Journal of Child Language (2020), 1–25
doi:10.1017/S030500092000063X

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Testing the validity of the Cross-Linguistic Lexical Task as a measure of language proficiency in bilingual children

Elise VAN WONDEREN and Sharon UNSWORTH*

Centre for Language Studies, Radboud University, Nijmegen, Netherlands
*Address for correspondence: Centre for Language Studies, Radboud University, Postbus 9103, 6500HD Nijmegen, the Netherlands. E-mail: s.unsworth@let.ru.nl

(Received 11 December 2019; revised 20 May 2020; accepted 25 August 2020)

Abstract

The Cross-linguistic Lexical Task (CLT; Haman, Łuniewska & Pomiechowska, 2015) is a vocabulary task designed to enable cross-linguistic comparisons both across and within (bilingual) children. In this paper we assessed the validity of the CLT as a measure of language proficiency in bilingual children, by determining the extent to which (i) age-matched, monolingual Spanish-speaking and Dutch-speaking children obtained similar scores, (ii) the CLT correlated with other measures of language proficiency in monolingual and bilingual children, and (iii) whether the factors underlying the CLT's construction, i.e., target words' estimated Age of Acquisition and Complexity Index, were predictive of children's scores. Our results showed that, while the CLT correlated with other measures and is therefore a valid means of tapping into language proficiency, caution is required when using it to compare children's language proficiency cross-linguistically, as scores for Dutch-speaking and Spanish-speaking monolinguals sometimes differed.

Keywords: language proficiency; vocabulary task; child bilingualism; test validity

Introduction

A complete understanding of a bilingual child's linguistic development requires examining their skills in both languages. In certain circumstances, it is arguably essential to do so. For example, measuring children's proficiency in both languages is crucial for accurate diagnoses regarding language impairment to be made. Both underdiagnosis and overdiagnosis of language impairment are reported in the literature (Boerma, Chiat, Leseman, Timmermeister, Wijnen & Blom, 2015), indicating that language impairment is either overlooked – because language delays in the language of schooling are mistakenly ascribed to the child's bilingual status – or that bilingual children are erroneously diagnosed with language impairment because their proficiency in the other language has not been taken into account.

Measuring proficiency in both languages is also relevant for researchers of child bilingualism, where language proficiency is of interest both as a predictor and as an outcome. Relative language proficiency, i.e., how proficient children are in one language compared to the other, is often taken as a proxy for language dominance, and as such is hypothesized to (partly) explain the magnitude and direction of cross-linguistic influence (e.g., Yip & Matthews, 2000), and the direction and type of code-switching (e.g., Paradis & Nicoladis, 2007). The extent to which bilingual children are proficient in both of their languages has also been used as a predictor for the extent of cognitive advantages compared to monolingual children (e.g., Blom, Küntay, Messer, Verhagen & Leseman, 2014). Conversely, researchers have tried to explain children's (relative) language proficiency across development by looking at child-internal factors, such as age of acquisition, and child-external variables, such as amount of exposure (e.g., Bedore, Peña, Griffin & Hixon, 2016; Bedore, Peña, Summers, Boerger, Resendiz, Greene, Bohman & Gillam, 2012; Chondrogianni & Marinis, 2011; Hoff, Core, Place, Rumiche, Señor & Parra, 2012; Unsworth, Chondrogianni & Skarabela, 2018).

When trying to measure bilingual children's proficiency in both languages, researchers and clinicians are faced with a challenge: as for many of the (standardized) proficiency tasks, only a limited number of language versions exist (but cf. Bilingual Verbal Ability Tests, available in 18 languages; Muñoz-Sandoval, Cummins, Alvarado & Ruef, 2005). Furthermore, as different language versions of the same task are often direct translations of each other they may not be entirely comparable, because they do not always take into account cross-cultural differences and they may differ in complexity (Gathercole, Thomas & Hughes, 2008; Haitana, Pitama & Rucklidge, 2010; Peña, 2007).

Language proficiency is typically measured by tasks targeting (morpho)syntactic skills or vocabulary size. Vocabulary size has been shown to correlate with a range of language skills, including grammatical ability (e.g., Meara, 1996; Miralpeix & Muñoz, 2018) and, as such, vocabulary tests are often used as a proxy for general language proficiency. However, direct translations of vocabulary tests are often not entirely comparable, as so-called translation equivalents may differ in terms of frequency (e.g., *curve* is a high frequent word in English, but the Welsh translation *cromlin* has a very low frequency, Gathercole *et al.*, 2008), specificity (e.g., in Dutch only the verb *scheuren* "to rip" is used for ripping paper, whereas in Spanish you can also use the more general verb *romper* "to break"), number of synonyms (Dutch only uses the verb *aanbellen* for ringing the doorbell, whereas Spanish includes *llamar*, *llamar a la puerta/al timbre*, and *tocar/picar el timbre*), and morphological and phonological complexity (i.e., *vrachtwagen* in Dutch is a trisyllabic compound noun with two consonant clusters, whereas the English *truck* is a monosyllabic noun with only one consonant cluster).

To address some of these issues, Haman, Łuniewska and Pomiechowska (2015) developed the Cross-linguistic Lexical Task (CLT). The CLT is an attempt to create cross-linguistically comparable vocabulary tests for a large number of languages: there are 29 language versions readily available, and new language versions can be easily constructed in consultation with the original authors. As part of the LITMUS test battery (Armon-Lotem, Meir & de Jong, 2015), the CLT was originally created as a diagnostic tool to identify language impairment in multilingual populations. Its aim is to be fully comparable across languages, to enable comparison between bilingual and monolingual populations, between the two languages of a bilingual child, and between typically developing and language-impaired populations (Haman *et al.*, 2015). To this end, the construction of the CLT in each language is based on

Table 1. Variables included in the complexity index (CI). Table adapted from Hansen *et al.* (2017) such that loan word status is no longer excluded (Magdalena Łuniewska, p.c.).

	Measure	Contribution to CI
Phonology	Word length in phonemes	$2 * \frac{\text{no. of phonemes} - \text{mean}_{\textit{word class}}}{\text{SD}_{\textit{word class}}}$
	Word-initial fricative or affricate?	Yes = 1 point
	Word-initial consonant cluster?	Yes = 1 point
	Word-internal consonant clusters?	Yes = 1 point
Morphology	How many stems?	1 point per stem
	Is the word a derivation?	Yes = 1 point
	Prefix or suffix?	Yes, both = 2 points Yes, either = 1 point
Exposure	Is the object/action available to direct experience in your country?	No = 1 point
	How often would preschool children in your country have access to the object/action?	Not at all/rarely = 1 point Quite often = 0.5 points Very often = 0 points

two language-specific properties: an estimation of the target words' age of acquisition (AoA) and a composite measure of the target words' complexity.

In the present study we assess the validity of this task for comparing language proficiency across and within bilingual children, by investigating the performance of both monolingual and bilingual children on the Spanish and Dutch CLT. We test whether monolingual Dutch and Spanish children obtain comparable scores; whether CLT scores correlate with other measures of language proficiency; and, following Hansen, Simonsen, Łuniewska and Haman (2017), whether the factors underlying the construction of the CLT – i.e., word complexity and AoA – are predictive of children's scores.

Properties underlying the CLT

The CLT consists of a picture naming and a picture selection task, each further divided into a noun and verb subtask. There are thus four subtasks with 30 items in each. The construction of the CLT in each language is based on: (a) subjective estimations of AoA for each target word, and (b) a target word's complexity index (CI).

For each language, subjective AoA estimates were obtained by asking at least 20 native speakers to estimate the age at which they could understand each word, ranging from 0 (i.e., during the first year of life) to 18 (i.e., at age 18 or later) (Łuniewska, Haman, Armon-Lotem, Etenkowski, Southwood, Anđelković, Blom, Boerma, Chiat, de Abreu, Gagarina, Gavarro, Hakansson, Hickey, de Lopez, Marinis, Popovic, Thordardottir, Blaziene, Sanchenz, Dabasinskiene, Ege, Ehret, Fritsche, Gatt, Janssen, Kambanaros, Kapalkova, Kronqvist, Kunnari, Levorato, Nenonen, Fhlannchadha, O'Toole, Polišenská, Pomiechowska, Ringblom, Rinker, Roch, Savic, Slancova, Tsimpli & Ünal-Logacev, 2016a). This measure was deemed a valid measure of a word's AoA as the estimates correlated with child data on MacArthur-Bates Communicative Development Inventories (CDI) and with previous AoA estimates.

The CI is a composite measure consisting of both phonological and morphological aspects, and exposure to the depicted object or action (see Table 1). Phonological aspects have the largest impact on the CI score and, in particular, word length in phonemes (normalized within word class).

Based on these two measures, words are divided into four difficulty levels (early/late AoA, high/low CI). Target words are selected from a list of 158 nouns and 142 verbs, chosen for consistency in naming agreement across languages (Haman *et al*., 2015). From each difficulty level, 7 to 8 words are randomly selected as target words for both comprehension and production.

