 2012). Inspection of the data by child gender revealed (Table 3) that although both mothers and fathers were in agreement when reporting girls' data, there was a strong tendency for fathers to report larger vocabulary for boys. This contrasts with studies looking at the effect of the gender of the reporting parent on the children's scores, where if anything, mothers tend to report more words than fathers, but do so equally for girls and boys (e.g. De Houwer et al., 2005). Table 3 shows that the boys whose fathers filled in the CDI were on average 40 days younger than those whose mothers filled it in, and yet the former were reported to produce 12 words more (out of

100) and understand 7 more than the latter. To give a perspective on these numbers, in the Oxford CDI norms (Hamilton et al., 2000), a 17-month-old English toddler would on average produce 12 words more than a 12-month-old infant. The size of the difference between father and mother reports for male children is such in the Jordanian data that we would have to compute different norms depending on the reporting parent gender and the child's gender, which would not have been possible given the relatively small number of father reporters as compared to mothers. Therefore, we took the difficult decision to exclude the data reported by fathers for all three countries. The reasons behind this observed behaviour will be addressed in detail in the general discussion.

Insert Table 3 here

Consequently, in Jordan the total usable data was 455 children (whose mothers filled in the CDI). The distribution of child gender in each age bin is displayed in Table 4.

Insert Table 4 here

Palestine Descriptive Data

Of the 539 Palestinian children, 347 (64%) lived in urban areas, 138 (26%) in rural areas, and 54 (10%) in camps. More participants were from the West Bank (56%) than Gaza (44%), reflecting the population distribution. In the West Bank, 59% lived in urban areas, 38% in rural and 3% in camps. Informed by population demographics our targets for these groups were 75% in urban areas and 25% in rural, with no census data on camps. In Gaza, proportions in urban, rural and camp areas were 72%, 9% and 19% respectively. Given that

the census data reported no rural area in Gaza (or camps), we did not set up a target in terms of rural/urban contexts. In terms of geographical areas, our distribution broadly reflects the demographics of the country: 43% of our sample came from the three most populated governorates (Gaza, Hebron and Jerusalem, which represent 37% of the population), with representation from each of the other less populated areas (see Table A2).

In terms of maternal education, 12% of mothers had a primary level of education (target 14%), 36% had a secondary level (target 56%), and 49% had a degree or above (target 25%). Despite our best efforts, we could not recruit illiterate mothers (target 5%). This distribution is broadly similar in the West Bank and Gaza.

Most parents who filled the CDI were mothers (82%). Because of the problem identified in the Jordanian data, we removed from further analyses the data reported by the fathers, which left us with 440 data points. It must be noted from Table 3, which reports CDI scores for girls and boys as a function of the gender of the reporting parent, that the effect observed in the Jordanian data (fathers over-reporting their male child's word knowledge) was not found here. Table 5 displays the distribution of girls and boys per age bin in the final Palestinian sample.

Insert Table 5 here

Egypt Descriptive Data

The aim of data recruitment in Egypt was to augment the data published in Abdelwahab et al. (2021) with a smaller sample (*N* = 201) which itself would be more representative of the demographics of the country. Indeed in 2021 the majority of

participants were from Lower Egypt (99%) and most mothers were highly educated (Table 6). The first step in the analyses here was to examine the population representativeness of the current sample and check how the current data compared to the 2021 data. From the 2021 paper's original 436 data points, we retained 416, due to a server error in retrieving the other 20 data points.

Out of the 201 newly recruited families, 39% lived in rural areas and 61% in urban areas (this information was not required from participants in the 2021 data). This is approaching our target of 50% in each type of area. Still in the same sample, 70% lived in Lower Egypt versus 30% in Upper Egypt, which reflects the demographic of the country (ratio of 2 to 1). In terms of geographical areas, Table A3 in the Appendix displays the proportion of children in each governorate, as well as the estimated distribution of the population across governorates. Although we did not manage to cover all governorates of Egypt, and sometimes overrepresented some areas as compared to the demographics of the country (e. g. Aswan, Daqahliya), we achieved overall a better representation of the Lower/Upper Egypt divide and the urban/rural distinction. It must be noted that due to the relative sizes of the three countries, it was less easy to cover the full geographical spread in Egypt.

In terms of maternal education, in the new sample 24% of all mothers were illiterate (target 26%), 4% of mothers had a primary level of education (target 8%), 23% had a secondary level (target 44%), and 49% had a degree or above (target 22%). The distribution of maternal education between rural and urban areas can be found in Table 6, together with the original targets. Although we did not achieve the trends seen in the targets in terms of a larger proportion of illiterate mothers in the rural sample as compared to the urban sample, we are much closer to the demographics stratification than the 2021 sample.

Insert Table 6 here

Most parents who filled in the CDI were mothers (83% in the new sample, and 62% in the 2021 sample). Because of the problem identified in the Jordanian data, we removed from further analyses the data reported by the fathers, which left us with 427 data points (168 from the new sample and 259 from the 2021 sample; in the new sample, the data of two reporting parents who identified themselves as "other" gender were kept). It must be noted from Table 3 which reports CDI scores for girls and boys as a function of the gender of the reporting parent, that the effect observed in the Jordanian data was not found here; CDI scores seem to be aligned with the age of the children more than with the relationship between the reporting parent's gender and the child's gender. Table 7 (left part) displays the distribution of girls and boys per age bin in the final Egyptian samples where mothers reported the data (new and 2021).

Insert Table 7 here

Do the New Egyptian Sample and the 2021 Sample Produce Similar Data?

The main differences between the two Egyptian samples were maternal education and spread of geographical locations. Although we did not collect information about rural versus urban environments in 2021, we could first check whether this factor explains some of the variance in the new sample, above and beyond the other factors (age, gender and maternal education).

In the new sample, 98 children lived in urban areas versus 70 in rural areas. The distribution of maternal education was quite similar in both areas (comparing urban and rural: primary education: 36% and 29%; secondary education: 14% and 36%; tertiary: 50% and 36%). In the 2021 sample, nearly all families were from Lower Egypt and presumably mostly in urban areas, although this cannot be upheld with certainty. Maternal education was primarily tertiary (87%), then secondary (10%) and primary (3%).

In the new sample, a linear regression on CDI production scores[1] with children's age, gender, maternal education (1: primary; 2: secondary; 3: tertiary) and environment (urban versus rural as a dummy variable) shows that only age (unstandardized $B = 1.98$, $p < .001$) and environment (unstandardized $B = 6.86$, $p = .046$) significantly contribute to the model ($R^2 = .29$, $F(4, 167) = 16.25$, $p < .001$). This means that children living in rural area are reported to produce nearly 7 more words than those in urban areas. For CDI comprehension scores, the significant predictors are again age ($B = 2.09$, $p < .001$), maternal education ($B = 4.57$, $p = .018$) with environment failing to reach significance ($B = 6.54$, p = .057) (model: $R^2 = .32$, $F(4, 167) = 19.27$, $p < .001$). In summary, environment seems to have an impact on word knowledge – yet statistically modest – in the opposite direction to that predicted by the literature (e. g. Al Lawati & Al Mukhaini, 2022).

Given that the original sample was collected mostly in urban areas, we might expect environment to lower the scores in the 2021 sample as compared to the new sample, once accounting for maternal education.

A linear regression on CDI production scores with age, gender, maternal education and dataset (old versus new) shows a significant model for production ($R^2 = .40$, $F(4,426) = 70.82$, $p < .001$) with age (unstandardized $B = .076$, $p < .001$), and dataset (unstandardized $B = 5.04$, $p = .023$) being the only significant contributors. For comprehension scores, the

model was also significant (R^2 = .41, F(4,426) = 72.40, p < .001), with the significant contribution of age (unstandardized B = .085, p < .001), maternal education (unstandardized B = 4.22, p = .012) and dataset (unstandardized B = 10.94, p < .001). This means that children in 2021 were reported to understand on average 10 fewer words and produce 5 fewer words than those assessed in the current study.

Could the difference between the two datasets be due to children from rural areas in the current study providing higher scores than urban ones? Assuming that all children from the 2021 dataset came from urban areas, a new regression on comprehension scores with all children from the 2021 dataset and only the urban children from the current study (N = 98) was conducted with age, gender, maternal education, and dataset. The model was significant for comprehension and production (R^2 = .43, F(4,356) = 65.30, p < .001; R^2 = .45, F(4,356) = 70.73, p < .001). For comprehension, age, maternal education and dataset were significant predictors (dataset: unstandardized B = 9.87, p = .001), whereas for production, only age was a significant predictor (effect of dataset: unstandardized B = 3.38, p = .17). Therefore, our hypothesis is only partially confirmed: data from the previous and current samples are not distinguishable on production when considering only urban children, but they are still different in comprehension, with the new scores being higher overall than the 2021 one.

Another difference between the two samples is the distribution of maternal education (87% of mothers in the 2021 sample were highly educated, against 44% in the new sample). What could partially explain the difference between the two datasets is lower SECs families over-reporting CDI scores (Feldman et al., 2000; Reese & Read, 2000). Indeed, when considering only the highly educated mothers, regression analyses do not reveal any significant contribution of dataset for production, but still for comprehension.

Figures 2 and 3 illustrate the dispersion of data in each data set, for comprehension and production. Visual inspection shows that scores reach similar levels from the ages of 22 months onwards in the two datasets, with the differences mainly due to the ages 8 to 17 months where scores are higher in the new dataset, and especially in comprehension. It is difficult to pinpoint exactly why this happened. The two CDIs were exactly the same, but the sampling was different, both in terms of geographical areas, environmental factors and maternal education; the platform was also slightly different, being more user-friendly in the current dataset. What is reassuring is that when the samples are equated on environment background (only considering the urban families) or on maternal education level (only considering tertiary education), then there is no longer an effect of dataset for production (but still for comprehension). In what follows, because there is no reason to exclude one set and not the other, we decided to pool together the two samples to what will be referred to as the Egyptian data (*N* = 427).

Insert Figures 2 and 3 here

Comparison of the Three Datasets: Egypt, Jordan and Palestine

Can we produce a single set of norms for the three countries? To answer this question, we would need to show that the factor country does not explain a significant part of the variance once age, gender, environment and maternal education have been accounted for. Table 8, which summarizes the descriptive data for each country, shows that the distribution of maternal education, the proportion of girls and the CDI scores are very similar in Jordan and Palestine, but seem different to Egypt.

Insert Table 8 here

A regression analysis on comprehension scores with age, gender, maternal education, environment and country (each country coded as a dummy variable and Egypt set as the baseline) revealed a significant model (R^2 = .49, F(6,1315) = 219.77, p < .001), with all factors but gender contributing significantly (see Table 9). The same analysis on production also provided a significant model (R^2 = .42, F(6,1315) = 158.32, p < .001), with all factors but maternal education significantly contributing.

Insert Table 9 here

The direction of predictions was as expected for age, gender (girls produce more words than boys) and education (more educated mothers report more words, at least in comprehension). The effect of environment suggests that children in urban areas know and produce fewer words than those in rural areas, which was also observed in the Egyptian data. The effect of country was due to the Egyptian scores being higher than those in Jordan and Palestine. When comparing the countries two by two (e.g. Jordan and Egypt in the same model with country as a dummy variable), Jordanian and Palestinian scores for comprehension and production are not significantly different (see Table 10). The analyses show an expected effect of age, gender (girls produce more words than boys) and maternal education (more educated mothers report more words in comprehension and production) but no effect of environment. These results suggest that norms could be calculated by pooling together Jordanian and Palestinian data.

Insert Table 10 here

In contrast, Egyptian scores are significantly different from those of each of the two other countries (see Table 10). Children in Egypt are reported to understand 14 words fewer than those in Jordan and 7 fewer than those in Palestine; they are also reported to produce 6 words fewer than those in Jordan and 2 fewer than those in Palestine. It is clear from Table 8 that the sampling was different in Egypt and the two other countries: most Egyptian children are in urban areas and have highly educated mothers. Whereas high education has been seen to lead to higher CDI scores (e.g. Cadime, et al., 2018; Kas et al., 2022), we have seen that overall children in urban areas tend to be reported to know less words than in rural areas, and the Egyptian data might be highly colored by the interaction between these two opposing factors. Therefore we decided to calculate separate norms for Egypt.

Psychometric properties of the Arabic CDI Toddler

Psychometric properties of the CDI Toddler were evaluated in two steps: first exploratory Principal Component Analyses (PCA) were run to evaluate whether trends and patterns could be observed within the 100-word list, and second McDonald's omegas were calculated to examine internal consistency. Each word received a score of 0 (not understood), 1 (understood but not produced) or 2 (produced).

A PCA using the varimax method (SPSS 25.0.02) using the screen plot test identified 3 factors explaining 56.7% of the variance, with Eigenvalue above 2.7. Factor 1 was comprised of 43 items explaining 22.6% of the variance with factor loadings from .389 to .788. Factor 2 was represented in 29 items, explaining 18.3% of the variance, with loadings from 0.381 to 0.727. Factor 3 was comprised in 28 items, explaining 15.7% of the variance with factor loadings from .452 to 0.678. When analyzing each country separately, the PCAs produced

very similar results. Inspection of the different groupings did not reveal any obvious classification.

McDonald's omega for the full dataset was 0.990, indicating acceptable internal consistency (Jordan: 0.991; Palestine: 0.990; Egypt: 0.989).

Norms of Arabic CDI Toddler Scores for Arabic-Speaking Infants Aged 8 to 30 Months

Jordan and Palestine

When pooling together data from Jordan and Palestine (N = 895), a regression on comprehension scores with age, gender, maternal education and environment (urban, rural and camp) gave a significant model (R^2 = .51, F(4,894) = 230.53, p < .001) with age and maternal education significantly contributing (age: unstandardized B = .096, p < .001; education: unstandardized B = 3.70, p < .001). For production, a significant model was also calculated (R^2 = .43, F(4,894) = 167.81, p < .001) with age, maternal education and gender significantly contributing (age: unstandardized B = .098, p < .001; education: unstandardized B = 2.91, p = .008; gender: unstandardized B = -5.34, p < .001). These results are very similar to those of the wider literature: girls produce more words than boys, but do not have a larger receptive vocabulary; children from mothers with higher education understand and produce more words than children from mothers with lower education.

Norms were created in the manner suggested by Fenson et al (1994) and used in the previous sample (Abdelwahab et al, 2021). That is, participants' age in months was calculated based on rounding to the nearest month, then in each one-month age range the data for comprehension and production were placed into 10 quantiles, from 10th to 90th, with the addition of a 16th percentile line (where the 10th percentile indicates the lowest scoring 10% of participants, and the 90th percentile indicates the highest scoring 10% ; see

the General Discussion for the justification of the 16th percentile). Then a binomial generalized linear model (GLM) was fitted for each of the 10 quantiles across age, such that the dependent variable was the proportion of words either comprehended or produced, and the independent variable was age in months. For this dataset, the age range was 7-31 months of age. Figure 4 indicates the comprehension (left) and production (right) norms for the pooled Jordanian and Palestinian data.

Insert Figure 4 here

Egypt

First, for the 427 Egyptian children, a regression on comprehension scores with age, gender, maternal education and environment (urban vs rural) gave a significant model (R^2 = .40, $F(4,426) = 70.01$, $p < .001$) with only age and environment contributing (age: unstandardized $B = .085$, $p < .001$; environment: unstandardized $B = 11.08$, $p < .001$). Similarly, the model for production was significant ($R^2 = .41$, $F(4,426) = 72.10$, $p < .001$) with only age and environment contributing (age: unstandardized $B = .076$, $p < .001$; environment: unstandardized $B = 7.69$, $p < .001$). In both cases, mothers report more words for children in rural areas than urban areas. Figure 5 displays the scores for comprehension and production as a function of maternal education and illustrates the fact that although there is no statistical effect of education, mothers with primary education tend to over-report in the earliest age range (under 17 months; see also Feldman et al., 2000, for similar effects). In contrast, for the older children, effect of maternal education tends to resemble what would be expected from the literature, with mothers with primary education reporting less words understood and produced than mothers with a higher education.

Insert Figure 5 here

Norms for the pooled Jordan and Palestine data were constructed as above, with separate binomial GLMs fitted across each of the 10 quantiles, using the same formula as the Jordan-Palestinian dataset. In the case of the Egyptian data, participant age in months ranged from 8 to 31 months. Figure 6 shows the norms for comprehension (left) and production (right) over age.

Insert Figure 6 here

Discussion of Study 1

Through the standardization of the Arabic CDI Toddler in Jordan, Palestine and Egypt, the primary aim of this paper was to examine if data could be similar enough across the three countries for a single set of norms to be produced. This would pave the way for potential further use of the Arabic CDI beyond the borders of these three countries. Other aims were to obtain a representative sample of the population in terms of socio-economic circumstances and geographical spread, and examine the effects of gender and socio-economic circumstances on language development in non-Western and non-Asian communities.

The production of single norms across multiple Arabic dialects was successfully achieved for Jordan and Palestine, where CDI scores were strikingly comparable, once accounting for age, child gender, maternal education, and environmental context. In addition, these data showed a strong effect of gender on production, with girls reported to

say more words than boys, which is similar to what is usually reported in young children (Bornstein et al., 2004; Fenson et al., 2007; Floccia et al., 2018; Frank et al., 2017). An effect of maternal education was also observed, with more educated mothers reporting more words being understood and said than less educated mothers (Bornstein et al., 1998; Hoff, 2003). These two effects replicate and extend what has already been observed in Western or Asian communities (Frota et al., 2016; Kas et al., 2022), first grounding the effect of gender on language development in biological origins (Eriksson et al., 2012; Lange et al., 2016), or at least in (broadly) universal differential socialization. Second, they demonstrate the impact of maternal education on children's language skills – including the fact that mothers with low education may over-report their child's language skills (Fenson et al., 2007). It is widely accepted that the impact of maternal education, or socio-economic circumstances in general, are due to families from more disadvantaged background having more difficulties in providing language enriching environments (e.g. Hoff, 2003; Wareham et al., 2021).

The Egyptian data collection, which complemented the data published in Abdelwahab et al. (2021), brought unexpected findings: significant differences between the 2021 and the current dataset were observed, and when data were amalgamated – since there was no reason to disqualify one set of data over another – the comparison between Egypt and the two other countries lead us to provide separate norms for Egypt. The puzzling result is actually that whether it was in Jordan, Palestine or Egypt, newly collected data showed higher scores than the 2021 dataset from Egypt. What had changed? The main difference is that a third of newly collected data in each country were collected face to face, on the phone, or sometimes reading out the words to parents when required. This was particularly the case in rural areas, where research assistants travelled to meet with

nurseries and parents. In contrast, all 2021 data were collected online. We tentatively suggest that in general, parents would be more inclined to pay more attention to each word when filling in the questionnaire in the presence of a research assistant, rather than online. A clearer sign that this is what happened comes from a closer look at what appears to be an environmental effect.

Recall that in Jordan and Palestine, environment (urban, rural or camp for Palestine) did not modulate the vocabulary scores. In contrast, in Egypt, urban children were reported to know less words than rural children, which contradicts initial expectations (Al Lawati and Al Mukhaini, 2022). However, children in the 2021 dataset were primarily recruited in urban areas, whereas in the new dataset, it was 60% urban and 40% rural. Two comparisons are of interest here: the comparison between children from urban areas in the two Egyptian datasets, and the comparison between urban and rural children in the new dataset. We showed in the results section that urban children in the new dataset outperformed the entire (urban) 2021 dataset. We can add here that when considering the new dataset only, the difference between urban and rural only just reaches significance. Therefore, it looks like the difference between the 2021 dataset and the current Egyptian one is not due to a difference between urban and rural environment, but to the fact that about a third of all data in the new Egyptian set have been collected face to face, and even more so in rural areas. If the environmental effect is actually a “research assistant” effect, one might ask why then we did not observe an environment effect in all countries, given that there was more face-to-face data collection in rural versus urban environment overall. The main reason why the environmental effect was only evident in the new Egyptian dataset is probably because in rural areas, the general level of education is lower than in urban areas, as shown in the census and reflected in our sampling; the effect of maternal education would act against the

effect of research assistant. As there were more illiterate mothers in the new Egyptian dataset than in the two other countries (29% in Egypt, 2% in Jordan and none in Palestine), the proportion of face-to-face interviews was necessarily higher in Egypt, pushing the effect of "research assistant" higher than the counteracting effect of maternal education. In short, this means that the differences between CDI scores in Egypt versus Jordan and Palestine are likely due to a combination of SEC factors and data collection mode, more than to genuine country related effects. Having said that, these differences also reflect country-specific economic contexts, because the need for more face-to-face data collection in Egypt was driven by the larger proportion of families who needed this level of input. The general message here is that CDI users should be aware of the possible "research assistant effect" described above, which could in principle inflate the scores for any language or country tested.