Previous findings of studies using the CLT

Studies using the CLT have found significant effects of age on children's performance (Haman, Łuniewska, Hansen, Simonsen, Chiat, Bjekić, Blažienė, Chyl, Dabašinskienė, Engel de Abreu, Gagarina, Gavarró, Håkansson, Harel, Holm, Kapalková, Kunnari, Levorato, Lindgren, Mieszkowska, Montes Salarich, Potgieter, Ribu, Ringblom, Rinker, Roch, Slančová, Southwood, Tedeschi, Tuncer, Ünal-Logacev, Vuksanović & Armon-Lotem, 2017), except when bilingual children were tested in the minority language (Bohnacker, Lindgren & Öztekin, 2016; Lindgren & Bohnacker, 2019), echoing previous findings showing that language development in the minority language is more dependent on exposure patterns than age (e.g., Hoff, Welsh, Place & Ribot, 2014). Accordingly, amount of – cumulative or current – exposure and dominance have both been found to predict bilingual children's CLT scores (Bohnacker *et al*., 2016; Gatt, Attard, Łuniewska & Haman, 2017; Potgieter & Southwood, 2016). Given enough variation in the sample, SES was also a significant predictor (Potgieter & Southwood, 2016; cf. Bohnacker *et al*., 2016).

In line with the frequently made observation that bilingual children may lag behind monolingual peers in at least one of their two languages on (expressive) vocabulary, some studies found significant differences between monolinguals and bilinguals on the CLT (Altman, Goldstein & Armon-Lotem, 2017; Hansen *et al*., 2017). At the same time, other studies have failed to find such a difference, perhaps because they included simultaneous (rather than sequential) bilingual children (Lindgren, 2018), or children with substantially more input in the language of testing than in their other language(s) (Potgieter & Southwood, 2016). Crucial for the validity of the CLT as a diagnostic tool, both quantitative and qualitative differences were found between typically-developing children and children with language impairment, both in monolingual and bilingual populations (Kapalková & Slančová, 2017; Khoury Aouad Saliby, Dos Santos, Kouba Hreich & Messarra, 2017).

Two other trends extensively reported in the literature on children's (early) lexical development are the precedence of nouns over verbs, and comprehension over production (e.g., Bornstein, Cote, Maital, Painter, Park, Pascual, Pêcheux, Ruel, Venuti & Vyt, 2004; Clark & Hecht, 1983; Gentner, 1982). Accordingly, many CLT studies have found higher accuracy for nouns over verbs, and for comprehension over production (e.g., Altman *et al*., 2017; Gatt *et al*., 2017; Haman *et al*., 2017).

Important for the validity of the CLT as a measure of vocabulary skills specifically, and overall language proficiency more generally, scores on the CLT have been found to correlate with parental report of children's overall language proficiency and with other standardized vocabulary tasks, although only consistently so for comprehension (Hansen, Łuniewska, Simonsen, Haman, Mieszkowska, Kołak & Wodniecka, 2019; Khoury Aouad Saliby *et al*., 2017). Note, however, that Khoury Aouad Saliby *et al*. (2017) used conceptual scoring for production, i.e., correct responses in either language were scored as correct; it is unclear how this may have affected the results.

Hansen *et al.* (2019) found moderate correlations between children's CLT scores and parental estimates of overall language proficiency, but only for the majority language, and not the minority language.

To investigate the validity of the task's construction, Hansen *et al.* (2017) investigated whether its two components, i.e., target words' AoA and CI, were robust predictors of children's scores. This was the case for AoA, for both monolinguals and bilinguals, but not for CI. Likewise, Haman *et al.* (2017) compared monolinguals' performance on the CLT across 17 languages, and found moderate-to-strong negative correlations with AoA for at least one of the four subtasks in each language (Haman *et al.*, 2017). In contrast, the CI revealed low-to-moderate negative correlations with children's scores for only one or two subtasks in seven of the languages and no significant correlations in the other 10 languages.

Despite little or no relation between children's scores and CI, Hansen *et al.* (2017) argued that the robust effect of AoA may nevertheless be enough to render cross-linguistically comparable CLT versions. Their observation that Polish and Norwegian age-matched monolingual children scored similarly supports this claim. Comparing monolingual children's performance across 17 languages, Haman *et al.* (2017) found that only isiXhosa-speaking children knew significantly fewer words than the rest, most likely due to their comparatively low SES. When this group was removed from the analysis, a significant effect of language remained. This effect was rather small, and, as suggested by the authors, may have been caused by differences across languages in sample size and age range.

In sum, studies using the CLT have been able to replicate numerous findings from the literature, including the asymmetry for production versus comprehension, and nouns versus verbs, as well as effects of age, SES and relative amount of input on (bilingual) children's vocabulary scores. These studies have also found, however, that the rationale behind the CLT's construction is partially undermined by the lack of a reliable effect for the complexity index. Nonetheless, the CLT's construction procedure may still be successful in creating cross-linguistically comparable vocabulary tasks, as similar scores have been obtained for age-matched monolingual children, at least for Polish and Norwegian. Whether this holds for other language pairs, and whether the CLT correlates with other measures of children's language skills, remains an open question.

The present study

In this paper we report two studies investigating performance on the CLT by monolingual Dutch and Spanish children, and by bilingual Spanish–Dutch children, between the ages of 3 and 9. We chose a wide age range to assess the suitability of the CLT in and beyond the pre-school years. We investigate the validity of the CLT by (a) comparing the scores of the two monolingual groups, (b) comparing children's performance on the CLT with their performance on other measures of language proficiency, and (c) investigating the effects of the two factors underlying the CLT's construction, i.e., the target words' estimated AoA and CI.

Haman *et al.* (2017) suggested that phonological complexity may exert less influence on children's lexical development when they are older; for this reason, we also investigate whether there is an interaction between the CI and children's age at testing. In addition, we might expect an interaction between target words' estimated AoA and children's age at testing: since most of the target words of the CLT have an estimated AoA lower than 6 (Łuniewska et al., 2016a), AoA should matter less for children beyond this age.

Study 1

Method

Participants

In the first study we tested 32 monolingual Dutch-speaking children (14 girls, 18 boys) and 32 monolingual Spanish-speaking children (17 girls, 15 boys), between the ages of 4;0 and 9;7. One additional Spanish-speaking child was tested, but later excluded because of exposure to English. The Dutch-speaking children were sampled from a public primary school in Gelderland, a province in the eastern part of the Netherlands, and the Spanish-speaking children were sampled from one public and one private primary school in Granada, in southern Spain. Ethical approval was obtained following the standard procedures in each location. All parents provided written informed consent prior to testing. The two groups were matched on age and parental education (see Table 2). Parental education was very high in the current sample, with almost all of the parents holding a bachelor's or master's degree.

Materials

CLT

The construction of the Dutch CLT was finalized in 2017 in consultation with the CLT developers (van Wonderen, Blom, Boerma, Janssen, Unsworth & van Dijk, 2017). For the current study, we adapted the Spanish production part of the CLT based on pilot data of monolingual six-year-olds (Myriam Cantú Sánchez, personal communication). Ten words that did not consistently evoke the target word in these pilot data, or constituted a phonologically identical cognate between Dutch and Spanish, were replaced by other words from the same difficulty level.

Although the CLT was originally intended for use with preschoolers, it has been used for monolingual children up to the age of seven (Haman *et al.*, 2017) and for bilingual children up to a mean age of eight (Ringblom & Dobrova, 2019). Because the present study includes children beyond that age, we decided to add 10 nouns and 10 verbs from the highest difficulty level in an attempt to prevent ceiling effects. Because there was only a limited number of words in the highest difficulty level – and the comprehension subtask requires the additional selection of three distractors per target item – we were only able to do so for the production subtasks. Thus, the total number of target items was 80 for production and 60 for comprehension. To ensure that none of the results presented in this paper were specific to the 40-item version of the production subtask, we additionally ran all analyses on the first 30 items only.

Sentence repetition

As a second measure of children's language proficiency, we included a sentence repetition task (SRT), designed to tap into children's overall language proficiency in both comprehension and production (Marinis & Armon-Lotem, 2015), and shown to be particularly sensitive to children's lexical and morphosyntactic knowledge (Polišenská, 2011).

The SRT used in this study was constructed in such a way that both language versions contained shared and language-specific structures, and that sentence complexity varied enough to accommodate the wide age range investigated. To achieve this, we selected 30 sentences in total from the Repeating Sentences subtasks

Table 2. Monolingual children's age (years;months) and parental education (means and SDs).

	Dutch		Spanish			
	M (SD)	Range	M (SD)	Range	Test estimate	*p*-value
Age	6;5 (1;8)	4;0–9;7	6;4 (1;8)	4;1–9;1	$t(62) = -0.03$	.98
Maternal education[a]	4.0 (0.5)	2–5	3.9 (0.4)	2–4	$\chi^2(3) = 4.28$	.23
Paternal education[a]	3.9 (0.6)	2–5	3.7 (1.0)	1–5	$\chi^2(4) = 7.69$	.10

[a]Level of education was measured on a 5-point scale: 1 = primary education, 2 = secondary education, 3 = post-secondary non-tertiary education, 4 = bachelor or master's degree or equivalent, 5 = doctoral degree or equivalent.

of the Spanish and Dutch CELF-Preschool-2 (Wiig, Secord & Semel, 2009; Wiig, Secord, Semel & De Jong, 2012), the CELF-4 (Wiig, Secord & Semel, 2006; Kort, Schittekatte & Compaan, 2010) and from the Spanish and Dutch LITMUS-SRTs (Marinis & Armon-Lotem, 2015). Care was taken to ensure that both language versions had an equal number of relatively simple and relatively complex sentences, and that both versions were comparable in terms of average sentence length (mean number of words was 9.0 ($SD = 1.8$) for Spanish and 9.6 ($SD = 2.3$) for Dutch, $t(58) = 1.12$, $p = .27$). See the Supplementary Materials (Supplementary Materials) for all items included in the SRTs.

Procedure

Children were tested individually in a separate, quiet room at school. The CLT was always administered before the SRT, to prevent children from hearing words in the SRT that they would have to produce in the CLT. For the CLT, the order of the production and comprehension tasks was counterbalanced across children. For the picture selection task, the experimenter would ask "Where is a [NOUN]?" or "Who is [VERB]?" and then children were instructed to point to – or say the number of – the corresponding picture. For the picture naming task, the experimenter would ask "What is this?" for nouns, and "What is the [boy/girl/woman/man] doing?" or "What is happening to the [NOUN]?" for verbs. When children were unclear about what was depicted in the picture they were asked to name, the experimenter tried to briefly describe it (e.g., for *scales* the experimenter would say "this can show you how heavy something is"). If children gave an incorrect answer, the experimenter would ask if they knew another word for the picture before continuing with the next picture.

For the SRT, sentences were pre-recorded by a female native speaker and inserted in the PowerPoint developed for the LITMUS-SRT (Marinis & Armon-Lotem, 2015). Children wore headphones and were told they would go on a treasure hunt with Teddy the bear. In order for Teddy to move a step forward they had to repeat everything Teddy said. Children's responses were recorded with a digital voice recorder. Children were praised for trying to repeat the sentence, regardless of how well they did. When a child could not hear the sentence because of interruptions or loud noises, the sentence was played one more time after completion of the task. After each task, children could select a sticker. In total, testing took approximately 30 minutes.

Scoring

For the CLT, children's responses were noted down on the answer sheet by the experimenter.[1] The original scoring guidelines for production (see Kapalková & Slančová, 2017) were deemed too generous because all responses were scored as correct as long as they contained the target's stem (i.e., including innovations, changes of word class, and one-stem answers for compounds, e.g. "sewing" for "sewing machine"); therefore, the Uppsala guidelines developed by Ute Bohnacker and colleagues were used instead (Bohnacker *et al.*, 2016). This meant that an answer was scored as correct when it contained the target word or a (regional) synonym of the target word. Alternative answers were scored as correct when they corresponded with the picture and were at least as specific as the target word (e.g., *peddelen* "to paddle" for *roeien* "to row"). Mispronunciations were allowed if the word was still recognizable as the target (e.g., *lipstip* for *lippenstift* "lipstick"). For Spanish, answers containing a wrong determiner were counted as correct (e.g., *ver el tele* instead of *ver la tele* "watch television").

Incorrect answers included words that were too general (e.g., *limpiar* "to clean" instead of *aspirar* "to vacuum"), words from the other language, innovations (e.g., *zager* "saw+suffix" instead of *zaag* "saw"), and words from a different word class (e.g., a noun in the verb subtask and vice versa; an exception was made in the case of periphrastic verbs with the verb *hacer* "to make/do" in Spanish; as this verb is already included in the elicitation question, *¿Qué está haciendo?* "What is s/he doing?", it is pragmatically correct to answer with only the noun). For Spanish, periphrastic verbs containing a preposition were scored as incorrect if the child used the wrong preposition (e.g. *hablar con el teléfono* "to talk with the phone" instead of *hablar por teléfono* "to talk on the phone"). For a complete list of accepted alternative answers, see the Supplementary Materials (Supplementary Materials).

For the analysis investigating whether the CI and AoA are predictive of children's scores, we only coded target responses as correct (i.e., disregarding synonyms or alternative correct responses), as the CI and AoA values are based on the target word only.

For the SRT, all children's responses were transcribed by (near-)native research assistants and these transcriptions were checked by the first author using the recordings. Responses were scored as correct if the sentence was (a) grammatical, AND (b) contained the target structure. Semantic anomalies or incomplete sentences were scored as incorrect. For Dutch, grammatical gender errors resulting in the wrong determiner (*de* instead of *het* or vice versa), demonstrative pronoun (*deze/die* instead of *dit/dat* or vice versa) or relative pronoun (*die* instead of *dat* or vice versa) were allowed, as gender is acquired relatively late by monolingual Dutch children (e.g., Van der Velde, 2004). For Spanish, dative constructions without a clitic (*Di un regalo a mi mamá* vs. *LE di un regalo a mi mamá* "I gave a present to my mom") were accepted, as the Real Academia Española (RAE, 2005) – a largely prescriptive language body and the official authority on the Spanish language in the Spanish-speaking world – considers it grammatical (though infrequent).

Responses were scored by two different coders in each language, who reached 94.3% agreement for Dutch and 91.4% for Spanish. Cohen's kappa for interrater reliability,

[1] For the Spanish comprehension subtask, there was one item for which two responses were eventually scored as correct, as in the original Spanish CLT the target *hundir* "to sink" is accompanied by a distractor picture depicting the verb "to drown". As these pictures are very similar, both answers were accepted.

which controls for the agreement expected by chance, was 0.88 and 0.79, respectively, which is high. In the case of disagreements, a third coder functioned as arbiter.

Analyses

To assess whether CI and AoA were predictive of children's scores (correct vs. incorrect), we performed Bayesian generalized linear mixed models in R (R Core Team, 2018) using the brms package (version 2.8.6; Bürkner, 2017). While in most cases Bayesian and frequentist analyses lead to very similar conclusions when relatively uninformative priors are used (Albers, Kiers & van Ravenzwaaij, 2018), a Bayesian approach has the advantage that it provides the entire posterior distribution with which we can say something about the (un)certainty of our estimate and what the 95% most credible values are. In other words, the claim that there is a 95% chance the true estimate is in the credible interval is warranted on a Bayesian approach but not on frequentist approaches (often leading to misinterpretations as researchers interpret frequentist confidence intervals as Bayesian credible intervals; see e.g., Kruschke & Liddell, 2018; Morey, Hoekstra, Rouder, Lee & Wagenmakers, 2016).

The models included CI, AoA, word class (nouns vs. verbs), modality (comprehension vs. production), and children's age as fixed effects, as well as the interactions of CI and AoA with age, and the interactions between all previously mentioned predictors and group (Spanish vs. Dutch monolinguals). Factors were coded with sum-to-zero contrasts, comparing level estimates to the grand mean. Children's age was mean-centered, and CI and AoA were standardized. The model used a maximal random effects structure as recommended by Barr, Levy, Scheepers, and Tily, (2013), including random intercepts for participants and items, by-item and by-participant random slopes, and the correlations between by-item random effects. Correlations between by-participant random effects were removed as they were all close to zero and unnecessarily complicated the model. The model was fit using four chains, with 8,000 iterations each (2,000 warm-up). We used brms' weakly informative default priors, and coefficients were deemed statistically "significant" if the associated 95% credible intervals were non-overlapping with zero, thus indicating a 95% chance that the effect of a predictor was different from zero. For factors, estimated marginal means were computed using the emmeans package (Lenth, 2019) to directly compare the factor levels to each other. Interactions were explored by follow-up models for the relevant subsets of the data.

To check for undue influence from items with no variation, we also ran the models on a reduced dataset excluding items on which all children scored correct (production: $n_{\text{Dutch}} = 17/80$, $n_{\text{Spanish}} = 17/78$;[2] comprehension: $n_{\text{Dutch}} = 39/60$, $n_{\text{Spanish}} = 46/59$[2]) or incorrect (production: $n_{\text{Spanish}} = 2/78$), according to the strict scoring scheme. The pattern of results remained the same.

Results

CI and AoA in the Dutch and Spanish CLT

Although the CLT is constructed in such a way that each language version contains an equal number of items from each difficulty level, it still may be the case that the versions

[2] There were three words in Spanish for which no AoA or CI values were available, so these items were already excluded from the models.

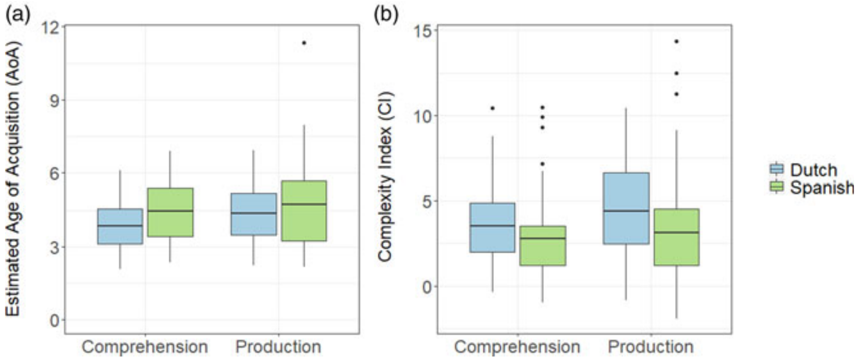

Figure 1. Boxplots of (a) AoA, and (b) CI, for the production and comprehension subtasks of the Dutch and Spanish CLT.

Table 3. AoA and CI in the Dutch and Spanish CLT (means and SDs).

		Dutch		Spanish			
	subtask	*M*	*SD*	*M*	*SD*	t^a	*p*
CI	Prod	4.44	2.65	3.35	3.04	2.40	.017
	Comp	3.72	2.38	2.80	2.45	2.09	.039
AoA	Prod	4.38	1.07	4.71	1.66	-1.48	.142
	Comp	2.80	2.45	3.72	2.38	-2.89	.004

[a] *t*-tests with Welch-adjusted degrees of freedom are reported to correct for unequal variance.

differ in the mean AoA and CI of the items. For this reason, before examining children's responses, we compared the two CLT versions on these two variables (see Figure 1).