Another unexpected finding was that Jordanian fathers tended to over report male children's word knowledge as compared to mothers; in contrast, their responses for female children were very similar to those of mothers. This left us with the difficult decision to remove the data from fathers' reports in Jordan, and further in Palestine and Egypt, in order to work with comparable datasets. Keeping the data would have required us to calculate separate norms for father-son dyads, which would not have been feasible given the relative paucity of such data points. Were fathers reporting high language scores for boys because of their expectations and beliefs about their sons' achievements, or did it reflect more advanced language skills in boys due to more complex and richer father-son communication opportunities? Evidence so far suggests the former. First it must be noted that cross-cultural observations – as well as the current data collected with mothers in Jordan and Palestine - show that when a gender effect is observed (which seems to depend on the age with a

greater likelihood from the age of 2; Fenson et al., 1994), it is always in the direction of girls being ahead of boys. Second, the effect of the reporter's gender has not been found to systematically modulate the data: De Houwer et al. (2005) examined Dutch CDI reporters' agreement by comparing scores given by mothers and fathers from the same families. They found that although agreement may not necessarily be perfect, there was no systematic tendency for one category of parents to over-report as compared to the other gender; in fact, if anything, mothers tended to report more words than fathers. Third, when looking at the interaction between reporter gender and child gender in American English toddlers aged between 18 and 26 months, Lovas (2011) observed that mothers tended to report higher CDI scores than fathers, as found in De Houwer et al. (2005). But the effect of child gender was similar for both mothers and fathers, with girls reported to know more words and produce longer utterances than boys. It follows that father-son dyads were actually the cases where the lowest number of words and the lowest MLUs were reported, which according to Lovas, may offer a new perspective on gender socialization and in particular, they hypothesize, in men's difficulties to engage in complex intimate conversations (Lovas, 2011). Perhaps such differences also reflect the contexts within which fathers tend to interact with their children providing fewer language rich interaction opportunities.

Having said that, the aforementioned studies were conducted in the Western world, and it is established that parenting style can differ between Eastern and Western countries. For example, Dwairy and Achoui (2010), analyzing questionnaires from 2884 participants from Arab, Argentinian, French, Indian, and Polish adolescents, reported that Eastern parents exert more authority and are more likely to enforce rules and boundaries in their children's lives than Western parents. Arab countries being traditionally patriarchal (e.g. Olmsted, 2005), it could be that fathers spend more time with their infant sons than with

their daughters, resulting in more communication opportunities, which would enhance the quantity and quality of language input and boost language skills. However, if that were the case, one would expect mothers to report more word knowledge for their sons than their daughters, albeit not to the same extent perhaps. Our data do not suggest this, and in the absence of a systematic within-family comparison of parents' reports as a function of their child gender, as was completed by Lovas (2011), we can only reasonably conclude that fathers in Jordan tend to over-report their sons' progress most likely due to their high expectations for their sons. This is a finding that should be kept in mind for professionals assessing young children's language skills through parental questionnaires.

In the next section, we will present the development and norming of the Arabic CDI Child for children aged 30 to 48 months. The stratification and recruitment processes were similar to those used in the first study, but one important difference is that there was no pre-existing dataset for the Arabic CDI Child in Egypt, Jordan or Palestine, as there was for the Arabic CDI Toddler in Egypt; in addition, there was no CDI Child for Arabic. After developing the Arabic CDI Child, the objective was to recruit most data in Egypt, and compare smaller datasets from Jordan and Palestine, with the hope that there would be no significant differences between countries and that single norms could be produced. Note that this plan was designed before the full data from the Arabic CDI Toddler were collected. Of interest will be to again examine effects of gender and socio-economic circumstances, as well as a possible interaction between reporter gender and child gender.

The Arabic CDI Child is based on the CDI-III (Fenson et al., 2007), which has been adapted into several languages (see below), but did not gain the same popularity as versions for younger children. One reason for this, as we will see below, is that in many CDI-III adaptations, a ceiling effect has been observed for its different sections and in particular for

the word list. To avoid this, we chose to follow the recommendations of the Swedish CDI-III (Eriksson, 2017) which avoided this pitfall by selecting words from four difficult semantic categories instead of covering a range of more familiar and concrete domains.

Study 2 – Arabic CDI Child

The American English CDI-III, originally normed with children between 30 and 37 months of age (Fenson et al., 2007), has a high level of validity and reliability, making it an effective tool for identifying children who may need further diagnostic evaluation (Feldman et al., 2003, 2005; Perra et al., 2015). Based on our review of the different language adaptations (see Table S4 in the Supplemental Material), we decided to retain the following core components: a 100-word list, a filtering question about word combinations and questions about longest utterances for mean length of utterance (MLU) evaluation, a grammatical complexity part (23 sentence pairs), and a language use section (11 yes/no questions).

The following section describes the preliminary steps to develop the Arabic CDI Child and establish its validity and reliability. This is followed by the main standardization phase in Egypt, Jordan and Palestine, which leads to the analyses of the effects of factors known to modulate language learning (gender, socio-economic circumstances indexed by maternal education and environmental context), and the computations of norms for Arabic-speaking children aged 30 to 48 months. As per Study 1, our main research question was to ask whether a pan-Arabic lexeme approach to language assessment tools is realistic.

Preliminary Phase

Developing the Arabic CDI Child

After obtaining authorization from the MacArthur-Bates Board to develop and adapt the multi-dialect Arabic CDI Child into Egyptian, Jordanian, and Palestinian[3] Arabic, we

adapted and translated the sections covering general communication abilities and vocabulary from the Swedish and the Basque CDI-III versions (Table S4 in Supplemental Material), as well as for a Syrian version developed by Mårten Eriksson (personal communication, 28/10/2019). Various CDI adaptations have suffered from ceiling effects, in particular in the word list section (Table S4), with the notable exception of the Swedish CDI-III (Eriksson, 2017), which has inspired the development of the Estonian (Tulviste & Schults, 2020), the Mexican Spanish (Jackson-Maldonadoet al., 2022), and the European Portuguese CDI-III (Cadime et al., 2021). We started with translating the list from Eriksson (2017) which contains 100 words covering 4 semantic domains (food, body parts, cognitive words, and emotion words). All these words describe concepts that are universal (e.g. love, compare, knee), even for food (e.g. chop, salt, spicy), and as pointed out by Eriksson (2017), they have been showed to expand slowly in the pre-school years (food: Clark, 1995; body parts: Fenson et al., 1994; cognitive and emotion words: Naigles, 2000). Using Open Lexicon (Pallier et al., 2019), English and Arabic orthographic frequencies were highly correlated for the 50 words which had an entry in the Arabic database ($r = .57$, $p < .001$). The aim of this section is for parents to indicate whether their child says the word or not yet.

As is found across all CDI-IIIs, this section was followed by a filtering question about word combination (in some CDIs, this can be located before the word list, and contains statements rather than questions). If the parent indicated that their child did not combine words yet, the questionnaire would end. Otherwise, they were asked to write down the longest 3 sentences they had heard their child produce recently to estimate an MLU.

This was followed by the initial sentence complexity section, containing 42 pairs of sentences with a simple and a complex version. These sentences were specifically designed for Arabic, and covered the use of past tense, future tense, passive tense, pronouns,

comparatives, subordinates, and conjunctions of clauses. They were constructed based on a review of the (scarce) literature about Arabic acquisition, and on the intuitions of the Arabic-speaking team members (comprising linguists and speech and language therapists).

The final part about language use included 11 yes/no questions about the forms of language that express more abstract concepts (e.g. does the child make complete sentences using why, how, and who).

Versions in Egyptian, Jordanian, and Palestinian dialects were first generated by team members who are native to these countries, with the collaboration of local research assistants. Following this initial development, a first series of online reflection groups were run in each country to assess the relevance of the linguistic material, reduce the number of sentence pairs to 23, and clarify the formulation of instructions. Twenty participants in each country were divided into 4 groups of 5 participants each, with 2 groups for parents of children from the lower age range (30 to 36 months) and 2 groups for those from the higher age range (40 to 48 months). Using WhatsApp groups, parents were sent each section one after the other with 2 to 3 days in between, during which they were free to comment. They were asked to comment about the clarity and easiness of each section and suggest any improvements. They were also asked to comment on the sentences and their appropriateness for their child, following which we selected the 23 sentences that did not elicit a ceiling or a floor effect. Parents were also asked to suggest any dialectal/lexical varieties in the vocabulary section, which resulted in some variations being added to the words *boring, proud, hate, stupid*, and *evil* in the Jordanian dialect, and *brain, wonder, surprised, sad, boring*, and *sick* in the Palestinian dialect. No further changes were suggested to the sentence section or the language use section (the Arabic CDI Child can be found in the Supplemental material).

The Arabic CDI Child platform

The Arabic CDI Child was implemented on a bespoke online platform designed by the Psychology Technical Team at the University of Plymouth, very similar to that used for the CDI Toddler. After the initial part where the same questions were asked, the CDI questionnaire per se was presented with the following sections in that order: word production (100 words), word combination stage (1 question, which triggered the end of the questionnaire if parents answered that their child had not started combining words yet), longest sentences produced (3 questions), sentence complexity (23 pairs of simple/complex sentences), and language use (11 questions).

The platform was first tested with 30 users in total (10 users from each of the 3 countries) whose task was to assess the usability and robustness of the different functions. Each user was asked to test one particular function, in addition to a common final evaluation form. Examples of functions were setting the language of use, signing up, adding siblings, jumping backward/forward, and closing the app midway through. The common evaluation form asked questions about text clarity, access to login details, entering family details, and signing in again after a delay. They were also asked to suggest any alternative words and/or spelling for the linguistic content of the CDI. All issues and suggestions for improvement were addressed. For example, warnings about internet speed, connection safety and a requirement to switch language to Arabic before starting the platform were added.

Reliability and validity

Reliability of the Arabic CDI Child was established through a test–retest, together with a comparison of the modes of completion (online vs. paper). Twenty-eight parents of typically developing Egyptian children (*M* = 41.32 months, *SD* = 6.07, 12 females and 16

males) were asked to fill in the Arabic CDI Child on paper and then online (or vice versa), with a period of 2 to 3 weeks between the completion of the two versions. The children were recruited through the local RAs' contacts in different Northern Egypt urban districts and belonged to middle to high social classes. None of the children were reported as being born premature, having hearing problems nor having a diagnosed developmental delay. All parents reported that the children could combine words to form a sentence.

When filling in the CDI on their second attempt, parents dropped on average 2.96 words (*SD* 5.39). At the same time, they also added 1.43 words (*SD* 2.27) that they had not previously reported. When filling in the sentence section on their second attempt, parents dropped on average .68 sentences in the complex sentences (*SD* 1.25), and added .71 sentences that they had not previously reported (*SD* 1.58).

A repeated measures ANOVA was used to assess the word production scores in relation to two factors: the order of completion as a between-participant variable (online version first or second), and the mode of completion (online versus paper) as a within-participant variable. The analysis indicated no significant effect of the mode of completion (paper: $M = 54.11$, $SD = 18.33$; online: $M = 55.14$, $SD = 19.33$; $F(1, 26) = 1.28$, $p = .27$, $\eta 2 = .05$), nor any significant effect of the order of presentation (first completion: $M = 55.23$, $SD = 22.82$; second completion: $M = 55.07$, $SD = 16.57$; $F(1, 26) < 1$). Additionally, no interaction between order and mode of completion was observed, $F(1, 26) < 1$. Similar findings emerged in the analysis of complex sentence scores, revealing no significant main effect of the mode of completion (paper: $M = 11.04$, $SD = 5.65$; online: $M = 10.79$, $SD = 5.51$, $F(1, 26) < 1$), no significant main effect of the order of completion (first completion: $M = 10.54$, $SD = 4.94$; second completion: $M = 11.0$, $SD = 6.13$; $F(1,26) < 1$), and no interaction ($F(1, 26) < 1$).

These results establish the reliability of the Arabic CDI Child in a test-retest situation, in both online and paper formats of presentation, and across all sections.

The validity of the Arabic CDI Child was established through concurrent validity with the Arabic Language Test (ALT) (Rifaie et al., 1994, 2004, 2021), as for the Arabic CDI Toddler (see Abdelwahab et al., 2021). The ALT, originally developed for children aged 2 to 8 years and standardized with 120 Egyptian children, provides a comprehensive assessment of a child's language comprehension and production. It contains 5 sections including semantic understanding as measured by picture vocabulary items, receptive and expressive syntactic abilities, pragmatic skills, and phonological proficiency. In the original paper (Rifaie et al., 1994), the validity of the test was established through six different methods, including factorial validity, while reliability was assessed through test-retest. It was then revised in 2004 and in 2021 (Rifaie et al., 2004, 2021) to improve its ability to diagnose language difficulties and obtain a more accurate language age, based on a larger sample, in particular for the age range 2 to 4 years (N = 400). It showed again good validity (internal consistency, contrasted group, factorial, face and judgment) and reliability.

A new group of 20 toddlers aged 39 months +/- 2 months (8 male and 12 female) from rural areas in Northern Egypt were tested, with most mothers having at least a technical qualification. None of the children were born prematurely, nor had any diagnosed developmental delay or hearing problems. The test took between 45 to 60 minutes to administer (5 sections with 8 to 16 questions per section), in a face-to-face interaction with the child. The ALT was typically administered on the same day as the Arabic CDI Child completion by the parent.

The results showed strong positive correlations between the CDI vocabulary production scores and the ALT word production scores (r = .95, p = .000), the CDI sentence

complexity scores and the ALT expressive syntax scores (r = .86, p = .000), and the CDI language use section and the pragmatic section of the ALT (r = .71, p = .000). These results suggest that the two tests measure the same constructs, and establish the validity of the Arabic CDI Child.

Standardization Phase

Participants

The data from a total of 2153 children were collected between 2021 and 2022, out of which 719 were rejected for the reasons described in the participant flow chart in Figure 7, leaving 1434 usable data. The target age range was 29.5 to 48.4 months. The majority of these data were collected online using a variety of social media specific to each country. About a third were collected by local research assistants in face-to-face interviews (84 in Jordan, 470 in Egypt and 53 in Palestine), most of whom were living in rural areas, and/or with parents who were either illiterate or with primary education, and/or unable to access the online website themselves, or without internet connection enabling them to sign up.

Figure 7 here

Stratification and sample sizes

Census data was used as in Study 1 to create a representative sampling frame with regards to region, environment (urban/rural) and maternal education (see Supplemental Material and Table 2). The initial target sample sizes were 200 children in Jordan, 200 in Palestine and 900 in Egypt, with the prospect of using Egyptian data as a benchmark. A sample size of 200 Jordanian and Palestinian children was needed for modelling an effect of ~0.04 (the smallest effect found in the 2021 Egyptian data) with a 0.80 power (Berndtson,

1991). As for the Egyptian target, typical CDI standardizations collect between 20 and 50 datapoints per one-month age bin (e.g. 22 in Slovenian, Marjanovič-Umek et al., 2013; 33 in Estonian, Urm & Tulviste, 2021; 46 in Mandarin Chinese, Hao et al., 2008). With 900 data points in Egypt, we would obtain about 47 children per monthly age bin (N=19), and 68 once pooling together data from Egypt, Jordan and Palestine, which is in the top range of what is usually found in the literature in terms of sample size.

Procedure

All aspects of the procedure are identical to those in Study 1. Completion of the Arabic CDI Child took approximately 20-40 minutes.

Results

Plan of analyses

The first step was to examine how each sample represented the country's population in terms of geographical spread, environmental context and maternal education. The second step was to compare the three datasets to one another (Egypt, Jordan and Palestine), to determine if a single set of norms could be produced. As in Study 1, linear regressions were used to analyze the effects of the different factors (SEC, gender, age) and to compare the datasets with one another. Psychometric properties were also examined, and binomial generalized linear models were used to generate final norms. Please note that the MLU data were not analyzed in the current paper.

Jordan Descriptive Data

Out of the 309 Jordanian children, 268 (87%) came from urban areas and 41 (13%) from rural, which is very close to our initial targets (respectively 90% and 10%). In terms of education, 8% of mothers were illiterate (target 7%), 10% had a primary level of education (target 11%), 47% had a secondary level of education (target 54%) and 36% had a degree or

above (target 28%). The final distribution is slightly skewed towards highly educated mothers, although our sample is very close to target in terms of the proportion of the less educated mothers (illiterate and primary), and therefore more representative of diversity than many previous CDI standardizations.

In terms of geographical areas, Table A4 in the Appendix displays the proportion of children in each governorate, as well as the estimated distribution of the population across governorates. As planned, most of the children (67%) live in Amman, Irbid and Zarqa, with all other areas containing between 1 and 7% of the population (no families from At-Tafileh could be recruited).

Most parents who filled in the CDI were mothers (81%). Because of the interaction between reporter gender and child gender found in the Arabic CDI Toddler, where fathers were found to over-report boys' language skills, we inspected the CDI scores as a function of reporter and child gender (Table 11). Both mothers and fathers appeared to be in agreement when reporting boys and girls' data, with a slight advantage for girls, for both vocabulary count and sentence complexity. Therefore, contrary to the Arabic CDI Toddler data where only data from female reporters were kept, here all data were retained. The distribution of child gender in each age bin is displayed in Table 4 (right part).

Insert Table 11 here

Palestine Descriptive Data

Of the 295 Palestinian children, 78% lived in urban areas, 20% in rural areas, and 2% in camps. We had more participants from the West Bank (64%) than Gaza (36%), reflecting the population distribution. In the West Bank 70% lived in urban areas (target 75%), 29% in

rural areas (target 25%) and 1% in camps. In Gaza, 92% lived in urban areas (target 100%), versus 4% in rural areas and 4% in camps. In terms of geographical areas, Table A5 in the Appendix displays the proportion of children in each governorate, as well as the estimated distribution of the population across governorates. Overall, our distribution reflects the population distribution: 34% of our sample live came from the three most populated governorates (Gaza, Hebron and Jerusalem, which represent 37% of the population), with representation from every other area.

In terms of maternal education, 17% of mothers had a primary level of education, 54% had a secondary level, and 28% had a degree or above. Only two (0.7%) illiterate mothers were recruited (our target was 5%). This distribution is broadly similar in the West Bank and Gaza. We reached our targets to have more mothers with secondary education than degree (targets were 56% and 25% respectively) and close to target for the proportion with a primary education (target 14%).

Most parents who completed the CDI were mothers (87%). To provide a comparison for what had been observed for the Arabic CDI Toddler in Jordan, Table 11 provides the word production and sentence complexity scores reported by mothers and fathers for boys and girls. Visual inspection does not suggest any interaction between reporter gender and child gender. Table 5 (right part) displays the distribution of girls and boys per age bin in the Palestinian sample.

Egypt Descriptive Data

Of the 830 newly recruited families, 61% lived in rural areas (target 50%) and 39% in urban areas (target 50%). About 64% lived in Lower Egypt (target 67%) versus 36% in Upper Egypt (target 33%), which reflects the demographic of the country. In terms of geographical areas, Table A6 in the Appendix displays the proportion of children in each governorate, as

well as the estimated distribution of the population across governorates. Although we did not manage to cover all governorates of Egypt, and sometimes overrepresented some areas as compared to the demographics of the country (e.g. Ismailiya, Fayyoum), we achieved overall a fair representation of the Lower/Upper Egypt divide and the urban/rural distinction. As in Study 1, due to the relative sizes of the three countries, it was less easy to cover the full geographical spread in Egypt.

In terms of maternal education, 20% of all mothers were illiterate (target 25%), 10% of mothers had a primary level of education (target 8%), 53% had a secondary level (target 44%), and 18% had a degree or above (target 22%). The distribution of maternal education between rural and urban areas can be found in Table 12, together with the original targets. Inspection of the table shows a remarkable fit between the target distribution and the current sample.

Insert Table 12 here

Most parents who filled in the CDI were mothers (91%). As in Jordan and Palestine, visual inspection of these data (Table 11) did not reveal any interaction between these two factors. Table 7 (right part) displays the distribution of girls and boys per age bin in the Egyptian sample.