Analyses revealed that there were indeed some small differences between the two language versions in terms of mean CI and AoA (see Table 3). Mean CI differed for both subtasks, with a higher mean CI for Dutch than for Spanish in both the production and the comprehension subtask. In contrast, whilst mean AoA differed for the two languages in the comprehension subtask, with a higher mean AoA for Spanish than for Dutch, mean AoA in the production subtask did not.

Monolingual children's performance on the CLT

Figure 2 presents children's scores on each subtask for each age group and language. In the Dutch group, there were 15 children aged 4 to 5, 10 children aged 6 to 7 and 7 children aged 8 to 9; for Spanish, the numbers per age group were 16, 8 and 8, respectively.

Children performed at ceiling on the comprehension subtask, although there was still some variation in the youngest age group, especially on verbs. When looking at production, we see that children's scores increased – and variability decreased – across age groups, but none were at ceiling.

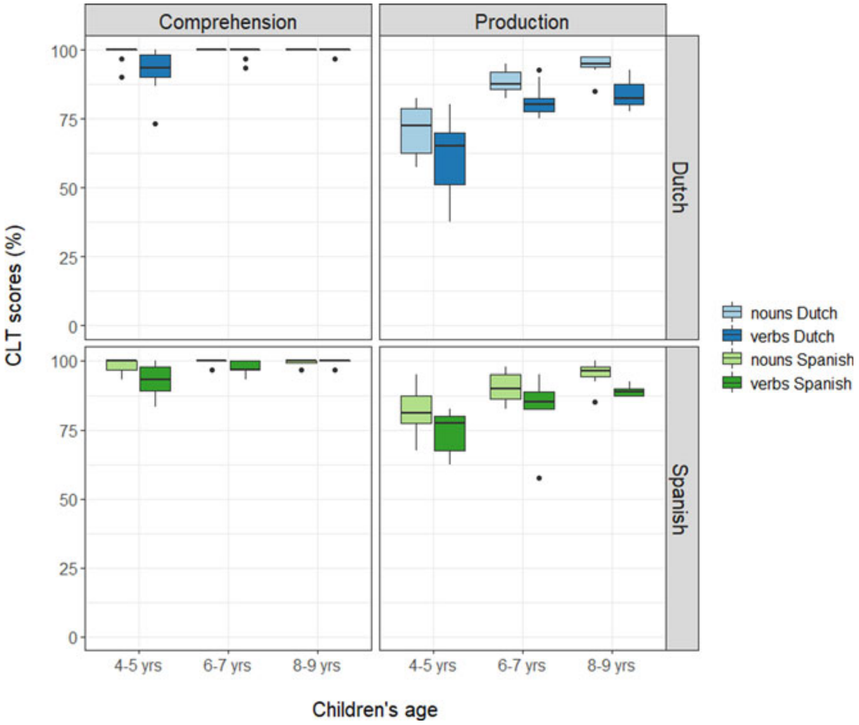

Figure 2. Boxplots of monolingual children's CLT scores divided into three age groups.

Table 4. Monolingual children's scores on the CLT and SRT (percentages).

		Dutch			Spanish				
	#item	*M*	*SD*	*Range*	*M*	*SD*	*Range*	*t*[a]	*p*
CLT prod	80	77.0	12.9	50.0–95.0	83.7	8.7	66.3–95.0	2.44	.018
Nouns	40	81.6	12.2	57.5–97.5	87.2	8.6	67.5–100	2.14	.037
Verbs	40	72.3	14.5	37.5–92.5	80.1	9.8	57.5–95.0	2.51	.015
CLT comp	60	97.9	3.3	86.7–100	97.4	2.9	90.0–100	-0.66	.510
nouns	30	99.6	1.9	90.0–100	98.9	1.8	93.3–100	-1.58	.119
verbs	30	96.3	6.0	73.3–100	95.9	4.8	83.3–100	-0.23	.819
SRT	30	61.5	28.8	10.0–100	70.2	26.6	13.6–100	1.27	.211

[a] *t*-tests with Welch-adjusted degrees of freedom are reported to correct for unequal variance.

Comparing Spanish and Dutch monolingual children's CLT and SRT scores

Children's scores on the CLT and the SRT are presented in Table 4, along with independent *t*-tests comparing the mean scores of the Dutch and Spanish monolingual children.

When comparing the mean scores of the two monolingual groups, we found that the Spanish group significantly outperformed the Dutch group for the CLT production subtask, scoring on average 6.7% higher (which is equivalent to 2 to 3 of the 40 items). The two groups did not significantly differ in their comprehension scores, nor in their performance on the SRT.

The difference between the two groups for the production subtask of the CLT could not be attributed to the higher mean CI of the items in the Dutch as compared to the Spanish production subtask, as a logistic regression analysis in which CI was controlled for still revealed a significant effect of group, with the odds of answering correctly being 2.40 (95% *CI* = 2.03, 2.83) times higher for the Spanish than for the Dutch monolinguals, $z = 10.3$, $p < .001$ (see Supplementary Materials, Supplementary Materials).

When we divide the data into three different age groups (4–5 year-olds, 6–7 year-olds, and 8–9 year olds), we found that the differences between the two monolingual groups for the production subtask were almost entirely driven by the youngest age group, with the Spanish-speaking children scoring 12.1% higher than the Dutch-speaking children (equivalent to 4 to 5 of the 40 items), whereas there were no significant differences for the older age groups (all *p*'s > .05, see Supplementary Materials, Supplementary Materials).

Finally, we checked whether children's scores differed for the first thirty items (i.e., the official CLT), as compared to the final ten items that were added for this study and which were all from the highest difficulty level. As expected, children indeed scored lower on the final ten items (Dutch: $M = 70.0\%$, $SD = 19.1$, Spanish: $M = 65.3\%$, $SD = 19.4$), than on the first thirty items (Dutch: $M = 79.3\%$, $SD = 11.2$, Spanish: $M = 89.8\%$, $SD = 5.9$). Looking at these averages, one may also note that whilst the Spanish-speaking children as a group outperformed the Dutch-speaking children on the first thirty items, the Dutch-speaking children numerically outperformed the Spanish-speaking children on the last ten items. This means that the overall difference between the two groups reported above (cf. Table 4) was slightly attenuated by the addition of the ten extra items in this study (for the 30-item version, the Spanish-speaking children outperformed the Dutch-speaking children with 10.5% on average, equivalent to 3 to 4 of the 30 items; also compare Tables S5 and S6 in Supplementary Materials, Supplementary Materials).

Correlations between the CLT and the SRT

While controlling for children's age, we found moderate correlations between children's SRT scores and their performance on the CLT's production and comprehension subtasks, for both Dutch (production: $r(30) = .62$, $p < .001$; comprehension: $r(30) = .51$, $p = .003$) and Spanish (production: $r(30) = .74$, $p < .001$; comprehension: $r(30) = .52$, $p = .003$).

Predictive power of CI and AoA

Estimated AoA was a strong predictor of children's scores, with the odds of children scoring correctly being 7.57 times lower for each standard deviation increase in AoA, 95% credible interval (95% *CrI*) [5.05, 11.66] (as AoA has a negative relationship with children's scores, the inverse odds ratio is reported for AoA throughout this paper for ease of interpretation). CI, on the other hand, was not a significant predictor of children's scores, odds ratio (OR) = 1.27, 95% *CrI* [0.88, 1.83]. Other significant predictors of children's scores included children's age, with the odds of a correct response becoming 1.05 times higher for a one month increase in age (95% *CrI* = 1.04, 1.07), subtask (nouns > verbs, *OR* = 3.20, 95% *CrI* = 1.38, 5.74), and modality (comprehension > production, *OR* = 48.2, 95% *CrI* = 21.2, 89.9). There were no significant interactions with group, and, crucial to our research questions, there were no interactions between age and CI (*OR* = 1.00, 95% *CrI* = 0.99, 1.01), nor age and AoA (*OR* = 1.01, 95% *CrI* = 0.996, 1.018).

When removing all items with zero variance (121/277 items), the pattern of results remained the same in the sense that AoA was still a strong predictor of children's scores (*OR* = 3.05, 95% *CrI* = 2.15, 4.41) whilst CI was not (*OR* = 1.28, 95% *CrI* = 0.90, 1.82), and there were no interactions with children's age. The same held for the analysis with the first thirty items only (AoA: *OR* = 6.85, 95% *CrI* = 4.46, 11.01; CI: *OR* = 1.08, 95% *CrI* = 0.71, 1.66).

To see if any of these results differed for production and comprehension, we ran a separate model including all interactions with modality. The only significant interaction was with CI, showing that CI did not have an effect on the production scores (*OR* = 1.10, 95% *CrI* = 0.70, 1.72), whereas an increase in CI seemed to have a positive, rather than a negative, effect on the comprehension scores (*OR* = 5.28, 95% *CrI* = 1.78, 22.05). However, the minimal variation on children's comprehension scores makes any estimate very sensitive to item- or participant-related noise, as there were very few items for which accuracy was NOT close to 100%. Because the results for comprehension may prove unreliable due to these ceiling effects, we also ran a model on the production data only. The pattern of results was the same as for the analysis on the complete dataset reported above.

Discussion

The findings so far suggest that the CLT is indeed indicative of children's language proficiency, as the correlation with the SRT suggests that both tasks are tapping into the same underlying construct. Furthermore, the CLT's construction was partly validated as AoA was a strong predictor of children's scores, in line with previous findings (e.g., Haman *et al.*, 2017). In contrast, CI did not seem to be predictive of children's scores.

We also found that, whereas there was still some variability in comprehension scores in the youngest age group, the older age groups performed at ceiling. This was not the case for production, where even the oldest age group did not perform at ceiling. This was likely in part due to the addition of the ten extra items, as scores were on average lower for these than the first thirty items. At the same time, it is worth noting that the absence of ceiling effects could partly be explained by the effect of specific items: for Dutch, there were three verbs that were correctly produced by 2 or 3 children only (viz. *liften* "to hitchhike", *bedelen* "to beg", and *vijlen* "to polish"), and for Spanish there were two verbs that only 5 children named correctly (viz. *taladrar* "to drill", and *operar* "to operate").

The finding that the monolingual Spanish-speaking children scored significantly higher on the production subtask than the monolingual Dutch-speaking children was unexpected, and raises the question whether this can be explained by differences induced by the sample or by the task. In other words, the two groups may have differed on some other factor than age or parental education, such as (non)verbal intelligence, and this may have caused the difference in scores. Without any data specifically targeting such variables, we cannot rule out sample-related differences for certain. However, the fact that children's scores did not significantly differ for the SRT makes this explanation less likely. At the same time, mean AoA did not significantly differ for the Spanish and Dutch production subtask of the CLT, and the higher mean CI for the Dutch task could also not explain the difference, as controlling for CI did not change the group effect. If, for some reason, the Dutch production subtask is more difficult than the Spanish one, it is as yet unclear why this may be so. It does however raise some doubts about the validity of the CLT, which we consider in the General Discussion.

In sum, the correlation between the CLT and SRT seems to suggest that the CLT is indeed tapping into children's language proficiency, but the significant difference between the Dutch and Spanish monolinguals for the production subtask raises doubts about the validity of the CLT for comparing language proficiency across and – in the case of bilinguals – within children. In the second study we were again interested in the correlation between the CLT and other measures of language proficiency, as well as whether the CI and AoA would be predictive of children's scores. In this second study we tested a group of Spanish–Dutch bilingual children in the same age range as the children tested in the first study. Given the ceiling effects and very limited amount of variance on the comprehension subtask in the monolingual children, we decided to only administer the production subtask of the CLT. A subset of the bilingual children completed an additional, standardized language test targeting morphology and morpho-syntax as a third measure of their language proficiency.

Study 2

Method

Participants

We tested 54 Spanish–Dutch bilingual children (28 girls, 26 boys), between the ages of 3;5 and 9;11. The bilingual children tested for this study came from two different samples ($n_1 = 35$, $n_2 = 19$), and were matched with the monolingual groups from the first study on age ($F(2, 115) = 0.04$, $p = .97$) and parental education (maternal: $\chi^2(6) = 9.65$, $p = .14$, paternal: $\chi^2(8) = 15.14$, $p = .06$), see Table 5.

Participants were recruited via expat groups on Facebook and by word of mouth. Children were included in the study if they (a) had been in regular contact with both Dutch and Spanish prior to school entry (i.e., before the age of 4), (b) did not have contact with a third language, and (c) did not have a history of language disorders, hearing problems, or attention or learning problems. Eight additional children were initially tested but later excluded because they did not complete all the tasks in one or both languages, either because their knowledge of Spanish was too passive to be able to complete the tasks ($n = 5$), or because they lost concentration during the session ($n = 3$).

All children lived in the Netherlands at the time of testing and attended schools or day care centres where Dutch was the language of instruction and communication. The vast majority of the children ($n = 46$) came from mixed families, in which one of the parents was a native speaker of Dutch and the other of Spanish, each speaking their native language to varying degrees. In one of these families, Spanish input came from an au-pair and the Spanish-speaking father was reported to speak only Dutch to his child. The remaining eight children came from families in which both parents were native speakers of Spanish, in which case the children had attended a Dutch day care before the age of 4.

Table 5. Bilingual children's age (years;months), relative exposure to Dutch, and relative use of Dutch in the home (%), and parental education (means and SDs).

	M	*SD*	*Range*
Age	6;4	1;6	3;5–9;11
Exposure to Dutch at home	49	23	0–87
Use of Dutch at home	61	28	0–100
Maternal education[a]	4.1	0.6	2–5
Paternal education[a]	4.0	0.6	2–5

[a] Level of education was measured on a 5-point scale: 1 = primary education, 2 = secondary education, 3 = post-secondary non-tertiary education, 4 = bachelor or master's degree or equivalent, 5 = doctoral degree or equivalent.

The Spanish-speaking parents in the study came from various countries of origin, including Spain and several Latin-American countries. Information on children's language experience was collected via a parental questionnaire (BiLEC, Unsworth, 2013). There was a wide range in how much input children received in each language at home (see Table 5). On average, and consistent with the fact that Dutch was the majority language, children spoke more Dutch than they were spoken to.

Materials

The bilingual children also completed the SRT, which for Spanish was the same task as described in the first study. For Dutch, the same task as in the first study was used for the first sample of bilingual children ($n = 35$), whereas for the second sample ($n = 19$), the LITMUS-SRT was used (see the Supplementary Materials, Supplementary Materials for a list of all target sentences for both SRTs). Because these two tasks were not exactly the same, we will report results for the SRT for the two samples separately.

The first sample of bilingual children also completed the Word Structure (WS) subtasks of the Spanish and Dutch Clinical Evaluation of Language Fundamentals 4 (CELF-4; Wiig *et al.*, 2006; Kort *et al.*, 2010). These tasks consist of 29 and 30 items, respectively, eliciting words and phrases to assess lexical, derivational and inflectional morphology and morphosyntax, and were used in this study as a third – standardised – measure of the children's language proficiency.

Both groups of bilingual children additionally completed a Digit Span task to measure non-verbal working memory, as part of a larger test battery.

Procedure

Children were tested at home by (near-)native research assistants on two separate occasions: once in Dutch and once in Spanish. Sixteen children completed the Spanish WS task for a second time during a third session, between 5 and 8 weeks after the original session. This was deemed necessary because some of the items (n = 3–17 out of 29) had not been elicited – or elicited incorrectly – during the first session, thus yielding unreliable and incomparable results. The fact that these children completed the Spanish WS task twice was unlikely to have affected their final test scores, as for the items that were elicited correctly during the first session a paired t-test did not show any significant difference between children's scores on the two test moments ($M_1 = 9.8$, $SD_1 = 5.8$; $M_2 = 10.5$, $SD_2 = 5.5$), $t(15) = -1.02$, $p = .33$.

Thirty-eight of the children were tested by a different experimenter in each language, whereas the other sixteen children were tested in both languages by the first author. Order of test language was counterbalanced across children. Testing took 45 minutes to one hour, and completing the questionnaire with one of the parents took an additional 20 minutes.

Scoring

The CLT was scored as described for the first study. The WS task was scored during testing according to the guidelines given in the CELF manual. For the SRT, responses were transcribed by (near-)native research assistants and checked by the first author. Responses were scored by two different coders for each language, who reached 96.7% agreement for Dutch and 94.1% for Spanish. Cohen's kappa was 0.92 and 0.87, respectively, which is very high. In the case of disagreements, a third coder functioned as arbiter.

Analyses

We again performed Bayesian generalized linear mixed models to investigate whether the CI and AoA were predictive of children's scores. The models had the same specifications as in the first study, except that for these analyses we added proportion of Dutch input at home as a fixed effect, and language (Dutch vs. Spanish) was now a within-subjects factor. Estimated marginal means were again computed to directly compare factor levels to each other, and separate models were run for each language to explore interaction effects that involved language. Following the strict scoring scheme, there were again some items with zero variance (all correct: $n_{\text{Dutch}} = 7/80$; $n_{\text{Spanish}} = 2/78$; all incorrect: $n_{\text{Spanish}} = 5/78$), and we thus ran additional analyses without these items. This did not change the pattern of results.

Results

Bilingual children's CLT scores

Children's scores on each subtask are provided separately for each age group and language in Figure 3. There were 23 children aged 4 to 5, 22 children aged 6 to 7, and 7 children aged 8 to 9.

As shown in Figure 3, the bilingual children showed a similar pattern as the monolingual children for Dutch, in the sense that children's scores increased, and variability decreased, across age groups. No such pattern was observed for Spanish, where there was great variability within each age group, and no observable differences in mean scores.

Comparison with other language proficiency measures

Children's scores on the CLT, SRT and CELF-WS task are provided in Table 6.

While controlling for children's age, we found moderate correlations between children's CLT scores and the SRT, both in Spanish ($r(48) = .74$, $p < .001$), and for both SRTs in Dutch (SRT: $r(30) = .62$, $p < .001$; LITMUS-SRT: $r(17) = .66$, $p = .003$). Likewise, moderate-to-strong correlations were found between the CLT and the standardized CELF-WS task in both Spanish ($r(30) = .83$, $p < .001$), and Dutch ($r(30) = .65$, $p < .001$).