Comparison of the Three Datasets: Egypt, Jordan and Palestine

Can we produce a single set of norms for the three countries? To answer this question, we would need to show that the factor "country" does not explain a significant part of the variance on the CDI scores, once age, gender, environment and maternal education have been accounted for. Table 13, which summarizes the descriptive data for

each country, shows that the socio-economic circumstances in Jordan and Palestine are rather similar to each other but differ from Egypt (majority of urban population in Jordan and Palestine, higher level of education).

Insert Table 13 here

Overall, 8.1% of children were reported to not combine words yet (in which case parents were not asked to complete the sentence complexity and language use sections), with proportions similar for the three countries (χ^2 (2, 1434) = 5.79, p = .055). Figure 8 shows the proportion of these children as a function of age and country. As would be expected, at least in Egypt and to some extent in Palestine, parents are more likely to report no word combination for the younger children. Responses for Jordan appear to be evenly distributed across ages, but it must be remembered the relative smaller number of responses overall for Jordan and Palestine as compared to Egypt. For a clinical use of the Arabic CDI Child, it will be advised to refer any child who does not combine words yet from the age of 30 months (which is the minimal age in our data). In what follows, all analyses were conducted on the subset of children reported to combine words (N = 1328).

Figure 8 here

A regression analysis on word production scores with age, gender, maternal education (1, 2 and 3 for primary, secondary and tertiary), environment and country (each country coded as a dummy variable and Egypt set as the baseline) revealed a significant model (R^2 = .121, F(6,1321) = 30.19, p < .001), with every factor but child gender contributing

significantly (see Table 14). The direction of predictions was as expected for age and education, with children with more educated mothers reported to produce more words. As was found in Study 1, there was a trend for women in rural areas to report more words than those in urban areas. Finally, children in Jordan were reported to know more words than those in Egypt and Palestine, who did not differ from each other.

Insert Table 14 here

A regression analysis on sentence complexity scores with the same variables as above revealed a significant model (R^2 = .155, F(6,1321) = 40.36, p < .001), with all factors but gender and maternal education contributing significantly (see Table 14). Environment effects were similar to what was observed for word production. However, differences between countries were due to Egyptian children being reported to know more complex sentences than children in Jordan and Palestine, who did not differ from each other. For language use scores, the model was also significant (R^2 = .140, F(6,1321) = 35.76, p < .001), with all factors but gender and environment contributing significantly. Here the effect of country was due to Egyptian children scoring lower than those in Jordan and Palestine, who did not differ from each other (Table 14).

Psychometric properties of the Arabic CDI Child

Word production

As in Study 1, exploratory PCA were run to evaluate whether trends and patterns could be observed within the 100-word list, followed by McDonald's omegas to examine internal consistency. Each word received a score of 0 (not produced) or 1 (produced).

Analyses were carried out first on the entire dataset (N=1328), with results per country reported afterwards.

A PCA with the varimax method (SPSS 25.0.02) identified 4 factors using the scree test (Osborne & Costello, 2009) explaining 34.05% of the variance, all with an Eigenvalue between 2.57 and 21.70. Factor 1, which represented mostly cognitive and emotion words (e.g. believe, curious), was comprised of 46 items explaining 13.5% of the variance with factor loadings from .333 to .657 (they were known by 42.8% of children). Factor 2, mainly containing food words (e.g. cook, wash), was represented in 23 items, explaining 9.2% of the variance, with loadings from 0.301 to 0.620 (known by 74.0% of children). Factor 3, containing mainly body words (e.g. head, cough) comprised 20 items, explaining 7.2% of the variance with factor loadings from .308 to 0.556 (known by 82.2% of children). The last factor grouped 8 words and explained 4.1% of variance (loadings from 0.301 to 0.797) with no obvious similarities (known by 68.6% of children). This suggests the possibility to shorten the word list to less than 100 words using semantic categories, and focusing on difficult categories such as emotions and cognition.

McDonald's omega for the full dataset was 0.963, indicating acceptable internal consistency (Jordan: 0.969; Palestine: 0.977; Egypt: 0.955).

Grammatical complexity

Each sentence received a code of 0 (simple version produced) or 1 (complex version produced). The scree plot of a varimax PCA for the full dataset produced two main factors with an Eigenvalue of 9.19 and 1.35 explaining 45.8% of the variance. The first component comprised 15 items and explained 26.1% of the variance (factors loading 0.427 to 0.718), while the second comprised 8 items and explained 19.7% (0.482 to 0.715). Inspection of the data revealed that the main distinction between the two components is the number of

morphemes in the simple version (4.53 in the first component, and 2.75 in the second) and in the complex version (7.2 in the first component versus 3.88 in the second). The first component primarily contains changes in subordinates, complements and tenses, whereas the second component comprises changes in internal clause structure (e. g. negation, wh-question, subject dropping). Ratings for the first component are lower than for the second one (50.9% of responses for the first component are for the complex version, versus 71.8% for the second component). McDonald's omega for the full dataset was 0.933, indicating acceptable internal consistency (Jordan: 0.913; Palestine: 0.933; Egypt: 0.936).

Language use

Each statement received a code of 0 (no) or 1 (yes). The scree plot of a varimax PCA for the full data set produced two main factors with an Eigenvalue of 1.15 and 3.86 explaining 45.5% of the variance. The first component comprised 8 items and explained 25.3% of the variance (factors loading 0.434 to 0.671), while the second comprised 3 items and explained 20.4% (0.616 to 0.760). The second group contained the first 3 questions of the list, and they scored higher (88.0%) than the next 8 ones (60.0%), with no obvious grouping other than position in the list. It must be remembered that this part is the last one in the Arabic CDI Child, therefore that might be simply a fatigue effect. Results per country are very similar with 2 main components identified on the scree plot (Egypt: Eigenvalue above 1.22, 46.0% variance; Jordan: 1.23, 43.7%; Palestine: 1.11, 47.2%). McDonald's omega for the full dataset was 0.811, indicating acceptable internal consistency (Jordan: 0.783; Palestine: 0.826; Egypt: 0.811).

Comparison between the three sections

As can be seen in Table 15, all three sections correlated with one another, even after controlling for age. That was the case when all data were collated, and for each country

individually. Similar results are found in each country (all correlations, bivariate or partial when controlling for age, are significant to 0.01).

Table 15 here

Norms of Arabic CDI Child Scores for Arabic-Speaking Children Aged 30 to 48 Months

Given the complexity of the effect of countries on the different datasets, we decided to calculate norms for each country separately. Norms for each country and measure were developed in the same manner as above, by calculating the deciles of each measure, as well as the 16th percentile, and then modelling each quantile independently with a binomial generalized linear model. Figure 9 indicates the norms for productive vocabulary, where proportion of words produced in Egypt, Jordan and Palestine can be seen in (A), (B) and (C) respectively. Immediately apparent is that there is a larger spread in the Palestine scores than in the other two countries.

Figure 9 here

Similarly, norms for the complex sentences can be seen in Figure 10. Here we can see that the overall norms for Jordan and Palestine look similar, but Egypt has a higher mean and more rapid slope.

Figure 10 here

Finally we also see that the language use score shows a difference between Egypt and the other two countries, displayed in Figure 11. Egypt shows a much greater spread between deciles, where a tighter fit is shown in the other two countries.

Figure 11 here

Discussion of Study 2

This second study aimed at developing and standardizing an Arabic version of the CDI-III (Fenson et al., 2007) across Jordan, Palestine and Egypt, for children aged 30 to 48 months, with the idea to produce a single set of norms. Effects of child gender and socio-economic circumstances were also examined in light of the results of Study 1, which showed robust effects of these factors on vocabulary scores with the CDI Toddler. The interaction between parent reporter gender and child gender which was observed with the CDI Toddler was also briefly examined here. A last objective was to determine whether we could avoid ceiling effects which are not uncommon in CDI-III tools (Fenson et al., 2007; Garcia et al., 2014; Kas et al., 2022), by using an approach similar to that developed by Eriksson (2017) in Swedish.

Results showed that the production of single norms across the three countries was not possible. In principle however, collating data across two countries could have been an option, but the pairing varied from one section to the next: for the vocabulary production component of the CDI Child, Egyptian and Palestinian datasets were similar, whereas for grammar and language use, Jordanian and Palestinian datasets were similar. To avoid complexity, we decided to produce norms for each country separately.

As for Study 1, the socio-economic profile of the mothers was different in Egypt than in the other countries, reflecting the population characteristics of these countries. The striking difference in the data was that there was no effect of maternal education in Egypt for the grammar and language use section, whereas such an effect was found in the two other countries. Exactly like in Study 1, this is likely to be due to an experimenter effect: harder to reach families (because of the rurality and/or their education level) were more likely to require face-to-face help from the experimenters when filling in the forms, and Study 1 showed that this seemed to result in higher scores than when parents did not require any support. For the CDI Child, four times more families were in that situation in Egypt than in the two other countries, probably resulting in the experimenter effect compensating for the maternal education effect.

Another, related, result was that children in rural areas were reported to know more complex sentences and have a more advanced use of language than those in urban areas, in the three countries. A similar effect (for vocabulary) was also observed with the CDI Toddler, and as before, the most reasonable explanation is that it is due to the aforementioned experimenter effect, that is, the fact that filling in the CDI with a research assistant results in higher scores than without, as most face-to-face support was provided in rural areas.

Contrary to Study 1, no effect of child gender was found with the Arabic CDI Child, which contrasts with our results with the Arabic CDI Toddler, but is not uncommon in CDI-IIIs (see Table S4 in Supplemental material), suggesting that from the age of 36 months, boys' production tends to catch up with girls' (e.g. Cadime et al., 2021; Simonsen et al., 2014).

In addition, contrary to Study 1, there was no tendency for fathers to over-report their sons' language skills. If anything, both mothers and fathers tended to report higher

scores for girls than for boys (although not consistently). The fact that the reporter gender effect was restricted to children aged 8 to 30 months suggests that perhaps fathers spend more time with their child as they grow older, so that their expectations about their sons' achievements can be tempered with knowledge derived from day-to-day interactions.

The resulting norms (provided in the Appendix, and figures 9 to 11) show no evidence of a ceiling effect for word production, which is a positive outcome given that we adopted Eriksson's (2017) approach for choosing words according to four semantic categories that have been showed to populate slowly during development. However, a ceiling effect was observed for sentence complexity in Egypt (Figure 10a) and language use in Palestine (Figure 11c). It must be noted that this effect has been found too in other CDI-III versions, especially for the grammar sections (see Table S4). However, as a whole, the combinations of the three scores (word production, sentence complexity and language use) should provide informative evidence of children's language skills between 30 and 48 months.