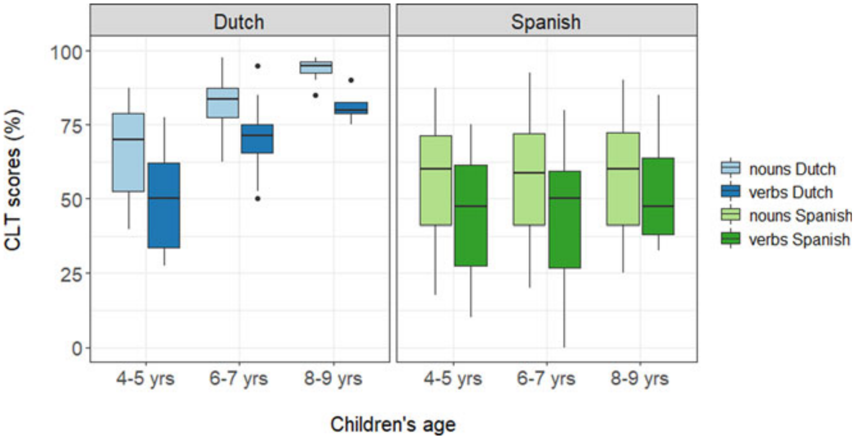

Figure 3. Boxplots of bilingual children's scores on the production subtasks of the CLT divided into three age groups. The two three-year-olds are not presented in this graph.

Table 6. Bilingual children's scores (%) on the language proficiency tasks. Number of items was 40 per subtask for the CLT (production only), 30 for both SRTs, and 30 and 29 for the Dutch and Spanish CELF-WS task, respectively.

	Dutch				Spanish			
	#child	*M*	*SD*	*Range*	*#child*	*M*	*SD*	*Range*
CLT	54	68.2	17.3	26.3–93.8	54	50.8	19.3	10.0–87.5
Nouns	54	75.5	16.3	37.5–97.5	54	55.4	19.2	17.5–92.5
Verbs	54	61.0	19.5	15.0–95.0	54	46.2	20.4	0.0–85.0
SRT	32	62.0	31.4	3.3–100	50	35.4	26.0	0.0–93.3
LITMUS-SRT	19	77.9	19.5	26.7–100	-	-	-	-
CELF-WS	32	67.0	23.8	16.7–96.7	32	47.4	21.2	17.2–92.9

Predictive power of CI and AoA

Estimated AoA was a strong predictor of children's scores, with the odds of children scoring correctly being 10.41 times lower for each standard deviation increase in AoA, 95% *CrI* [7.21, 15.39]. CI, on the other hand, was not a significant predictor of children's scores (*OR* = 0.99, 95% *CrI* = 0.72, 1.38), and a small interaction with children's age (*OR* = 0.99, 95% *CrI* = 0.987, 0.999) showed a counterintuitive pattern: whereas there was no effect of CI in the older children, the younger children's scores slightly increased for words with a higher, rather than a lower, CI value.

Other significant predictors included the interaction between children's age and language, with an increase in odds of 1.07 (95% *CrI* = 1.05, 1.09) in Dutch for a one month increase in children's age, and no significant increase with age for Spanish (*OR* = 1.02, 95% *CrI* = 0.99, 1.04). Proportion of Dutch input at home was also a significant predictor in both languages, with the odds of scoring correctly being 1.56 times higher in Dutch (95% *CrI* = 1.22, 2.00), and 2.09 times lower in Spanish (95%

CrI = 1.36, 3.24), for each standard deviation increase in the proportion of Dutch input (i.e., 23%). The only other significant predictor was subtask, with children scoring higher on nouns than on verbs (*OR* = 2.17, 95% *CrI* = 1.09, 3.68). There was no interaction between children's age and AoA (*OR* = 1.01, 95% *CrI* = 0.998, 1.017).

When removing all items with zero variance (14/158 items), the pattern of results remained the same. AoA remained a strong predictor of children's scores (*OR* = 6.83, 95% *CrI* = 4.87, 9.69), whilst a small interaction between CI and children's age (*OR* = 0.99, 95% *CrI* = 0.986, 0.999) showed that younger children's scores slightly increased for items with a higher CI. The same held for the analysis with the first thirty items only (AoA: *OR* = 10.42, 95% *CrI* = 6.82, 16.05; age*CI: *OR* = 0.99, 95% *CrI* = 0.983, 0.997).

Discussion

The second study replicated the findings from the first study, in that AoA was a strong and reliable predictor of children's scores, whereas CI was not: although we found a small interaction between CI and children's age, this interaction seemed counterintuitive as it showed that younger children's scores slightly improved for words that were more, rather than less, complex. In line with previous research, we also found that age was a predictor of children's scores in the majority language but not in the minority language. In contrast, proportion of input was a strong predictor of children's scores in both languages.

In addition, we found that children's performance on the CLT correlated with two different versions of the SRT, as well as with a standardized measure of children's language proficiency, the CELF-WS task, demonstrating again that the CLT is indeed an indication of children's language proficiency.

Finally, there was quite some variability in children's scores on the CLT, especially in the younger age groups, and none of the age groups were at ceiling, showing that the CLT is sensitive enough to capture differences between bilingual children of a wide age range.

General discussion

In this paper we assessed the validity of using the CLT as a measure of language proficiency in bilingual children, by looking at the extent to which: (a) monolingual Spanish-speaking and Dutch-speaking children matched on age and SES obtained similar scores; (b) the CLT correlated with other measures of language proficiency in monolingual and bilingual children; and (c) the factors underlying the CLT's construction – i.e., target words' estimated AoA and CI – were predictive of children's scores. In addition, we tested children within a wider age range than previously tested in CLT studies, to explore whether the CLT could be used with children in and beyond the pre-school years.

We found that the CLT correlated with several other measures of children's language proficiency, including two different sentence repetition tasks, and a standardized test targeting morphosyntax. This shows that the CLT taps into children's language proficiency, irrespective of whether they are acquiring one language or two. We also found that, for the production part of the CLT, there was much variability in children's scores in the younger age groups, and none of the age groups were at ceiling. These findings show that the production subtask of the CLT is sensitive enough to capture differences between both monolingual and bilingual children, at least up to the age of eight (although it is worth noting that the absence of ceiling effects in the older children may have primarily been due to a few specific items, i.e., items that virtually none of the children knew). In contrast, almost all monolingual children performed at ceiling on the comprehension subtask, suggesting that the comprehension part of the CLT may only be suitable for children younger than the age range investigated here, at least when used with monolinguals.

Thus far, the results concerning the validity of the CLT as a measure to compare children's language proficiency cross-linguistically sound promising. There were, however, two findings that require us to temper this conclusion. The first is that monolingual Spanish-speaking and Dutch-speaking children matched for age and parental education did not obtain similar scores, with the former outperforming the latter on the CLT's production subtask. Secondly, the rationale behind the CLT's construction procedure was partially undermined by the lack of a reliable effect of CI on children's scores. We now discuss these two findings in more detail.

Difference between the Spanish and Dutch monolinguals

The difference between the Spanish and Dutch monolingual children held for both the 40-item version and the "official" 30-item version. On average, the Spanish-speaking children outperformed the Dutch-speaking children by 2 to 3 items on the 40-item version and by 3 to 4 items on the 30-item version. Although it is possible that this difference is sample-related – i.e., the two groups of monolingual children may have differed on another relevant factor besides age and level of parental education – the fact that the two groups did not differ on their SRT scores rather suggests that the Spanish production subtask may be easier than the Dutch one.

Łuniewska, Haman and Hansen (2016b) suggested that there may also be real differences in the timing and pace of lexical development across languages and cultures (e.g., Bleses *et al.*, 2008), which are related to properties of the languages themselves as well as contextual differences, such as characteristics of children's language input and parental practices. For example, they suggest that children's lexical development may be slower in phonologically non-transparent languages (e.g., Bleses *et al.*, 2008), as well as in highly inflected languages which are not completely regular (e.g., Thordardottir, Weismer & Evans, 2002; cf. Aksu-Koç & Ketrez, 2003). Furthermore, based on the finding that quantity of child-directed speech predicts the size of children's vocabulary at a later age (e.g., Schneidman & Goldin-Meadow, 2012), Łuniewska *et al.* (2016b) suggest that children's lexical development may be slower in cultures where children have been found to hear less child-directed speech (e.g., Lieven, 1994; Schneidman & Goldin-Meadow, 2012).

Only some of these factors are relevant in the present context. Both monolingual groups came from Western-European countries, and thus differences in the type of input and parental practices would presumably not have been very large. In contrast, Dutch and Spanish do differ in terms of phonology and morphology, but it is difficult to predict whether these differences would favour more rapid word learning in Spanish over Dutch. Although Dutch is phonologically more complex (it possesses more phonemes in general, and many words contain consonant clusters, cf. Spanish), it is unknown whether this would result in slower vocabulary development in Dutch-speaking children compared to their Spanish-speaking peers. The available data, albeit tentative, suggest this is not the case: in the CDI data reported in Bleses *et al.* (2008), Dutch-speaking children's production vocabulary at age 2;6 is numerically larger, rather than smaller, than that of Spanish-speaking children (but no statistical tests are reported). At the present time, it remains a possibility that the Dutch and Spanish CLT are not equally complex due to factors other than CI or AoA that are presently unknown.

Validity of the CLT's construction procedure

In line with previous findings, the estimated AoA of the target words was a strong predictor of children's scores, both for the monolingual and bilingual children in this study. As there appear to be strong correlations between adults' estimated AoA and parental reports on children's early lexical development (Łuniewska *et al.*, 2016a), AoA was previously assumed to be a good measure of the words children have acquired (Haman *et al.*, 2015; Hansen *et al.*, 2017). We therefore hypothesized that AoA may have a smaller effect in older children; as most of the CLT's target words have an estimated AoA of six or lower (cf. Figure 1), children beyond this age should already have acquired these words. However, in the present study, we did not find such an interaction with children's age. In other words, AoA was as strong a predictor in the younger as in the older children. This suggests that AoA may not necessarily be a good indicator of the EXACT age at which words are acquired, but rather of the order in which words are acquired relative to each other.