General Discussion

The challenging task of developing and norming two CDIs in three different Arabic dialects was rewarded by two main results: first, the two CDIs were found to have excellent psychometric properties. For the Arabic CDI Child, reliability was established through test-retest, McDonald's omegas, and between-scale correlations, whereas validity was successfully assessed by concurrent validity (see Abdelwahab et al., 2021, for the Arabic CDI Toddler). Word production scores showed no evidence of ceiling effects across the two CDIs. Second, and most importantly, a pan-Arabic lexeme approach proved to be successful to some extent, with common norms being calculated for Jordan and Palestine for the Arabic CDI Toddler. The fact that Egyptian data needed to be processed separately from the

Jordanian and Palestinian ones reveals two things: first, it suggests that for a pan-Arabic lexeme approach to work, the population profile of the countries must be as aligned as possible, in terms of stratification of adult education. Statistics can of course account for differences in SEC across countries, so in theory it should not be an obstacle to producing similar norms. But what unexpectedly happened is that the greater need for face-to-face data collection in Egypt, where a larger proportion of parents have received little education and/or live in remote areas, created a source of bias in reporting language skills, as was apparent in the two CDIs. This experimenter effect should not be taken lightly not only for future CDI standardization projects, but also for clinical use. Indeed, children of mothers who have received little education might require support in filling in the CDI, which might lead to overinflated scores due to social desirability bias, with important consequences for clinical decision making. Another issue to consider is whether the CDI word list is sufficiently representative of the language of children from various SES background, and whether it biases against working class families (Figueroa, 2024). We have tried to mitigate for this by piloting our word lists with a wider range of families and included variants of the lexemes wherever users suggested these. A limitation to this study is that data collection started during the Covid pandemic, especially for the Arabic CDI Toddler which started 6 months earlier than the Arabic CDI Child. During the pandemic, access to preschools and nurseries were mainly restricted, with many parents working from home or losing their jobs. Kartushina et al. (2022) collected CDI data from 13 different countries at the start and the end of lockdowns, and found that as a whole, children's language benefitted from daily parental presence. However, the self-selected sample was primarily parents with university education who probably engaged even more than usual in language-rich shared reading and screen activities. In contrast, Davies et al. (2021) compared 8-36-month-old children who

continued to attend early childhood settings during the restrictions to those who stayed at home, and found that the former over-performed the latter in terms of language skills and executive functions, and that the gap was especially noticeable for those children from less privileged socio-economic circumstances. Hence it is likely that the language scores we collected are lower than they would have been for children who were born outside of the pandemic era. Therefore, if children in the future are compared to the current norms, scores on the norms might not represent "typical" non-COVID scores. A child scoring in the 30% percentile in the current norms might have scored lower in non-COVID norms, leading to a risk in missing the identification of children with language difficulties.

Another limitation is that this tool is not appropriate for assessing multilingual children. The current norms were calculated for a sample of children not exposed to a language other than Arabic for more than 10% of the time (or to an Arabic dialect different from the one used in the country of residence). It must be noted however that the way the question was formulated does not exclude exposure to other languages outside from home, so our sample might include children exposed to other languages in preschool settings. CDI norms for bilingual children are rare: Floccia et al. (2018) designed the UKBTAT for use with 24-month-old bilinguals growing up with British English and an Additional Language, where CDI scores are adjusted by a measure of relative exposure to each language (see also bilingual CDIs in Irish-English: O'Toole & Fletcher, 2008; Maltese-English: Gatt, 2007; Welsh-English: Mills et al., 2024). We are currently adapting this approach to Lebanon where Arabic, English and French co-exist in a large proportion of the population (see Khattab et al., 2022). In the absence of such norms for Egypt, Jordan and Palestine, a possible approach advocated for in the UK by Cattani et al. (2014) is to use monolingual norms for children exposed to Arabic for at least 60% of the time (which can be measured with a language

exposure tool such as the Plymouth LEQ developed in Cattani et al., 2014), although this would need empirical testing in this context.

Recommendations for users

For both CDIs, the norms will be implemented on a freely accessible platform, usable by parents and Early Years professionals to reach an estimate of the percentile ranking of infants aged 8 to 48 months. The interpretation of results will be slightly different for parents and professionals who use the platform. Parents will be provided with general advice, tips, and useful links to professional local support. For professionals, the emphasis will be on using the CDI as a first step in a triaging process, whereby parental concern and other factors are also considered and the results of these together with the CDI used to determine whether further assessment is indicated. In that perspective - keeping in mind the warnings issued by Eriksson (2022) about the insufficient evidence for the CDIs to be used as screening tools, but also the pragmatic need for a solution where resources are scarce - we would initially recommend the conservative cutoff of the 10th percentile to identify late talkers (Dale et al., 2003; 11th percentile for Heilmann et al., 2005), around the age of 24 months (range of 18-30 months) for a more accurate predictive value (see Law & Roy, 2008, for a review). However, to maximize sensitivity, especially given the potential COVID effects on this norming sample, perhaps a less conservative cutoff of 16th-20th percentile might be indicated (further empirical study would be necessary to determine the most appropriate cutoff). Children aged at least 30 months whose parents report no word combination should be directed towards referral, following well established practice (e.g. Rudolph & Leonard, 2016).

It must be noted however that not all young children who score low on language scales or parental reports will grow up with language difficulties, but that the majority of

children with a diagnosed language disorder will have a history of protracted language development (Rice & Hoffman, 2015). However, the older children are, the more accurate the prognosis of persisting language difficulties will be: by 4 years of age a child with low language ability is highly likely to have persisting difficulties (see McKean et al., 2017), whereas low language skills at age 2 do not accurately predict children's skills at age 4 (Reilly et al., 2010).

This raises the more general question of the sensitivity of the CDI, or for that matter, any language assessment tool in the early years. Spaulding et al. (2006) reviewed 43 tests of early language development and found that measures of specificity and sensitivity were reported only for 9, with only 5 reaching a satisfactory level of sensitivity. McKean et al. (2017) indicated that language difficulties at age 4 was a better predictor of language skills at age 7, than language abilities at age 2, which points to the individual variability of language skills and outcomes in the early years. This does not mean that it is not useful to monitor language skills before the age of 4. The tools have a place within a public health approach enabling identification of those with other risks, such as socio-emotional difficulties, poor social communication skills and high parental concern who should receive specialist support; those without those other risks but low language development who should be monitored; and families who would benefit from advice regarding language enriching strategies (McKean & Reilly, 2023). The development of online/remote CDIs at multiple ages may allow this monitoring to be conducted in a cost-effective manner.

Acknowledgements

We are grateful to the families and the non-governmental organizations (NGOs) who have supported this work. We are also grateful to the Research Assistants in the three countries: Nour Alaryan, Juliette Hindaileh, Esraa Hamdy, Mariam Hamdy, Iman Saraj, Alaa El-ebbiny, Enjood Abu Qoleh. A special thank you to Mårten Eriksson for his invaluable input at the earliest stages of this project. This project was funded by the Economic and Social Sciences Research Council (ES/T003995/1). We are very grateful to our funder for having supported us extensively during the COVID pandemic, Boris Johnson's government's cuts to overseas funding in 2021, and the 2023 Palestinian-Israeli conflict. Our thoughts go to the Palestinian families in Gaza who contributed their time to this project, and might not be here any more today.

Data Availability Statement

Anonymized raw CDI data for the two age groups described in this paper can be accessed at https://osf.io/cwvx6/?view_only=67ecf1342fc94efdb32c15c60b59b20e, together with datafiles and R scripts used for the beta-binomial models.

References

Abdelwahab, A. G. S., Forbes, S., Cattani, A., Goslin, J., & Floccia, C. (2021). An adaptation of the MacArthur-Bates CDI in 17 Arabic dialects for children aged 8 to 30 months. *Language Learning and Development*, *17*(4), 425-446.

Abdul-Mageed, M. (2019). Modelling Arabic subjectivity and sentiment in lexical space. *Information Processing & Management*, *56*(2), 291-307.

Abu Haseeba, A., El Sady, S., Elshobary, A., Gamal, N., Ibrahim, M., & Abd El-Azeem, A. (2011). Standardization, translation and modification of the preschool language scale–4MD. Thesis submitted to phoniatric unit. *Ain Shams University, Cairo*.

Al Lawati, E., & Al Mukhaini, F. (2022). The psychological development of the early childhood language in Sultanate of Oman. *Journal of Positive School Psychology*, *6*(8), 6364-6374.

Andersson, I., Gauding, J., Graca, A., Holm, K., Öhlin, L., Marklund, U., & Ericsson, A. (2011). Productive vocabulary size development in children aged 18-24 months–Gender differences. *Proceedings from Fonetik 2011: Speech, Music and Hearing; Quarterly Progress and Status Report*, *51*, 109-112.

Bates, E., & Goodman, J. C. (2001). On the inseparability of grammar and the lexicon: Evidence from acquisition. In M. Tomasello & E. Bates (Eds.), *Language Development: The Essential Readings*. Oxford: Blackwell Publishers, pp. 134-162.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., ... & Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(1), 85-123.

Berndtson, W. E. (1991). A simple, rapid and reliable method for selecting or assessing the number of replicates for animal experiments. *Journal of Animal Science*, *69*(1), 67-76.

Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics*, *37*(6), 1461-1476.

Bishai, W. B. (1964). Coptic Lexical Influence on Egyptian Arabic. Journal of Near Eastern Studies, 23(1), 39–47. http://www.jstor.org/stable/543178.

Bornstein, M. H., Hahn, C. S., & Haynes, O. M. (2004). Specific and general language performance across early childhood: Stability and gender considerations. *First Language*, *24*(3), 267-304.

Bornstein, M. H., Haynes, M. O., & Painter, K. M. (1998). Sources of child vocabulary competence: A multivariate model. *Journal of Child Language*, *25*(2), 367-393.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, *53*(1), 371-399.

Bruce, T., & Spratt, J. (2011). *Essentials of Literacy from 0-7: A Whole-Child Approach to Communication, Language and Literacy*. Sage.

Brushe, M. E., Lynch, J., Reilly, S., Melhuish, E., Mittinty, M. N., & Brinkman, S. A. (2021). The education word gap emerges by 18 months: Findings from an Australian prospective study. *BMC Pediatics, 21(247).* https://doi.org/10.1186/s12887-021-02712-1

Cadime, I., Santos, A. L., Ribeiro, I., & Viana, F. L. (2021). Parental reports of preschoolers' lexical and syntactic development: Validation of the CDI-III for European Portuguese. *Frontiers in Psychology*, *12*, 677575.

Cadime, I., Silva, C., Ribeiro, I., & Viana, F. L. (2018). Early lexical development: Do day care attendance and maternal education matter? *First Language*, *38*(5), 503-519.

CAPMAS (2021). *Central Agency for Public Mobilisation and Statistics*. https://www.capmas.gov.eg/.

Cattani, A., Abbot-Smith, K., Farag, R., Krott, A., Arreckx, F., Dennis, I., & Floccia, C. (2014). How much exposure to English is necessary for a bilingual toddler to perform like a monolingual peer in language tests? *International Journal of Language & Communication Disorders*, *49*(6), 649-671.

Cepanec, M., Lice, K., & Šimleša, S. (2012). Mother–father differences in screening for developmental delay in infants and toddlers. *Journal of communication disorders*, *45*(4), 255-262.

Clark E. V. (1995). *The lexicon in Acquisition* (Vol. 65). Cambridge, UK: Cambridge University Press.

Cotter, W. M., & De Jong, R. (2019). Regional variation. In Al-Wer, E., & Horesh, U. (Eds.). The Routledge handbook of Arabic sociolinguistics (pp. 47-62). London: Routledge.

Cristia, A. (2023). A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations. *Developmental Science*, *26*(1), e13265.

Dale, P. S., Price, T. S., Bishop, D. V., & Plomin, R. (2003). Outcomes of early language delay. I. Predicting persistent and transient language difficulties at 3 and 4 years. *Journal of Speech, Language and Hearing Research, 46*, 544-650.

Davies, C., Hendry, A., Gibson, S. P., Gliga, T., McGillion, M., & Gonzalez-Gomez, N. (2021). Early childhood education and care (ECEC) during COVID-19 boosts growth in language and executive function. *Infant and Child Development*, *30*(4), e2241.

Davis, A. N., & Qi, C. H. (2020). A longitudinal examination of language skills, social skills, and behavior problems of preschool children from low-income families. *Topics in Early Childhood Special Education*, *40*(3), 172-186.

De Houwer, A., Bornstein, M. H., & Leach, D. B. (2005). Assessing early communicative ability: A cross-reporter cumulative score for the MacArthur CDI. *Journal of Child Language*, *32*(4), 735-758.

Dwairy, M., & Achoui, M. (2010). Parental control: A second cross-cultural research on parenting and psychological adjustment of children. *Journal of Child and Family Studies*, *19*, 16-22.

Eadie, P., Levickis, P., McKean, C., Westrupp, E., Bavin, E. L., Ware, R. S., Gerner, B., & Reilly, S. (2022). Developing Preschool Language Surveillance Models - Cumulative and clustering patterns of early life factors in the early language in Victoria Study Cohort [Original Research]. *Frontiers in Pediatrics*, *10*. https://doi.org/10.3389/fped.2022.826817

Eriksson, M. (2017). The Swedish Communicative Development Inventory III: Parent reports on language in preschool children. *International Journal of Behavioral Development*, *41*(5), 647-654.

Eriksson, M. (2022). Insufficient evidence for the validity of the Language Development Survey and the MacArthur–Bates Communicative Development Inventories as screening tools: A critical review. *International Journal of Language & Communication Disorders*, 58(2), 555-575.

Eriksson, M., Marschik, P. B., Tulviste, T., Almgren, M., Pérez Pereira, M., Wehberg, S., ... & Gallego, C. (2012). Differences between girls and boys in emerging language skills:

Evidence from 10 language communities. *British Journal of Developmental Psychology*, *30*(2), 326-343.

Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs-Lasky, M., Rockette, H. E., & Paradise, J. L. (2005). Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. *Child Development*, *76*(4), 856-868.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Colborn, D. K., Janosky, J., Kurs-Lasky, M., ... & Paradise, J. L. (2003). Parent-reported language skills in relation to otitis media during the first 3 years of life. *Journal of Speech, Language and Hearing Research, 46*(2), 273-287.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development*, *71*(2), 310-322.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... & Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, i-185.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., Pethick, S., & Reilly, J. S. (1993). *The MacArthur Communicative Development Inventories: User's guide and technical manual*. San Diego: Singular Publishing Group.

Fenson, L., Marchman, V. A., Thal, D., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories. User's Guide and Technical Manual*. Illinois: Paul H. Brookes Publishing Co., Inc.

Fenson, L., Bates, E., Dale, P., Goodman, J., Reznick, J. S., & Thal, D. (2000a). Reply: Measuring variability in early child language: Don't shoot the messenger. *Child Development*, *71*(2), 323-328.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000b). Short-form versions of the MacArthur communicative development inventories. *Applied Psycholinguistics*, *21*(1), 95-116.

Figueroa, M. (2024). Decolonizing (Psycho)linguistics Means Dropping the Language Gap Rhetoric. In Hudley, A. H. C., Mallinson, C., & Bucholtz, M. (Eds). Decolonizing linguistics. (New York, 2024; online edn, Oxford Academic, 30 Apr. 2024), https://doi.org/10.1093/oso/9780197755259.003.0009, accessed 20 Nov.

Floccia, C., Sambrook, T. D., Delle Luche, C., Kwok, R., Goslin, J., White, L., … & Plunkett, K. (2018). Vocabulary of 2-year-olds learning English and an Additional Language: Norms and effects of linguistic distance. *Monographs of the Society for Research in Child Development*, *83*(1), 1-108.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677-694.

Frota, S., Butler, J., Correia, S., Severino, C., Vicente, S., & Vigário, M. (2016). Infant communicative development assessed with the European Portuguese MacArthur–Bates communicative development inventories short forms. *First Language*, *36*(5), 525-545.

Garcia, I., Barreña, A., Ezeizabarrena, M. J., Almgren, M., Arratibel, N., & Barnes, J. (2014). Haur euskaldunen komunikazio-garapena neurtzen 30-50 hilabete bitartean: MacArthur-Bates CDI-III tresnaren euskal bertsioa. *Uztaro*, *88*, 33-72.

Gatt, D. (2007). Establishing the concurrent validity of a vocabulary checklist for young Maltese children. *Folia Phoniatrica et Logopaedica, 59*(6), 297-305.

Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a British communicative development inventory. *Journal of Child Language*, *27*(3), 689-705.

Hao, M., Shu, H., Xing, A., & Li, P. (2008). Early vocabulary inventory for Mandarin Chinese. *Behavior Research Methods*, *40*(3), 728-733.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But... *Communication Methods and Measures*, *14*, 1-24.

Heilmann, J., Weismer, S. E., Evans, J., & Hollar, C. (2005). Utility of the MacArthur—Bates Communicative Development Inventory in identifying language abilities of late-talking and typically developing toddlers. *American Journal of Speech-Language Pathology, 14*, 40-51.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*(5), 1368-1378.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*(4), 343-365.

Jackson-Maldonado, D., Marchman, V. A., Dale, P., & Rubio-Codina, M. (2022). The MacArthur-Bates CDI-III for Spanish-speaking Children. https://doi.org/10.31234/osf.io/7cw4v

Kartushina, N., Mani, N., Aktan-Erciyes, A., Alaslani, K., Aldrich, N. J., Almohammadi, A., ... & Mayor, J. (2022). COVID-19 first lockdown as a window into language acquisition: associations between caregiver-child activities and vocabulary gains. *Language Development Research*, *2*(1).

Kas, B., Jakab, Z., & Lőrik, J. (2022). Development and norming of the Hungarian CDI-III: A screening tool for language delay. *International Journal of Language & Communication Disorders*, *57*(2), 252-273.

Kaye, A. S. (2018). Arabic. In *The World's Major Languages* (pp. 577-594). Routledge.

Khattab, G., Abdelwahab, A. G. S., Al-Shdifat, K., Alsiddiqi, Z., Floccia, C., Kouba Hreich, E., ... & Trebacz, A. (2022). Promoting early language development in the Arab world and sustainable development goals 3, 4, 10 and 17. *International Journal of Speech-Language Pathology*, 1-6.

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research?. *First Language*, *42*(6), 703-735.

Kim, S.W., Kim, J.Y., Lee, S.Y. & Jeon, H.R. (2016) The comparison of M-B CDI-K Short Form and K-ASQ as screening test for language development. *Annals of Rehabilitation Medicine, 40*(6), 1108–1113.https://doi.org/10.5535/arm.2016.40.6.1108

Kristoffersen, K. E., Simonsen, H. G., Bleses, D., Wehberg, S., Jørgensen, R. N., Eiesland, E. A., & Henriksen, L. Y. (2013). The use of the Internet in collecting CDI data–an example from Norway. *Journal of Child Language*, *40*(3), 567-585.

Lange, B. P., Euler, H. A., & Zaretsky, E. (2016). Sex differences in language competence of 3- to 6-year-old children. *Applied Psycholinguistics*, *37*(6), 1417-1438.

Law, J., Charlton, J., & Asmussen, K. (2017). Language as a Child Wellbeing Indicator. London, Early Intervention Foundation.

Law, J., Charlton, J., Dockrell, J., Gascoigne, M., McKean, C., & Theakston, A. (2017). Early language development: Needs, provision and intervention for pre-school children from socio-economically disadvantaged backgrounds. London Education Endowment Foundation.

Law, J., & Roy, P. (2008). Parental report of infant language skills: A review of the development and application of the Communicative Development Inventories. *Child and Adolescent Mental Health*, *13*(4), 198-206.

Lovas, G. S. (2011). Gender and patterns of language development in mother-toddler and father-toddler dyads. *First Language*, *31*(1), 83-108.

Marjanovič-Umek, L., & Fekonja-Peklaj, U. (2017). Gender differences in children's language: a meta-analysis of Slovenian studies. *CEPS Journal*, *7*(2), 97-111.

Marjanovič-Umek, L., Fekonja-Peklaj, U., & Podlesek, A. (2013). Characteristics of early vocabulary and grammar development in Slovenian-speaking infants and toddlers: a CDI-adaptation study. *Journal of Child Language*, *40*(4), 779-798.