In contrast to AoA, CI was neither a strong nor reliable predictor of children's scores, and when it had an effect (i.e., for the monolingual children's comprehension scores, and for the younger bilingual children), the direction was opposite than expected: children's scores seemed to slightly increase, rather than decrease, as words became more complex. Because the CI is largely based on phonological complexity, and it may be the case that this is no longer relevant for children beyond the age the CLT was intended for (i.e., five-year-olds), we also included an interaction between CI and children's age in our analyses. There was a small but significant interaction for the bilinguals, but this showed the counterintuitive pattern described above. Although phonological complexity may influence lexical development in children younger than investigated here, we agree with Hansen *et al.* (2017) that at least some of the properties included in the CI may simply not play a decisive role in children's lexical development or may vary cross-linguistically. For example, word length has a great impact on the CI score, implicitly assuming that word difficulty increases linearly as a function of word length. As already discussed by Hansen *et al.*, this is probably not the case as cross-linguistic research shows that disyllabic, not monosyllabic, words seem to be the most prevalent in young children's vocabularies in most languages (Vihman & Croft, 2007). In addition, there seem to be substantial cross-linguistic differences in the average word length of children's early words: whereas Italian children's first fifty words contain mainly disyllables and polysyllables, Danish children for example mainly produce monosyllables (Garmann, Hansen, Simonsen & Kristoffersen, 2019). Likewise, the number of morphemes in a word may be less decisive for children's vocabulary development than, e.g., the morphological transparency of a language (as, for example, evidenced by the early acquisition of morphology in Turkish, a morphologically rich but also transparent language; Aksu-Koç & Ketrez, 2003).

If the CI were to be used for constructing new vocabulary tasks, including new versions of the CLT, it would thus be important to take these cross-linguistic differences into account. This may however prove too difficult a task, considering both the lack of relevant empirical data (e.g., on the preferred word length of children's early words in each language), and the difficulties involved in operationalising variables, such as morphological transparency across languages. Two other factors found by Hansen *et al.* (2017) to sometimes influence children's scores over and above the effect of AoA were a word's frequency in child-directed speech and a word's imageability. While imageability ratings would be fairly easy to obtain, frequency in child-directed speech would not, as only a few corpora exist and for a limited number of languages. However, since frequency and imageability did not always explain additional variance on children's scores, it remains unclear whether additional measures of complexity alongside AoA would notably improve comparability of different language versions.

Implications and conclusion

Hansen *et al.* (2017) concluded that although the CI does not seem to have any predictive value, the fact that AoA was a significant and robust predictor of children's scores may be enough to ensure cross-linguistically comparable CLT versions. The fact that in their study there was no difference between the Polish and Norwegian monolinguals' performance supports this conclusion. In contrast, even though the mean AoA of the items in the Dutch and Spanish production subtasks was the same, the Spanish monolinguals outperformed the Dutch monolinguals in the present study. This study thus shows that instruments developed by a shared protocol are not necessarily equivalent, and that each language version of an assessment instrument needs rigorous testing with monolingual speakers of each language before the instrument is used amongst bilingual children.[3]

In terms of possible comparisons, our findings imply that one should be careful when comparing children's CLT scores across languages, as well as when directly comparing the two languages of bilingual children (e.g., when using these scores to compute relative proficiency). Comparisons within languages are still possible, i.e., one can compare scores across monolingual children who speak the same language, and one can compare bilingual children from the same language pair, as long as no direct claims about language dominance are made. In other words, the CLT may still prove to be a useful tool for measuring children's language proficiency, but the present study casts some doubts on the extent to which the CLT can be used as a tool to compare children's language proficiency cross-linguistically. Further research with more languages and language combinations is needed to ascertain whether these doubts are well founded. Moreover, and more generally, this study underscores the need for large-scale norming studies in the development of tasks with the aim of cross-linguistic equivalence: these would not only determine whether the differences between the Dutch- and Spanish-speaking children in the current study (or any other study with similar results) should be attributed to sample or task characteristics, they would also provide the standardized scores required for cross-linguistic comparisons to be made in the eventuality that such differences remain.

Supplementary Material. Supplementary material can be found online at https://doi.org/10.1017/S030500092000063X

[3] We thank one of the reviewers for this well-formulated conclusion.

Acknowledgements. We would like to thank Teresa Quesada Calvo for testing the Spanish monolingual children, and Cristóbal Lozano for his help in coordinating the study in Spain. Furthermore, we are indebted to several team members and research assistants of the 2in1 project, for their help with testing, transcribing and scoring. We also thank two anonymous reviewers for their comments on a previous version of this manuscript. Finally, we would like to thank all children, parents and primary schools that participated. This study was funded by a VIDI grant (276-70-027) awarded by NWO, the Dutch Organization for Scientific Research, to the second author.

References

Aksu-Koç, A., & Ketrez, F.N. (2003). Early verbal morphology in Turkish: Emergence of inflections. In D. Bittner, W.U. Dressler, & M. Kilani-Schoch (Eds.), *Development of verb inflection in first language acquisition: A cross-linguistic perspective* (pp. 27–52). Berlin: Mouton de Gruyter.

Albers, C. J., Kiers, H. A. L., & van Ravenzwaaij, D. (2018). Credible confidence: A pragmatic view on the frequentist vs Bayesian debate. *Collabra: Psychology*, *4*(1), 31. doi: 10.1525/collabra.149

Altman, C., Goldstein, T., & Armon-Lotem, S. (2017). Quantitative and qualitative differences in the lexical knowledge of monolingual and bilingual children on the LITMUS-CLT task. *Clinical Linguistics & Phonetics*, *31*(11–12), 931–954.

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bedore, L. M., Peña, E. D., Griffin, Z. M., & Hixon, J. G. (2016). Effects of age of English exposure, current input/output, and grade on bilingual language performance. *Journal of Child Language*, *43*(3), 687–706.

Bedore, L. M., Peña, E. D., Summers, C. L., Boerger, K. M., Resendiz, M. D., Greene, K., Bohman, T. M., & Gillam, R. B. (2012). The measure matters: Language dominance profiles across measures in Spanish–English bilingual children. *Bilingualism: Language and Cognition*, *15*(3), 616–629.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., & Basboll, H. (2008). Early vocabulary development in Danish and other languages: A CDI-based comparison. *Journal of Child Language*, *35*(3), 619–650.

Blom, E., Küntay, A. C., Messer, M., Verhagen, J., & Leseman, P. (2014). The benefits of being bilingual: Working memory in bilingual Turkish–Dutch children. *Journal of Experimental Child Psychology*, *128*, 105–119.

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning Dutch as a second language. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1747–1760.

Bohnacker, U., Lindgren, J., & Öztekin, B. (2016). Turkish-and German-speaking bilingual 4-to-6-year-olds living in Sweden: Effects of age, SES and home language input on vocabulary production. *Journal of Home Language Research*, *1*, 17–41.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S. Y., Pascual, L., Pêcheux, M. G., Ruel, J., Venuti, P., & Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development*, *75*(4), 1115–1139.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Chondrogianni, V., & Marinis, T. (2011). Differential effects of internal and external factors on the development of vocabulary, tense morphology and morpho-syntax in successive bilingual children. *Linguistic Approaches to Bilingualism*, *1*(3), 318–345.

Clark, E. V., & Hecht, B. F. (1983). Comprehension, production, and language acquisition. *Annual review of psychology*, *34*(1), 325–349.

Garmann, N. G., Hansen, P., Simonsen, H. G., & Kristoffersen, K. E. (2019). The phonology of children's early words: trends, individual variation and parents' accommodation in child-directed speech. *Frontiers in Communication*, *4*, 10. doi: 10.3389/fcomm.2019.00010

Gathercole, V. C. M., Thomas, E. M., & Hughes, E. (2008). Designing a normed receptive vocabulary test for bilingual populations: A model from Welsh. *International Journal of Bilingual Education and Bilingualism*, *11*(6), 678–720.

Gatt, D., Attard, D., Łuniewska, M., & Haman, E. (2017). The effects of bilingual status on lexical comprehension and production in Maltese five-year-old children: A LITMUS-CLT study. *Clinical Linguistics & Phonetics*, *31*(11–12), 844–873.

Gentner, D. (1982). *Why nouns are learned before verbs: Linguistic relativity versus natural partitioning* (Technical Report No. 257). Cambridge, MA: Bolt Beranek and Newman Inc. Retrieved from <http://files.eric.ed.gov/fulltext/ED219724.pdf>

Haitana, T., Pitama, S., & Rucklidge, J. J. (2010). Cultural biases in the Peabody Picture Vocabulary Test-III: Testing Tamariki in a New Zealand sample. *New Zealand Journal of Psychology*, *39*(3), 24–34.

Haman, E., Łuniewska, M., & Pomiechowska, B. (2015). Designing Cross-linguistic Lexical Tasks (CLTs) for bilingual preschool children. In S. Armon-Lotem, J. De Jong, & N. Meir (eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 196–240). Bristol: Multilingual Matters.

Haman, E., Łuniewska, M., Hansen, P., Simonsen, H. G., Chiat, S., Bjekić, J., Blažienė, A., Chyl, K., Dabašinskienė, I., Engel de Abreu, P., Gagarina, N., Gavarró, A., Håkansson, G., Harel, E., Holm, I. E., Kapalková, S., Kunnari, S., Levorato, C., Lindgren, J., Mieszkowska, K., Montes Salarich, L., Potgieter, A., Ribu, I., Ringblom, N., Rinker, T., Roch, M., Slančová, D., Southwood, F., Tedeschi, R., Tuncer, A. M., Ünal-Logacev, Ö., Vuksanović, J., & Armon-Lotem, S. (2017). Noun and verb knowledge in monolingual preschool children across 17 languages: Data from Cross-linguistic Lexical Tasks (LITMUS-CLT). *Clinical linguistics & Phonetics*, *31*(11–12), 818–843.

Hansen, P., Łuniewska, M., Simonsen, H. G., Haman, E., Mieszkowska, K., Kołak, J., & Wodniecka, Z. (2019). Picture-based vocabulary assessment versus parental questionnaires: A cross-linguistic study of bilingual assessment methods. *International Journal of Bilingualism*, *23*(2), 437–456.

Hansen, P., Simonsen, H. G., Łuniewska, M., & Haman, E. (2017). Validating the psycholinguistic aspects of LITMUS-CLT: Evidence from Polish and Norwegian. *Clinical Linguistics & Phonetics*, *31*(11–12), 910–930.

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of Child Language*, *39*(1), 1–27.

Hoff, E., Welsh, S., Place, S., & Ribot, K. M. (2014). Properties of dual language input that shape bilingual development and properties of environments that shape dual language input. In T. Grüter & J. Paradis (Eds.), *Input and experience in bilingual development* (pp. 119–140). Amsterdam/Philadelphia: John Benjamins.

Kapalková, S., & Slančová, D. (2017). The vocabulary profile of Slovak children with primary language impairment compared to typically developing Slovak children measured by LITMUS-CLT. *Clinical Linguistics & Phonetics*, *31*(11–12), 893–909.

Khoury Aouad Saliby, C., Dos Santos, C., Kouba Hreich, E., & Messarra, C. (2017). Assessing Lebanese bilingual children: The use of Cross-linguistic Lexical Tasks in Lebanese Arabic. *Clinical Linguistics & Phonetics*, *31*(11–12), 874–892.

Kort, W., Schittekatte, M., & Compaan, E. (2010). *CELF-4-NL: Clinical Evaluation of Language Fundamentals*. Amsterdam: Pearson.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177.

Lenth, R. (2019). emmeans: Estimated marginal means, aka least-squares means. R package version 1.3.3. Retrieved from <https://CRAN.R-project.org/package=emmeans>

Lieven, E. V. M. (1994). Crosslinguistic and crosscultural aspects of language addressed to children. In C. Gallaway & B. J Richards (Eds.), *Input and interaction in language acquisition* (pp. 56–73). Cambridge: Cambridge University Press.

Lindgren, J. (2018). *Developing narrative competence: Swedish, Swedish-German and Swedish-Turkish children aged 4–6* (Doctoral dissertation). Retrieved from <http://uu.diva-portal.org/smash/person.jsf?pid=authority-person%3A20593&dswid=5064>

Lindgren, J., & Bohnacker, U. (2019). Vocabulary development in closely-related languages: Age, word type and cognate facilitation effects in bilingual Swedish-German preschool children. *Linguistic Approaches to Bilingualism*. doi: 10.1075/lab.18041.lin

**Łuniewska, M., Haman, E., Armon-Lotem, S., Etenkowski, B., Southwood, F., Anđelković, D., Blom, E., Boerma, T., Chiat, S., de Abreu, P. E., Gagarina, N., Gavarro, A., Hakansson, G., Hickey, T., de Lopez, K. J., Marinis, T., Popovic, M., Thordardottir, E., Blaziene, A., Sanchenz, M. C.,

Dabasinskiene, I., Ege, P., Ehret, I-A., Fritsche, N-A., Gatt, D., Janssen, B., Kambanaros, M., Kapalkova, S., Kronqvist, B., Kunnari, S., Levorato, C., Nenonen, O., Fhlannchadha, S. N., O'Toole, C., Polišenská, K., Pomiechowska, B., Ringblom, N., Rinker, T., Roch, M., Savic, M., Slancova, D., Tsimpli, I. M., & Ünal-Logacev, Ö. (2016a). Ratings of age of acquisition of 299 words across 25 languages. Is there a cross-linguistic order of words? *Behavior Research Methods*, *48* (3), 1154–1177.

Łuniewska, M., Haman, E., & Hansen, P. (2016b). Is there a road to universal assessment of lexical knowledge in multilingual children? Cross-cultural aspects of Cross-linguistic Lexical Tasks. In H.-O. Enger, M. I. N. Knoph, K. E. Kristoffersen & M. Lind (Eds.), *Helt fabelaktig! Festskrift til Hanne Gram Simonsen på 70-årsdagen [Absolutely fabulous! Festschrift for Hanne Gram Simonsen on her 70th birthday]* (pp. 147–165). Oslo, Norway: Novus Forlag.

Marinis, T., & Armon-Lotem, S. (2015). Sentence repetition. In S. Armon-Lotem, J. De Jong, & N. Meir (Eds.), *Assessing multilingual children: Disentangling bilingualism from language impairment* (pp. 95–122). Bristol: Multilingual Matters.

Meara, P. M. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjagr & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 33–51). Cambridge: Cambridge University Press.

Miralpeix, I., & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, *56*(1), 1–24.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123.

Muñoz-Sandoval, A. F., Cummins, J., Alvarado, C. G., & Ruef, M. L. (2005). *Bilingual Verbal Ability Tests – Normative Update*. Itasca, IL: Riverside Insights.

Paradis, J., & Nicoladis, E. (2007). The influence of dominance and sociolinguistic context on bilingual preschoolers' language choice. *International Journal of Bilingual Education and Bilingualism*, *10*(3), 277–297.

Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development*, *78*(4), 1255–1264.

Polišenská, K. (2011). *The influence of linguistic structure on memory span: repetition tasks as a measure of language ability* (Doctoral dissertation). Retrieved from <http://openaccess.city.ac.uk/id/eprint/682/1/Polisenska_PhD_Jan2012.pdf>

Potgieter, A. P., & Southwood, F. (2016). A comparison of proficiency levels in 4-year-old monolingual and trilingual speakers of Afrikaans, isiXhosa and South African English across SES boundaries, using LITMUS-CLT. *Clinical Linguistics & Phonetics*, *30*(2), 87–100.

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Real Academia Española. (2005). *Diccionario panhispánico de dudas*. Madrid, Spain: Santillana.

Ringblom, N., & Dobrova, G. (2019). Holistic Constructions in Heritage Russian and Russian as a Second Language: Divergence or Delay? *Scando-Slavica*, *65*(1), 94–106.

Schneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental Science*, *15*(5), 659–673.

Thordardottir, E. T., Weismer, S. E., & Evans, J. L. (2002). Continuity in lexical and morphological development in Icelandic and English-speaking 2-yearolds. *First Language*, *22*(1), 3–28.

Unsworth, S. (2013). Assessing the role of current and cumulative exposure in simultaneous bilingual acquisition: The case of Dutch gender. *Bilingualism: Language and cognition*, *16*(1), 86–110.

Unsworth, S., Chondrogianni, V., & Skarabela, B. (2018). Experiential measures can be used as a proxy for language dominance in bilingual language acquisition research. *Frontiers in Psychology*, *9*, 1809. doi: 10.3389/fpsyg.2018.01809

Van der Velde, M. (2004). L'acquisition des déterminants en L1: une étude comparative entre le français et le néerlandais [The acquisition of L1 determiners: a comparative study of French and Dutch]. *Acquisition et Interaction en Langue Etrangère*, *21*(1), 9–46.

van Wonderen, E., Blom, E., Boerma, T., Janssen, B., Unsworth, S., & van Dijk, C. N. (2017). Cross-linguistic Lexical Task – Dutch. Retrievable from <http://psychologia.pl/clts/>

Vihman, M., & Croft, W. (2007). Phonological development: Toward a "radical" templatic phonology. *Linguistics*, *45*(4), 683–725.

Wiig, E. H., Secord, W. A., & Semel, E. M. (2006). *CELF-4 Spanish: Clinical Evaluation of Language Fundamentals, Fourth edition, Spanish*. San Antonio, TX: PsychCorp.

Wiig, E. H., Secord, W. A., & Semel, E. M. (2009). *CELF Preschool-2 Spanish: Clinical Evaluation of Language Fundamentals - Preschool 2, Spanish edition*. San Antonio, TX: Pearson.

Wiig, E. H., Secord, W. A., & Semel, E. M., & De Jong, J. (2012). *CELF Preschool-2-NL: Clinical Evaluation of Language Fundamentals: Preschool - Nederlandstalige versie*. Amsterdam: Pearson.

Yip, V., & Matthews, S. (2000). Syntactic transfer in a Cantonese–English bilingual child. *Bilingualism: Language and Cognition*, *3*(3), 193–208.

Cite this article: Van Wonderen E, Unsworth S (2020). Testing the validity of the Cross-Linguistic Lexical Task as a measure of language proficiency in bilingual children. *Journal of Child Language* 1–25. https://doi.org/10.1017/S030500092000063X