Mayor, J., & Mani, N. (2019). A short version of the MacArthur–Bates Communicative Development Inventories with high validity. *Behavior Research Methods*, *51*, 2248-2255.

McKean, C., Morgan, A., Law, J. Reilly, S., (2018) Developmental Language Disorder in Shirley-Ann Rueschemeyer & M Gareth Gaskell (Eds.) *Oxford Handbook of Psycholinguistics*

McKean, C. & Reilly, S. (2023) Creating the conditions for robust early language development for all: Part two: Evidence informed public health framework for child language in the early years. *International Journal of Language & Communication Disorders*, 58, 2242–2264. https://doi.org/10.1111/1460-6984.12927

McKean, C., Reilly, S., Bavin, E. L., Bretherton, L., Cini, E., Conway, L., ... & Mensah, F. (2017). Language outcomes at 7 years: Early predictors and co-occurring difficulties. *Pediatrics*, *139*(3).

Mills, D., O'Riordan, C., Neophytou, E., & Roch, N. (publication expected summer 2024). A Welsh / English Bilingual Communicative Development Inventory (WEB_CDI): The importance of considering home language environment and socioeconomic status in developing norms. *Current Issues in Bilingualism. Special issue: Bilingual acquisition: Insights into language and cognition*.

Moore, G., Campbell, M., Copeland, L., Craig, P., Movsisyan, A., Hoddinott, P., ... & Evans, R. (2021). Adapting interventions to new contexts—the ADAPT guidance. *BMJ*, *374*.

Musa, S. M. K., Hassan, S. M., & Elrefaie, D. A. (2021). The impact of gender on the semantic skills of Arabic speaking Sudanese children. *Ain Shams Medical Journal*, *72*(4), 815-820.

Naigles, L. (2000). Manipulating the input: Studies in mental verb acquisition. In B. Landau, J. Sabini, J. Jonides, & E. Newport (Eds.), *Perception, cognition, and language. Essays in honor of Henry and Lila Gleitman* (pp. 245–274). Cambridge, MA: MIT Press.

Ohana, O., & Armon-Lotem, S. (2023). Parental reports on the lexicon of children from diverse bilingual populations. *Frontiers in Psychology*, *14*, 1123983.

Olmsted, J. C. (2005). Gender, aging, and the evolving Arab patriarchal contract. *Feminist Economics*, *11*(2), 53-78.

Osborne, J. W., & Costello, A. B. (2009). Getting the most from your analysis. *Pan*, *12*(2), 131-146.

O'Toole, C., & Fletcher, P. (2008). Developing assessment tools for bilingual and minority language acquisition. *Journal of Clinical Speech and Language Studies, 16*, 12-27.

Owens, J. (2013). *The Oxford handbook of Arabic linguistics*. Oxford Handbooks.

Özçalışkan, Ş., & Goldin-Meadow, S. (2010). Sex differences in language first appear in gesture. *Developmental Science*, *13*(5), 752-760.

Pallier, C., New, B., & Bourgin, J. (2019). Openlexicon, GitHub repository, https://github.com/chrplr/openlexicon

Perra, O., McGowan, J. E., Grunau, R. E., Doran, J. B., Craig, S., Johnston, L., ... & Alderdice, F. A. (2015). Parent ratings of child cognition and language compared with Bayley-III in preterm 3-year-olds. *Early Human Development*, *91*(3), 211-216.

Pérez-Pereira, M., & Resches, M. (2007). Elaboración de las formas breves del Inventario do Desenvolvemento de Habilidades Comunicativas. Datos normativos y propiedades psicométricas. *Infancia y Aprendizaje*, *30*(4), 565-588.

Petersen, I. T., & LeBeau, B. (2021). Language ability in the development of externalizing behavior problems in childhood. *Journal of Educational Psychology*, *113*(1), 68.

Reese, E., & Read, S. (2000). Predictive validity of the New Zealand MacArthur communicative development inventory: Words and sentences. *Journal of Child Language*, *27*(2), 255-266.

Reilly, S., Tomblin, B., Law, J., McKean, C., Mensah, F., Morgan, A., Goldfeld, S., Nicholson, J. and Wake, M. (2014). 'SLI: a convenient label for whom?', *IJLCD* 49(4), pp.416-451.

Reilly, S., Wake, M., Ukoumunne, O. C., Bavin, E., Prior, M., Cini, E., ... & Bretherton, L. (2010). Predicting language outcomes at 4 years of age: findings from Early Language in Victoria Study. *Pediatrics*, *126*(6), e1530-e1537.

Rice, M. L., & Hoffman, L. (2015). Predicting vocabulary growth in children with and without specific language impairment: A longitudinal study from 2; 6 to 21 years of age. *Journal of Speech, Language, and Hearing Research*, *58*(2), 345-359.

Rifaie, N., Hamza, T. M. A. W., & Elfiky, Y. H. (2021). Validity and reliability of the revised Arabic language test for 2–4-year-old children: Cross-sectional study. *The Egyptian Journal of Otolaryngology*, *37*, 1-8.

Rifaie, N., & Hassan, S. (2004). The Arabic language test-revised. *Benha Med J*, *21*, 205-216.

Rifaie, N., Kotby, M. N., Khairy, A., Baraka, M., & El-Shobary, A. (1994). The construction of an Arabic test to evaluate child language. *Unpublished PhD thesis, Ain Shams University, Cairo*.

Rudolph, J. M., & Leonard, L. B. (2016). Early language milestones and specific language impairment. *Journal of Early Intervention*, *38*(1), 41-58.

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., and Jorgensen, R. N. (2014). The Norwegian Communicative Development Inventories: reliability, main developmental trends and gender differences. *First Language, 34*, 3–23. doi: 10.1177/0142723713510997

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment. Is the low end of normal always appropriate? *Language, Speech and Hearing Services in Schools, 37*(1), 61-72.

Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. *International Journal of Bilingualism*, *15*(4), 426-445.

Tulviste, T., & Schults, A. (2020). Parental reports of communicative development at the age of 36 months: The Estonian CDI-III. *First Language*, *40*(1), 64-83.

Urm, A., & Tulviste, T. (2021). Toddlers' early communicative skills as assessed by the Short Form Version of the Estonian MacArthur-Bates Communicative Development Inventory II. *Journal of Speech, Language, and Hearing Research*, *64*(4), 1303-1315.

Vehkavuori, S.M. & Stolt, S. (2018) Screening language skills at 2; 0*. Infant Behavior and Development*, *50*, 174– 179

Versteegh, K. (2014). *Arabic Language*. Edinburgh University Press.

Vogt, P., Mastin, J. D., Knauff, M., Pauen, M., Sebanz, N., & Wachsmuth, I. (2013). Rural and urban differences in language socialization and early vocabulary development in Mozambique, *Proceedings of the 35th Annual Conference of the Cognitive Science Society (CogSci '13)*, July-August 2013, Austin, Tex, USA, Cognitive Science Society, 3787–3792.

Wareham, H., Huang, W., Labuschagne, L. & Law, J. (2021). The association between mother's education level and early child language skills; Findings from three European cohort studies. Publications archive – dynamicsofinequality.org

Wilson, P., Rush, R., Charlton, J., Gilroy, V., McKean, C., & Law, J. (2022). Universal language development screening: Comparative performance of two questionnaires. *BMJ Paediatrics Open, 6*(1), e001324. https://doi.org/10.1136/bmjpo-2021-001324

Figure Captions

Figure 1. The data from an additional 403 children were discarded as they did not live in the target countries (total 3084 children). Age range was from 7.5 months to 30.4 months. Dev delay: diagnosed developmental delay as reported by the parent. Hearing: hearing difficulties as reported by the parent. Multiling: more than 10% exposure to another language than Arabic. Diff dialect: if at least one of the parents uses another dialect than the country's dialect. Note that for Jordan, both Palestinian and Jordanian dialects were included. Father report: if the CDI was filled in by the father.

Figure 2. Current Egyptian dataset for Arabic CDI Toddler, with word comprehension (top) and production (bottom) scores as a function of age and maternal education.

Figure 3. Egyptian dataset from Abdelwahab et al. (2021) for the Arabic CDI Toddler, with word comprehension (top) and production (bottom) scores as a function of age and maternal education.

Figure 4. Norms from the Jordanian and Palestinian Arabic CDI Toddler dataset, for comprehension and production scores over age in months. Each line indicates a binomial model fit to the decile, where the solid black line indicates 50^{th} percentile, and the red one the 16^{th} percentile.

Figure 5. Arabic CDI Toddler data for Egypt, comprehension (top) and production (bottom) scores as a function of age and maternal education.

Figure 6. Norms from the Egyptian Arabic CDI Toddler dataset, for comprehension and production scores over age in months. Each line indicates a binomial model fit to the decile, where the solid black line indicates 50th percentile, and the red one the 16th percentile.

Figure 7. The data from an additional 102 children were discarded as they did not live in the target countries (total 2153 children). Age range was from 29.5 months to 48.4 months. Dev delay: diagnosed developmental delay as reported by the parent. Hearing: hearing difficulties as reported by the parent. Multiling: more than 10% exposure to another language than Arabic. Diff dialect: if at least one of the parents uses another dialect than the country's dialect. Note that for Jordan, both Palestinian and Jordanian dialects were included.

Figure 8. For the Arabic CDI Child and for each country, proportion of children who do not produce word combination according to parental report, per age bin (30 to 33 months, etc).

Figure 9. Arabic CDI Child, norms for productive vocabulary scores split into deciles for (A) Egypt, (B) Jordan, and (C) Palestine. Each line indicates a binomial model fit to the decile, where the solid black line indicates 50th percentile and the red one the 16th percentile.

Figure 10. Arabic CDI Child, norms for complex sentence use scores split into deciles for (A) Egypt, (B) Jordan, and (C) Palestine. Each line indicates a binomial model fit to the decile, where the solid black line indicates 50th percentile and the red one the 16th percentile.

Figure 11. Arabic CDI Child, norms for language use scores split into deciles for (A) Egypt, (B) Jordan, and (C) Palestine. Each line indicates a binomial model fit to the decile, where the solid black line indicates 50^{th} percentile and the red one the 16^{th} percentile.

Content of the Supplemental Material

Details of the population stratification process in Egypt, Jordan and Palestine (including Figures S1 to S3).

Summary of the main features of the language adaptations of the CDI-III, listed chronologically (Table S4)

Arabic CDI Toddler and Child